THÈSE

Pour obtenir le grade de

Docteur

Délivré par

L'Université de Montpellier

&

L'Université de Tunis el Manar

Faculté des Sciences de Tunis

Préparée au sein de l'école doctorale GAIA

Et de l'unité de recherche UMR5554-ISEM

Spécialité :

Génétique-Evolution

Présentée par Mr Ahmed Souissi

Analyse génétique et morphologique de l'isolement reproductif partiel dans la zone d'hybridation de *Solea senegalensis* et *Solea aegyptiaca* en Tunisie

Soutenue le 12/12/2016 devant le jury composé de

M^me^ Neila TRIFI, Professeur, Faculté des Sciences de Tunis	Présidente
Mr. Didier AURELLE, Maître de Conférences, Université Aix-Marseille	Rapporteur
M^me^ Karima FADHLAOUI, Maître de Conférences, Institut des Sciences Biologiques Beja	Rapporteur
Mr. François BONHOMME, Directeur de Recherches CNRS, Université de Montpellier	Co-Directeur de thèse
M^me^ Lilia BAHRI-SFAR, Maître de Conférences, Faculté des Sciences de Tunis	Co-Directrice de thèse
Mr. Nicolas BIERNE, Directeur de Recherches CNRS, Université de Montpellier	Examinateur
Mr. Pierre-Alexandre GAGNAIRE, Chargé de Recherches CNRS, Université de Montpellier	Membre invité

Table des matières

Résumé 3
Introduction 5
1. Spéciation et notion d'espèce : 5
1.1. Les mécanismes de l'isolement reproductif 6
2. Etude de la spéciation dans les zones d'hybridation 10
2.1. Flux génique au travers d'une zone d'hybridation 11
2.2. Inférence démo-génétique de l'histoire du flux génique 14
3 Génomique de la spéciation 17
4 Originalité des organismes marins et modèle d'étude 20
5 Objectifs de la thèse 22
Chapitre I : Hybridation introgressive et transgression morphologique dans la zone de contact de deux espèces méditerranéennes du genre Solea 24
1 Introduction 24
2 Article: Introgressive hybridization and morphological transgression in the contact zone between two Mediterranean Solea species 26
2.1. Introduction 27
2.2. Materials and Methods 30
2.3. Results 33
2.4. Discussion 39
Chapitre II : Acquisition d'un jeu de données de SNPs haute densité 43
1. Echantillonnage, Construction des librairies RAD et Séquençage 43
1.1. Echantillonnage 43
1.2. Construction des librairies et séquençage 44
2. Préparation des données et analyses bio-informatiques 45
3. Analyse descriptive du polymorphisme 48
4. Résultats 48
4.1. Obtention du jeu de données final 48
4.2. Distribution de la variabilité génétique 50
5. Discussion 53
Chapitre III : Histoire du flux génique et signature génomique des échanges génétiques à travers la zone d'hybridation semi-perméable 54
Introduction 54

1. Analyses démo-génétiques 55
1.1. Spectre joint des fréquences allèliques 55
1.2. δaδi et ajustement des modèles démographiques 57
2 Clines génomiques 59
3 Clines géographiques 61
4 Résultats 62
4.1. Ajustement du meilleur modèle démo-génétique de divergence 62
4.2. Probabilité de l'introgression 66
4.3. Clines génomiques 68
4.4. Géographie de l'introgression 71
4.5. Analyse comparative de l'introgression 73
5 Discussion 78
Conclusion 85
Bibliographie 91

Résumé

Analyse génétique et morphologique de l'isolement reproductif partiel dans la zone d'hybridation de *Solea senegalensis* et *Solea aegyptiaca* en Tunisie

Les processus d'hybridation et d'introgression occupent une place importante dans l'étude du mécanisme de spéciation, car ils permettent d'analyser les conséquences évolutives des échanges génétiques entre espèces partiellement isolées. Ici, nous nous sommes intéressés à la zone d'hybridation entre les soles *S. senegalensis* et *S. aegyptiaca* au niveau des côtes nord-tunisiennes, pour comprendre l'origine de leur diversification. Nous avons premièrement caractérisé les conséquences phénotypiques de l'hybridation sur la variabilité morphologique. Nos résultats montrent que si l'introgression provoque la convergence de certains caractères morphologiques, elle est en revanche à l'origine de transgressions et de distorsions morphologiques sur d'autres traits, pouvant refléter une condition plus faible des génotypes recombinants. Les phénomènes d'incompatibilité génétique associés à une éventuelle contre-sélection des hybrides sont supposés créer une perméabilité différentielle du génome face au flux génique entre espèces. Pour étudier cette semi-perméabilité à l'échelle du génome, nous avons établi un jeu de données de polymorphisme par la méthode RAD-seq. Ceci nous a permis de génotyper 200 individus pour 10 756 marqueurs SNP, qui nous ont permis de caractériser les flux géniques entre ces deux espèces à travers trois approches complémentaires. La première est basée sur une reconstitution démographique de l'histoire des échanges génétiques qui intègre les effets de la semi-perméabilité des génomes. La seconde approche se focalise sur l'évolution spatiale des fréquences allèliques à travers la zone d'hybridation. La dernière méthode, dite des clines génomiques, compare le comportement de chaque locus au patron d'introgression moyen attendu sous l'hypothèse de neutralité. Nos résultats indiquent que *S. senegalensis* et *S. aegyptiaca* ont subi une divergence ancienne en allopatrie suivie d'un contact secondaire récent. Seule une faible proportion du génome parvient à introgresser de manière asymétrique dans la zone hybride qui en résulte, selon une grande diversité de patrons d'introgression dont nous discutons les origines possibles.

Genetic and morphological analysis of partial reproductive isolation between *Solea senegalensis* and *Solea aegyptiaca* in the Tunisian hybridization area

Hybridisation and introgression processes have an important place in the study of speciation as they allow to analyse the evolutionary consequences of genetic exchanges between partially isolated species. Here we are interested in the hybrid zone resulting from the contact between the soles *S. senegalensis* and *S. aegyptiaca* along the North Tunisian coast. First, we studied the consequences of hybridisation on phenotypic variation. This allowed us to evidence that phenotypic convergence in hybrids was accompanied by phenotypic transgression and morphological distortions for certain traits that seem to reflect a reduced condition of hybrids. Possible genetic incompatibilities between species should be responsible for the differential permeability of the genome to gene flow, thereby creating a semi-permeable barrier. To study this barrier at the genome scale, we have produced a polymorphism dataset using the RAD-seq method. This allowed us to genotype 200 individuals at 10,756 SNP markers and to characterise genomic patterns of gene flow between the two species through three complementary approaches. The first is based on a reconstruction of the demographic history of the genetic and exchanges that incorporates the effects of the semi-permeability to gene flow. The second approach focuses on the spatial evolution of allele frequencies across the hybrid zone. The last method, called genomic clines, compares the behaviour of each locus to the average introgression pattern expected under the hypothesis of neutrality. Our results indicate that *S. senegalensis* and *S. aegyptiaca* underwent ancient divergence in allopatry followed by a recent secondary contact. Only a small proportion of the genome can asymmetrically introgress across the hybrid zone, resulting in a variety of introgression patterns of which we discuss the possible origins.

Introduction

1. Spéciation et notion d'espèce :

La formation des espèces - la spéciation - et la compréhension des mécanismes qui lui sont associés, reste l'une des principales préoccupations des biologistes de l'évolution plus de 150 ans après la publication de *L'origine des espèces* par Darwin (1959) (Sobel *et al.,* 2010). Lorsqu'on veut étudier la spéciation, il est impératif de la définir clairement et de savoir quand elle commence et quand elle finit. Le concept d'espèce biologique proposé par (Mayr, 1942) se base sur la notion de l'isolement reproductif pour définir une espèce. Il propose que « *L'espèce est la plus grande unité au sein de laquelle les individus peuvent* ***effectivement*** *ou* ***potentiellement*** *se reproduire entre eux et engendrer une* ***descendance viable*** *et* ***féconde****. Les individus d'une même espèce sont donc génétiquement isolés d'autres ensembles équivalents du point de vue reproductif* ». Cependant, cette définition peut montrer ses limites lorsque des événements d'hybridation se produisent naturellement entre espèces proches. C'est le cas dans cette présente étude portant sur deux espèces de soléidés ; *Solea senegalensis* et *Solea aegyptiaca* qui produisent des hybrides viables et féconds en conditions naturelles. On se confronte alors à la difficulté de définir la notion d'espèce qui reste encore un sujet de débat récurrent. Dans la mesure où la restriction du flux génique est une étape fondamentale du processus de spéciation, l'isolement reproductif absolu pourrait ne pas être un critère indispensable dans le cadre d'un concept biologique de l'espèce étendu. Ainsi, l'observation d'un phénomène d'hybridation rare entre deux entités qui maintiennent des différences phénotypiques et génétiques claires resterait compatible avec un processus de spéciation potentiellement irréversible. Du point de vue de la compréhension des mécanismes évolutifs impliqués dans la spéciation, il est donc nécessaire d'étudier les conditions dans lesquelles le flux génique se réduit. Cela nécessite d'identifier les mécanismes d'isolement reproductif mis

en jeu, afin de mieux comprendre le maintien de la divergence si celui-ci est avéré, en dépit de l'effet potentiellement homogénéisateur du flux génique dû à l'hybridation.

1.1. Les mécanismes de l'isolement reproductif

Etudier la spéciation revient donc en grande partie à étudier l'évolution des mécanismes de l'isolement reproductif (Coyne & Orr, 2004). Ces mécanismes sont nombreux et peuvent être classés en pré- ou post-zygotiques selon le stade de vie auquel ils interviennent (Dobzhansky, 1937).

- Isolement pré-zygotique : Les mécanismes pré-zygotiques sont ceux qui interviennent avant ou pendant la reproduction diminuant ainsi la probabilité de rencontre et de fécondation entre gamètes d'individus appartenant à des entités différentes. Parmi ces mécanismes, on peut citer le choix de l'habitat, la philopatrie (reproduction dans des lieux définis qui différent selon les entités), l'homogamie (choix préférentiel de partenaires de la même entité) ou bien encore la divergence morpho-anatomique des appareils génitaux (fréquente chez les insectes par exemple).

- Isolement post-zygotique Les mécanismes d'isolement post-zygotiques sont ceux qui s'expriment après la fécondation. Dans ce cas, l'existence de barrières d'isolement reproductif peut se traduire par la stérilité et/ou la non-viabilité des hybrides. Ces barrières peuvent être également subdivisées en fonction de leur degré de dépendance vis-à-vis de l'environnement. On parle alors de barrières extrinsèques lorsque la valeur sélective des hybrides dépend de l'environnement biotique et/ou abiotique, telles que la non viabilité écologique ou l'inadaptation comportementale. On parle de barrières intrinsèques ou endogènes quand il s'agit de problèmes indépendants des conditions environnementales qui touchent les hybrides comme la non viabilité ou la stérilité hybride.

La mise en place de barrières d'isolement reproductif est en étroite relation avec le contexte biogéographique expérimenté par les populations au cours de leur histoire. Par exemple les oscillations climatiques du Pléistocène au cours des 2,5 derniers millions d'années ont été à l'origine de profondes modifications des aires de répartitions de plusieurs espèces terrestres (Hewitt, 2000) et marines (Palumbi, 1994). Les périodes glaciaires ont notamment modifié les routes migratoires, et contraint de nombreuses espèces à migrer vers des régions refuges pour suivre leur préférendum thermique, provoquant une fragmentation de leurs aires de distribution. A l'inverse, les périodes interglaciaires (comme l'actuelle) ont favorisé l'extension des aires de répartition pour les espèces tempérées, provoquant ainsi la mise en contact de lignées glaciaires ayant survécu dans des refuges différents. Ces contacts entre lignées glaciaires d'une même espèce sont aujourd'hui visibles au niveau de zones de transition biogéographiques majeures, comme la transition Atlantico-Méditerranéenne matérialisées par le front courantologique Almeria-Oran chez certaines espèces marines (Patarnello *et al*., 2007). D'une manière plus générale, le contexte géographique dans lequel la spéciation se déroule peut varier en fonction de l'importance des événements biogéographiques et des adaptations écologiques qui accompagnent l'histoire évolutive des populations. Trois modes de divergence sont ainsi classiquement distingués en fonction du contexte spatial dans lequel les barrières d'isolement reproductif évoluent :

- Divergence allopatrique : la spéciation est initiée en l'absence de flux génique. Elle est le résultat de la mise en place de barrières géographiques à la migration (Mayr, 1942). En effet si deux populations se retrouvent séparées par de telles barrières, leur polymorphisme ancestral partagé va d'abord se fixer indépendamment dans chacune d'entre elles provoquant leur différentiation progressive, puis de nouvelles mutations vont se fixer de manière indépendante dans les deux populations qui continuent ainsi à diverger.

- Divergence sympatrique : elle se déroule en l'absence de barrière physique au flux génique. Ici, les individus peuvent donc se croiser librement, et les probabilités de croisement entre individus ne dépendent que de leurs génotypes (Kondrashov & Mina, 1986). Ce mécanisme nécessite l'évolution conjointe d'une adaptation locale et d'un choix d'habitat correspondant au génotype.

- Divergence parapatrique : dans ce mode de spéciation, la divergence se fait entre deux populations susceptibles d'échanger encore des migrants. Cependant les flux de gènes ne sont pas totalement libres, et la migration est donc intermédiaire entre les modes de spéciation sympatrique et allopatrique. Les différents modes de spéciation précédemment énoncés ne font pas intervenir les mêmes forces évolutives principales dans l'évolution des barrières d'isolement (Sobel *et al.,* 2010). Lors de la divergence allopatrique, c'est la dérive génétique et éventuellement la sélection qui jouent un rôle majeur dans l'accumulation des barrières reproductives. Dans la spéciation sympatrique, c'est l'antagonisme entre la migration qui s'oppose à la différentiation et la sélection qui détermine le processus de différentiation.

1.2. Bases génétiques de l'isolement reproductif

La dépression hybride est un fait d'observation courante (Dobzhansky, 1937; Coyne & Orr, 2004). Ce fait est la manifestation d'un isolement reproductif dont les bases génétiques peuvent être expliquées génétiquement par le biais de modèles comme celui de la sous dominance ou encore le modèle des incompatibilités dites de Dobzhansky-Muller (Dobzhansky, 1937; Muller, 1942; Coyne & Orr, 2004; Presgraves, 2010). Le modèle de Dobzhansky-Muller (Figure 1) permet notamment d'expliquer comment des mutations provoquant une diminution de la valeur sélective des individus hybrides peuvent se fixer indépendamment dans les populations parentales sans entrainer d'effet délétère. Dans ce

modèle, seules les nouvelles combinaisons alléliques produites par hybridation, et qui n'ont pas été soumises au filtre de la sélection lors du processus de divergence sont délétères. Ce modèle permet également d'expliquer l'observation courante d'une dépression hybride plus élevée dans les deuxièmes générations d'hybridation (Dobzhansky, 1937; Edmands, 1999) par simple effet de dominance (Turelli & Orr, 1995).

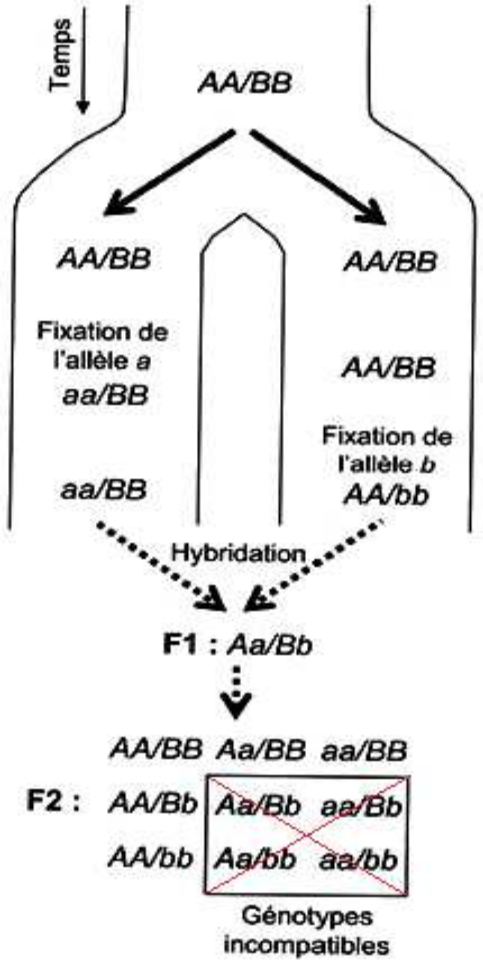


Figure 1 : Représentation schématique du modèle à deux locus (locus « A » et locus « B ») d'évolution de l'isolement post-zygotique de Dobzhansky-Muller. Deux lignées divergent indépendamment d'une lignée ancestrale unique (génotype aux deux locus AA/BB) et fixent différentiellement des allèles au cours du temps (l'allèle a au locus « A » pour la lignée à gauche et l'allèle b au locus B pour la lignée à droite). Lors de leur remise en contact, l'hybridation entre les deux lignées produit un génotype de type hybride F1 (i.e. hétérozygote aux deux locus). En F2, 9 combinaisons génotypiques sont théoriquement possibles dont 4 impliquant les allèles apparus dans chacune des lignées divergentes ; ces 4 combinaisons sont formées de génotypes incompatibles aux deux locus : les combinaisons des allèles a et b sont incompatibles ou entrainent une diminution de la valeur sélective des individus hybrides F2 (i.e. dépression d'hybridation). Source : (Ravigné *et al.*, 2010).

2. Etude de la spéciation dans les zones d'hybridation

Dans la nature, la rencontre de deux espèces partiellement isolées reproductivement se produit généralement dans une zone géographique restreinte où des hybrides sont créés par croisements interspécifiques. Ces zones intermédiaires appelées « zones de tension » sont maintenues par un équilibre entre les flux de combinaisons alléliques parentales qui arrivent par migration et la contre-sélection de certaines combinaisons hybrides qui ne sont ainsi pas ou peu exportées en dehors de la zone (Barton & Hewitt, 1985). Une zone de tension fonctionne alors comme un puits pour les génotypes hybrides (« hybrid sink », Barton, 1980). Ces zones, parfois qualifiées de laboratoires naturels pour étudier l'évolution, jouent un rôle important dans notre compréhension des mécanismes d'isolement reproductif (Barton & Hewitt, 1985). La théorie des zones d'hybridation offre une description mathématique des clines de fréquences allèliques le long d'un transect géographique passant d'une population parentale à l'autre en traversant la zone d'hybridation. L'existence d'un cline repose sur un équilibre entre deux forces antagonistes, la migration qui tend à homogénéiser les fréquences allèliques et la sélection qui tend à diminuer la fréquence des génotypes hybrides à chaque génération du fait de leur moindre valeur sélective. La forme sigmoïde des clines n'est que très peu affectée par le type de contre-sélection des combinaisons hybrides, qu'elle soit exogène (valeur sélective liée à l'environnement) ou endogène (valeur sélective liée aux interactions entre incompatibilités génétiques, (Barton & Gale, 1993; Kruuk *et al.,* 1999). Par contre, leur largeur dépendant de l'équilibre migration-sélection, les locus les plus contre-sélectionnés montrant les clines les plus abrupts (Figure 2).

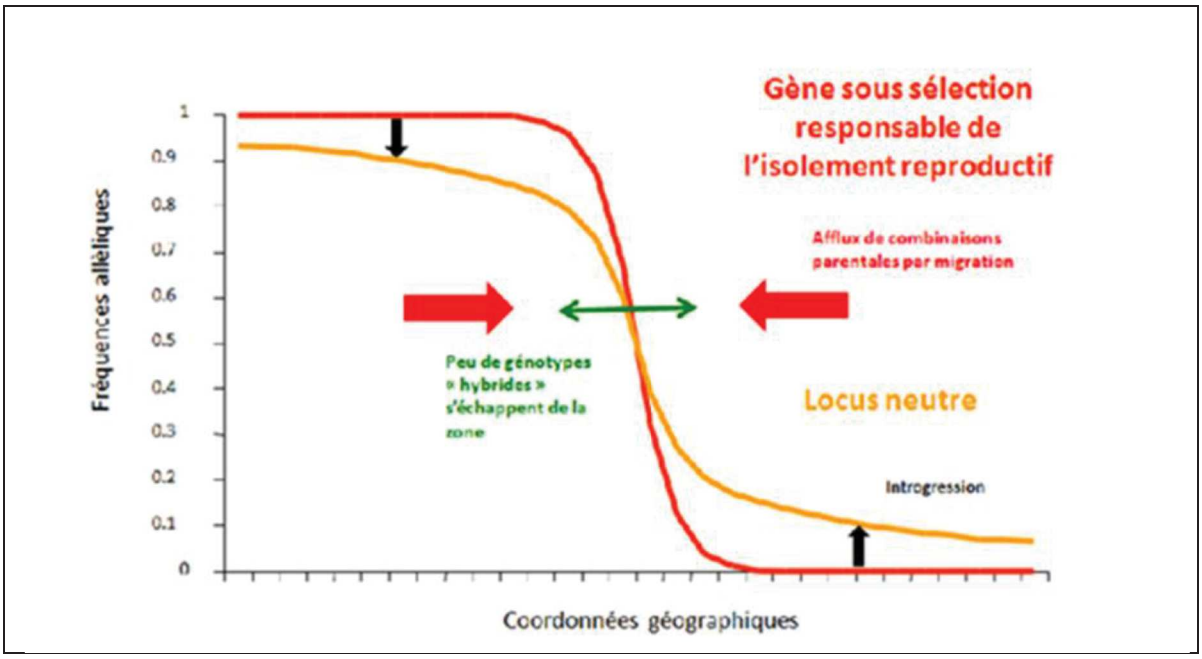


Figure 2 : Clines de variation de fréquences alléliques d'un locus neutre et un locus sous sélection.

2.1. Flux génique au travers d'une zone d'hybridation

Les zones d'hybridation sont connues pour être un ralentisseur du flux génique neutre (Barton & Hewitt, 1985). Pour expliquer la stabilité géographique des zones de tension, il est prédit que les clines de gènes sous sélection viennent se caler au niveau de barrières physiques où la migration est limitée où de zones à faible densité de populations (Barton, 1979). A cet effet mécanique s'ajoute le déséquilibre de liaison entre les gènes soumis à la contre-sélection chez les hybrides, qui est responsable de l'effet puits à hybrides et vient réduire d'autant plus la densité d'individus dans les zones d'hybridation (Barton, 1980). Ces deux effets sont pris simultanément en compte pour expliquer l'impact que jouent les zones d'hybridation sur la réduction du flux génique (Barton & Bengtsson, 1986). Les zones d'hybridation peuvent également se caler au niveau de barrières environnementales, lorsque des gènes d'adaptation à l'habitat parviennent à attirer les clines endogènes dont la localisation n'est pas contrainte par l'environnement. Ce phénomène de couplage entre clines explique l'existence de nombreuses zones de tension co-localisées au niveau de zones de transition écologique comme les écotones

(Bierne *et al.,* 2011). Le flux de gènes au niveau des zones de tension affecte les différentes régions du génome de façon hétérogène. On qualifie ainsi ces barrières génétiques de barrières semi-perméables car elles agissent comme un filtre sélectif au flux génique (Harrison, 1993). Le concept de barrière semi-perméable prédit en effet des comportements différents pour trois types de locus à travers la zone d'hybridation : les locus contre-sélectionnés, les locus neutres et les locus uniformément favorables. Les locus impliqués dans l'isolement reproductif présentent des combinaisons alléliques défavorables chez les individus hybrides, qui réduisent d'autant plus les échanges génétiques entre espèces qu'ils sont fortement contre-sélectionnés. On parle ainsi de gènes barrières au flux génique ou de gènes verrous. Les échanges aux locus neutres ne sont pas tout à fait indépendants de l'existence des gènes verrous, dont l'effet de barrière au flux génique s'étend aux marqueurs voisins en raison de la faible fréquence des événements de recombinaison. Ainsi, le flux génique aux marqueurs neutres est d'autant plus ralenti qu'ils sont fortement liés à des gènes verrous (Barton & Bengtsson, 1986; Feder & Nosil, 2010). A proximité des gènes verrous, le flux génique est donc fortement restreint voire nul, y compris pour des locus neutres, et au fur et à mesure que l'on s'en éloigne, le flux génique augmente vers son niveau génomique basal (Figure 3) qui dépend de la démographie des espèces en interaction. Sous certaines conditions, la variance de différenciation entre locus peut donc devenir très grande avec des régions complètement hermétiques aux échanges génétiques et des régions qui introgressent librement tout autour. Une telle variance favorise donc la détection des gènes localisés dans le voisinage chromosomique des locus d'isolement. Lorsque les gènes s'isolement sont très nombreux, leurs effets sélectifs cumulés peuvent suffire à réduire efficacement le flux génique sur l'ensemble du génome (Barton & Bengtsson, 1986), on parle alors de « genome hitchhiking » (Feder *et al.,* 2012) ou de « génome congelé ». Dans ce type de situation, seuls les locus soumis à une sélection positive indépendante de l'isolement reproductif (Pialek & Barton, 1997; Bierne *et al.,* 2002) parviennent à traverser la zone

d'hybridation (Barton & Bengtsson, 1986). Cet effet peut se produire quand un allèle issu du fond génétique d'une espèce A s'avère avantageux dans le fond génétique d'une espèce B qu'il va envahir via l'intermédiaire de la formation de génotypes hybrides. La catégorie des locus qui introgressent adaptativement parviennent donc à passer le filtre sélectif des barrières semi-perméables quel que soit le degré d'isolement reproductif tant que celui-ci n'est pas complet.

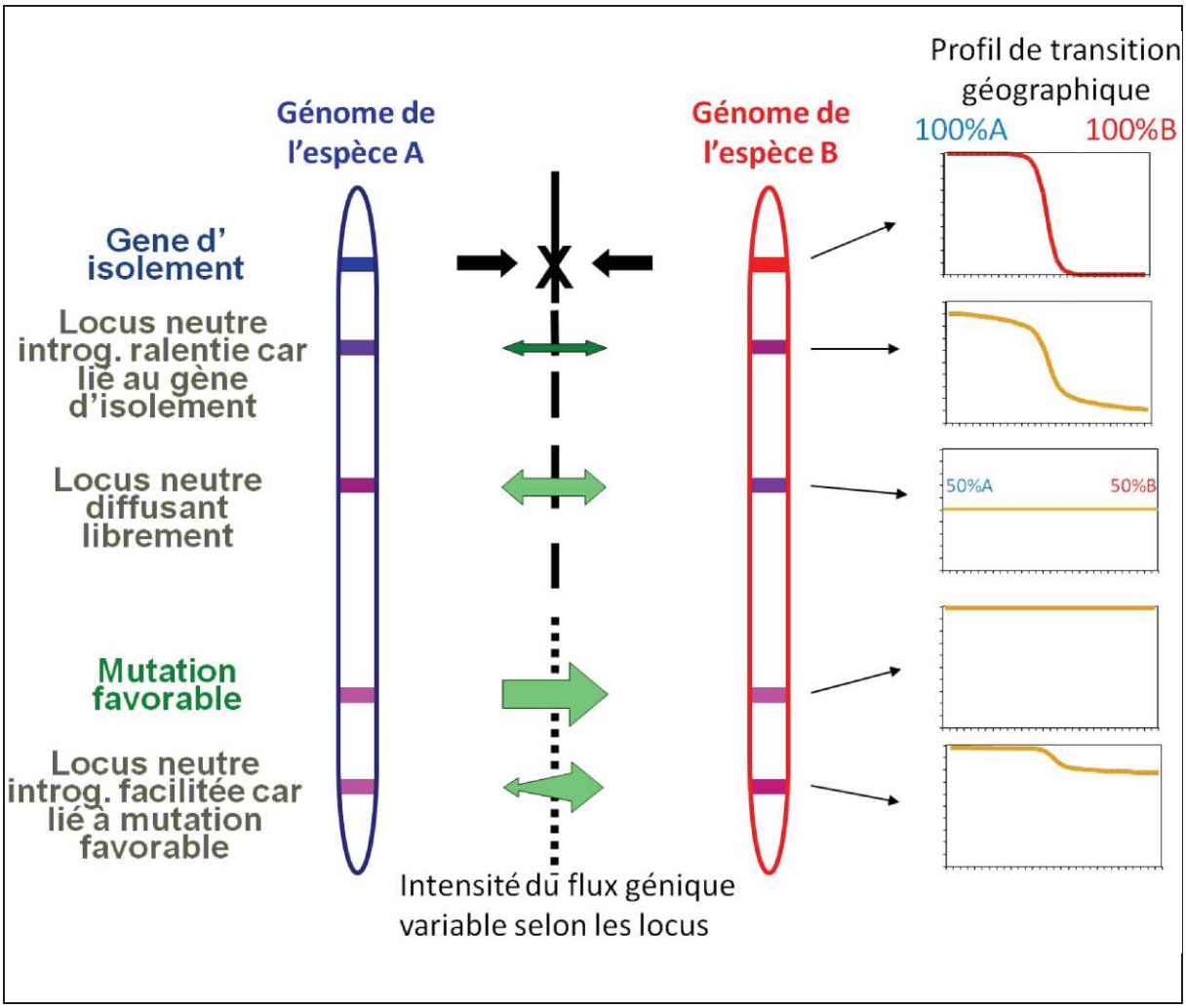


Figure 3 : Détail d'une barrière semi perméable.

2.2. Inférence démo-génétique de l'histoire du flux génique

Reconstituer l'histoire de la divergence entre espèces est une composante importante de l'étude de la spéciation, afin de déterminer quand les populations divergentes ont cessé et éventuellement recommencé à échanger des gènes. Cependant il n'est pas toujours facile d'entreprendre une telle démarche rétrospective, surtout chez de vraies espèces différenciées qui n'interagissent plus via des échanges génétiques. En effet, au cours du temps les populations peuvent être soumises à une variété de processus démographiques et migratoires. Des alternances peuvent avoir lieu entre des périodes où le flux génique est interrompu (périodes d'isolement) et d'autres périodes où les échanges génétiques entre populations sont rétablis (Feder *et al.,* 2012; Abbott *et al.,* 2013). Ces événements laissent néanmoins une empreinte sur le polymorphisme des génomes. Ainsi les données polymorphisme recueillies sur grand nombre de locus au sein d'une paire de populations divergentes contient des informations pouvant nous renseigner aussi bien sur les aspects temporels et démographiques de l'histoire de leur divergence. Ainsi, en comparant des modèles qui sont des représentations simplifiées mais intégrants les principales caractéristiques du processus de divergence, il est possible d'évaluer statistiquement la vraisemblance de différents scénarios alternatifs de spéciation.

Le premier modèle de divergence démographique à avoir été utilisé est celui représentant l'isolement strict entre deux populations divergentes issues d'une même population ancestrale (Figure 4a) (Wakeley & Hey, 1997). En 2001, une nouvelle méthode s'appuyant sur l'information généalogique contenue dans les données haplotypiques qui renseignent en particulier sur les patrons de coalescence a été développé par Nielsen et Wakeley (Nielsen & Wakeley, 2001) dans le but de distinguer entre un modèle d'isolement strict et un modèle d'isolement avec migration continue (modèle IM) entre deux populations divergentes (Figure 4b). L'avancée importante procurée par ce type de modèle est de permettre de distinguer les effets de la migration de ceux liés au tri incomplet du polymorphisme ancestral, essentiellement

sous l'influence de la taille efficace. Le modèle IM permet donc de décomposer le flux génique en deux composantes en prenant en compte séparément les effets de la dérive et de la migration. L'implémentation de méthodes statistiques permettant d'ajuster le modèle IM à des données dans les programmes IM et IMA (Hey & Nielsen, 2007) a contribué à son application pour étudier la spéciation chez de nombreuses espèces, comme la drosophile (Herrig *et al.,* 2014) ou encore le gobe-mouche *Ficedula* (Backstrom *et al.,* 2013).

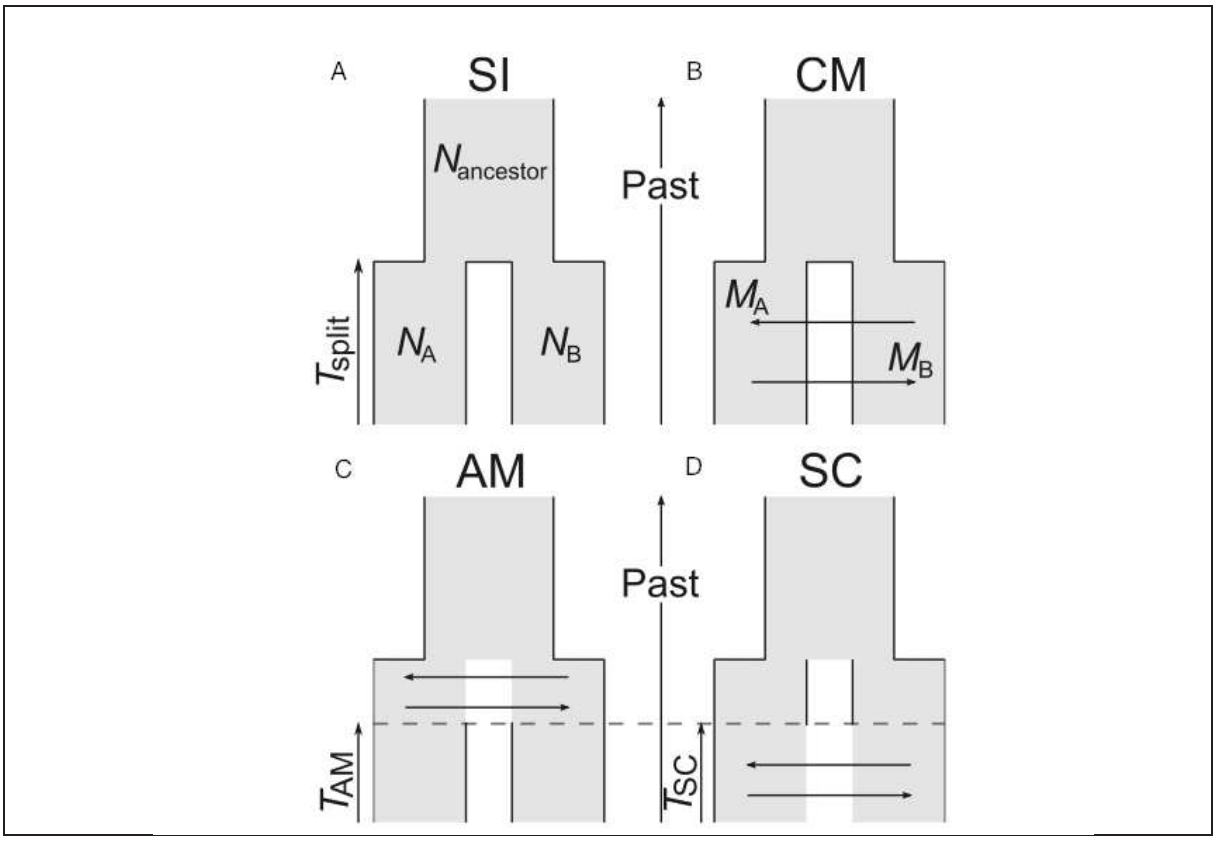


Figure 4 : Scénarios de spéciation. Quatre scénarios avec des patrons temporels de migration différents sont comparés : (A) SI, strict isolement ; (B) CM, migration continue au cours de la divergence ; (C) AM, migration ancienne et (D) SC, contact secondaire. Tsplit est le nombre de générations écoulées depuis la spéciation ; Nancestor, NA, et NB sont les tailles efficaces des populations ; TAM est le nombre de générations depuis que les deux espèces ont arrêté d'échanger des migrants ; et TSC est le nombre de générations depuis que les deux espèces échangent des migrants, après avoir été strictement isolées. Les taux de migration efficaces MA et MB sont exprimés en unités 4Nm, où M est la proportion de la population composée de migrants venant de l'autre population à chaque génération (d'après Roux *et al.,* 2013).

Avec le développement des nouvelles méthodes de séquençage à haut débit permettant l'acquisition de jeux de données génomiques, les méthodes classiques de coalescence développées pour l'analyse de quelques gènes sont inadaptées sur le plan computationnel. De plus, certaines de ces méthodes qui se basent sur les généalogies des gènes nécessiteraient des données haplotypiques phasées qui restent difficiles à obtenir à l'échelle génomique. Les nouvelles techniques moléculaires de génotypage par séquençage telles que le séquençage de fragments RAD (Rad sequencing) produisent des milliers de marqueurs SNPs (Polymorphisme Nucléotidique Simple) aléatoirement échantillonnés dans le génome, et le plus souvent avec une densité insuffisante pour reconstituer localement des généalogies. Pour exploiter ces nouvelles données, deux types de méthodes d'inférence ont été développées. La première appelée méthode ABC pour Aproximate Bayesian Computation (Beaumont *et al.,* 2002) compare un ensemble de statistiques résumant des jeux de données obtenus par simulations avec celles des données observées. Cette méthode est particulièrement appropriée pour la comparaison de modèles complexes dont la vraisemblance est incalculable numériquement et doit être approximée par simulation. La seconde méthode se base sur des équations de diffusion pour décrire analytiquement le polymorphisme sous un modèle de divergence donné. Dans les faits, c'est spectre joint des fréquences allèliques (ou JAFS pour joint site frequency spectrum) qui est utilisé pour décrire la façon dont le polymorphisme se réparti au sein et entre les populations, en représentant le nombre de copies d'allèles dérivés en fonction de leur fréquence dans chacune des deux populations. Cette méthode, proposée et implémentée dans le programme δaδi par Gutenkunst *et al.,*(2009), permet donc d'ajuster des spectres joints théoriques à celui observé en maximisant la vraisemblance d'un modèle de divergence démographique. De plus, sa flexibilité permet le calcul de la vraisemblance de modèles de divergence plus complexes que le modèle IM (Tine *et al.,* 2014; Le Moan *et al.,* 2016). Grace à ces avancées méthodologiques, de nouveaux scenarios de divergence se rapprochant encore

plus de la réalité peuvent être comparés à ceux d'un isolement strict et d'une divergence avec migration continue. Ainsi, des situations où une partie de la divergence entre populations se fait en présence de flux génique et l'autre en allopatrie peuvent être évaluées, comme la migration ancienne (le flux génique post divergence suivi d'un isolement allopatrique) (Figure 4c) ou le contact secondaire (isolement allopatrique suivi par une reprise des échanges génétiques) (Figure 4d).

Tous ces progrès en matière d'acquisition de données génomiques et d'analyses statistiques permettent à présent de confronter les inférences obtenues sous des modèles représentant différents modes de divergence avec les données observées (Sousa & Hey, 2013). Ainsi, en comparant les scenarios de divergence les plus probables et en particulier le moment où ils font intervenir de possibles flux de gène, il est possible d'évaluer l'importance relative de la spéciation allopatrique vs para/sympatrique dans la formation de la biodiversité actuelle ainsi que déterminer la nature du flux génique (homogène ou hétérogène) et son impact sur l'architecture génomique de l'isolement. Par exemple, les travaux basés sur une étude des transcriptomes complets des ciones *Ciona robusta* et *Ciona intestinalis* ont montré que le modèle d'un contact secondaire avec une migration hétérogène était le plus probable pour expliquer la distribution du polymorphisme partagé entre les individus des deux espèces (Roux *et al.,* 2013).

3 Génomique de la spéciation

Parallèlement aux études d'inférence de l'histoire démo-génétique de la spéciation, de nombreuses études se sont focalisées sur la description des patrons (ou paysages) génomiques de différenciation entre espèces ou populations en cours de spéciation. Ces études ont mis en évidence une forte hétérogénéité de la différentiation le long des génomes (Harr, 2006; Ellegren *et al.,* 2012; Parchman *et al.,* 2013; Gagnaire *et al.,* 2013; Tine *et al.,* 2014). Cette architecture

de la divergence génétique est en outre assez variable d'un modèle biologique à l'autre, et peut correspondre à des zones de différenciation étroites, nombreuses et fortement dispersées comme c'est le cas chez le manakin (Parchman *et al.,* 2013) ou le gobe-mouche (Ellegren *et al.,* 2012) ou larges et discrètes comme c'est le cas chez le papillon *Heliconius* (Martin *et al.,* 2013) ou encore la souris *Mus musculus* (Harr, 2006). Dans ce dernier cas, on qualifie ces régions différenciées « d'ilots génomiques de différenciation » (on trouve parfois également le terme « ilots de spéciation ») par comparaison métaphorique à un paysage d'îles montagneuses au milieu de l'océan. Pour expliquer la présence de tels patrons, trois mécanismes majeurs ont été mis en avant dans la littérature.

Le premier mécanisme permettant d'expliquer la formation des îlots de différenciation met l'accent sur l'effet relatif de la sélection et du flux génique dans un contexte de divergence écologique plutôt spatiale (Nosil *et al.,* 2009). Ainsi il considère que les îlots se construisent progressivement lors d'un processus de divergence primaire, en conséquence d'une augmentation du nombre de locus participant à l'adaptation locale. Lors de la première phase de ce phénomène, la sélection disruptive agirait sur un petit nombre de gènes impliqués dans l'adaptation locale et augmentant ainsi le niveau de différenciation entre populations uniquement dans les régions génomiques proches de ces gènes sélectionnés (Feder *et al.,* 2012). Suite à cette étape ces régions moins brassées par le flux génique pourraient faciliter l'établissement de nouvelles mutations qui, si elles sont localement avantageuses, augmenteraient la taille des îlots de différentiation (Smadja *et al.,* 2008; Via & West, 2008). Ce mécanisme d'accumulation de nouveaux gènes d'adaptation induisant une réduction locale de la migration efficace autour des îlots est appelé « divergence hitchhiking ». Enfin, avec l'augmentation du nombre de gènes sélectionnés, le cumul des effets sélectifs aboutirait à réduire la migration efficace sur l'ensemble du génome. Cette dernière phase est appelée « genome hitchhiking » (Feder *et al.,* 2012).

Par contraste avec le mécanisme précédent, un deuxième processus mis en avant par Bierne et collaborateurs considère que la formation des îlots résulterait de l'érosion différentielle de la divergence préexistante lors d'un contact secondaire entre populations différenciées (Bierne *et al.,* 2013). En effet, lors de la divergence en allopatrie des incompatibilités génétiques ou des mutations avantageuses peuvent être fixées (Bierne *et al.,* 2011). Lors de la remise en contact des populations et la reprise du flux génique, la différenciation précédemment accumulée serait alors érodée de façon hétérogène à cause des locus responsables de l'isolement reproductif ainsi que l'adaptation locale qui vont résister à l'introgression et réduire le flux génique localement dans leur voisinage chromosomique (Barton & Hewitt, 1985; Barton & Bengtsson, 1986). Par conséquent plus le taux de recombinaison serait faible localement plus l'effet barrière serait étendu, produisant une érosion lente de la différenciation autour des îlots.

Enfin la dernière hypothèse plus récemment proposée par Cruikshank & Han (2014) se distingue des deux premières par le fait qu'elle minimise l'effet du flux génique hétérogène pour expliquer la formation des îlots génomiques de différenciation. En effet grâce à la réanalyse de données de différenciation de plusieurs paires d'espèces, ces auteurs ont pu montrer que les régions à fort FST (forte différenciation) peuvent être le résultat de l'absence totale du flux génique lors d'une période de divergence allopatrique combiné à certains processus comme les balayages sélectifs (Smith & Haigh, 1974) et la sélection purificatrice (Charlesworth *et al.,* 1993). En accélérant localement le tri du polymorphisme ancestral partagé, ces effets sélectifs ont pour effet de réduire localement la diversité génétique des régions neutres voisines. Ainsi, une diminution locale du polymorphisme peut être due à l'effet d'autostop lors de la fixation d'une mutation avantageuse ou encore le résultat d'une sélection négative qui éliminerait les allèles neutres liés aux mutations désavantageuses, et ce indépendamment de tout mécanisme d'isolement reproductif. Par conséquent, ces effets indirects de la sélection sur les sites neutres voisins réduisent d'autant plus efficacement la diversité génétique intra-

populationnelle que la recombinaison est faible. Ainsi, ils contribuent à augmenter localement les valeurs de l'indice de différenciation génétique FST, qui est négativement corrélé à la diversité intra-populationnelle, particulièrement dans les régions à faible taux de recombinaison (Rieseberg, 2001; Noor & Bennett, 2009; Nachman & Payseur, 2012). Les diminutions du niveau de diversité intra-populationnelle local sont impliquées donc de la même façon que le flux génique hétérogène dans l'architecture du paysage génomique de la différenciation. Il est donc très difficile lorsqu'on détecte un îlot génomique de différenciation, de déterminer s'il résulte d'un processus de divergence primaire, d'une introgression différentielle, ou d'un tri de lignées localement accéléré. Une approche combinant des méthodes d'inférence de l'histoire du flux génique avec une description du paysage génomique de différenciation est donc souvent nécessaire pour comprendre l'histoire évolutive de la divergence entre espèces proches.

4 Originalité des organismes marins et modèle d'étude

La plupart des poissons marins ont un cycle de vie à phase larvaire planctonique (Gyllensten, 1985; Palumbi, 1994; Ward *et al.,* 1994; Waples, 1998) dont les caractéristiques impliquent d'importantes conséquences vis-à-vis de la génétique des populations :

- Une fécondation externe synonyme d'union au hasard des gamètes c'est-à-dire la panmixie à l'intérieur des populations.

- Une phase larvaire planctonique qui autorise des taux de migration (m) importants ainsi que des effectifs de populations (N) très grands ; on s'attend donc à des flux géniques (Nm) forts qui assureraient l'homogénéité génétique sur de longues distances.

- Une forte fécondité qui autorise un fort différentiel de sélection, et qui couplé à des effectifs importants augmente l'efficacité de la sélection même pour des coefficients faibles. La combinaison entre migration forte et efficacité de la sélection offre des conditions favorables pour observer les processus de sélection mis en œuvre.

Alors que ces caractéristiques sont censées favoriser le brassage génétique intra- et inter-populationnel, certaines études mettent à l'épreuve les idées émises a priori au vu du cycle de vie. Par exemple, la forte variance du succès reproducteur chez ces organismes ainsi qu'une mortalité importante lors des stades précoces font que la taille des populations pourrait être, à petite échelle et transitoirement, plus faible que prévue (Hedgecock & Pudovkin, 2011).

Ceci est d'autant plus marqué chez les Pleuronectiformes qui semblent avoir une structure génétique populationnelle plus prononcée que ce qui est prédite et qui serait principalement dû à leur cycle de vie et à leur écologie benthique. En effet, les Pleuronectiformes sont généralement confinés à un substrat particulier et semblent avoir des habitats plus fragmentés que ceux des autres espèces de poissons pélagiques (Quéro *et al.,* 1986). De plus que leurs capacités natatoires sont relativement limitées et leur attachement à un substrat et à un habitat bien spécifiques rendent leurs mouvements et leur migration plus difficiles (Whitehead *et al.,* 1986). Par ailleurs, ils utilisent les estuaires et les lagunes en tant que nourriceries, or ces milieux constituent, par nature, un habitat fragmenté fortement impacté par les activités anthropiques et les facteurs naturels (Boutier *et al* 2000 ; Roessig *et al* 2004). Ainsi, dans de précédentes études qui se sont intéressées à la génétique de quelques espèces de cet ordre de poissons plats ont révélé une nette différenciation génétique chez les soles (Kotoulas *et al.,* 1995; Rolland *et al.,* 2007) et le turbot (Nielsen *et al.,* 2004).

Dans la présente étude, nous nous sommes intéressés aux soléidés et plus précisément au genre Solea. Actuellement, ce genre comprend trois espèces (Desoutter Meniger, 1997). Il s'agit de *Solea solea* Linnaeus 1758, *Solea aegyptiaca* Chabanaud 1927 et *Solea senegalensis* Kaup 1858. La systématique du genre Solea a longtemps posé des problèmes et a été sujette à plusieurs révisions (Desoutter Meniger, 1997). En effet, sept espèces du genre *Solea* étaient encore récemment reconnues en Méditerranée : *S. aegyptiaca* ; *S. impar* Bennett 1831 ; *S. kleini* Bonaparte 1833 ; *S. lascaris* (Risso 1810) ; *S. nasuta* (Pallas, 1811) ; *S. senegalensis* et *S. solea*

(Quéro *et al.,* 1986). Ces distinctions se basaient sur des critères morphologiques parfois ambigus, comme la membrane joignant la nageoire caudale aux nageoires anale et dorsale (facilement altérable lors de la manipulation des poissons), la forme et la position de la tâche noire sur la nageoire pectorale, le nombre des rayons au niveau de la nageoire dorsale ou encore le nombre de vertèbres (Quéro *et al.,* 1986).

Ces critères sont souvent difficiles à évaluer et rendent l'identification des espèces délicate. Cependant, au sein même des trois espèces qui actuellement composent le genre Solea, il a été mis en évidence la présence d'un phénomène hybridation interspécifique entre *S. senegalensis* et *S. aegyptiaca* (She *et al.,* 1987; Ouanes *et al.,* 2011). En effet, les premiers travaux réalisés sur ces espèces le long des côtes septentrionales tunisiennes ont montré l'existence d'une hybridation entre *S. senegalensis* et *S. aegyptiaca* (She *et al.,* 1987). Ces travaux ont été repris par Ouanes *et al* en 2011 afin de délimiter cette zone hybride et d'évaluer ses évolutions potentielles qui auraient pu survenir au cours des dernières années (Ouanes *et al.,* 2011). Ce phénomène semble être limité géographiquement à une aire très proche de la limite supposée des aires de distribution des deux espèces : la lagune de Bizerte. Cette lagune semble être le principal habitat dans lequel des hybrides sont rencontrés et ce constat est soutenu par un indice hybride moyen élevé observé en 2011 (41,2%), et qui représente le triple de la valeur mesurée pour ce même milieu quelques années auparavant par She *et al* (1987). De cette observation est née l'hypothèse que les échanges génétiques entre les deux taxons sont en train d'évoluer.

5 Objectifs de la thèse

Dans la continuité des précédentes études menées sur la zone hybride des soles, nous nous somme proposé à travers ce travail de caractériser les échanges génétiques entre *Solea senegalensis* et *Solea aegyptiaca* ainsi que leurs signatures à travers cette zone hybride. Ainsi

au cours de cette thèse nous nous sommes focalisé sur deux principaux volets qui touchent à différents aspects de l'hybridation entre ces deux espèces.

Le premier s'intéresse aux relations entre génotype et phénotype en combinant approche morphométrique et génétique afin de mettre en évidence l'influence de l'hybridation et de l'introgression sur la variabilité morphologique.

Dans le deuxième volet nous aborderons la nature du contact entre ces deux espèces en faisant appel à une méthode d'inférence de modèles démo-génétiques (δaδi) et nous nous intéresserons aussi aux variations d'introgression entre locus à l'échelle géographique et génomique en adoptant une approche qui se base sur les méthodes de clines (clines géographiques et clines génomiques). Ainsi dans cette seconde partie nous allons essayer de répondre aux questions liées à l'histoire du contact entre ces deux espèces par exemple :

- S'agit-il d'un contact récent ou ancien ?
- Qu'elle serait le type de la zone hybride mise en place suite à ce contact ?
- Qu'elle est la fraction du génome qui est congelée (incompatibilités génomiques) et celle qui est perméable aux flux géniques ?
- Est-ce que la zone hybride s'est stabilisée ou elle est encore en mouvement ?
- Existe-il de l'asymétrie dans les flux géniques entre ces deux espèces ?

Chapitre I : Hybridation introgressive et transgression morphologique dans la zone de contact de deux espèces méditerranéennes du genre *Solea*

1 Introduction

Dans le contexte d'un contact secondaire, l'hybridation introgressive peut conduire à la convergence des phénotypes parentaux et l'observation de phénotype intermédiaire si les gènes impliqués sont neutres et introgressent facilement. Cependant, des phénotypes d'hybrides introgressés qui présentent un excès par rapport à la variation phénotypique parentale ont été reportés au niveau des zones hybrides. Ces observations s'expliquent par le fait qu'au niveau de ces zones, de nouvelles combinaisons génotypiques et de nouveaux phénotypes peuvent être produits. L'étude des remaniements génotypiques ainsi que les variations phénotypiques dans ces zones est donc d'un intérêt particulier afin de documenter la mise en place de l'isolement reproductif, mais aussi l'émergence de nouveautés évolutives. Dans cette première partie de la thèse et dans le but de contribuer à la compréhension des contraintes évolutives qui limiteraient le flux génique au niveau des barrières interspécifiques nous nous sommes proposés d'explorer les patrons de la variation morphologique associée à l'hybridation introgressive entre deux espèces de la famille des Soléidés, *Solea senegalensis* et *Solea aegyptiaca*. Nous nous sommes intéressés également à l'analyse de la co-segregation des variations génétiques et de la forme du corps. En effet les déviations des patrons de ségrégation pourraient refléter d'éventuel mécanismes sélectifs contre des combinaisons allèliques dans les génotypes introgressés quant aux variations de la forme du corps chez les hybrides pourrait être considéré comme un des moyens pour évaluer les conséquences du flux génique sur la fitness. La relation entre la

composition génétique au niveau de quatre locus nucléaires et la variation individuelle de la forme du corps a été étudiée dans quatre populations échantillonnées à travers la zone hybride située au nord de la Tunisie. L'analyse a révélé une forte corrélation, entre la variation génétique et phénotypique à l'échelle inter-populationnelle Cependant cette corrélation n'est plus présente lorsqu'elle est étudiée entre individus d'une même population. Une convergence morphologique entre les espèces parentales a été observée à proximité de la zone de contact. Cependant, les échantillons les plus proches de la zone hybride affichent en plus une déviation de la ségrégation génotypique et phénotypique ainsi qu'une transgression phénotypique. Dans ces échantillons, la déviation de la variation de la forme du corps pourrait être attribuée, en partie, à une réduction de l'indice de condition mais aussi à une distorsion de la composition génétique due probablement à la disparition de certaines combinaisons allèliques. Ces résultats sont interprétés comme étant une indication d'une dépression hybride, qui contribuerait à l'isolement reproductif entre les deux espèces.

Cette partie a fait l'objet d'un article qui a été publié dans Ecologie and Evolution.

2 Article: Introgressive hybridization and morphological transgression in the contact zone between two Mediterranean *Solea* species

Abstract

Hybrid zones provide natural experiments where new combinations of genotypes and phenotypes are produced. Studying the reshuffling of genotypes and remolding of phenotypes in these zones is of particular interest to document the building of reproductive isolation and the possible emergence of transgressive phenotypes that can be a source of evolutionary novelties. Here, we specifically investigate the morphological variation patterns associated with introgressive hybridization between two species of sole, Solea senegalensis and Solea aegyptiaca. The relationship between genetic composition at nuclear loci and individual body shape variation was studied in four populations sampled across the hybrid zone located in northern Tunisia. A strong correlation between genetic and phenotypic variation was observed among all individuals but not within populations, including the two most admixed ones. Morphological convergence between parental species was observed close to the contact zone. Nevertheless, the samples taken closest to the hybrid zone also displayed deviant segregation of genotypes and phenotypes, as well as transgressive phenotypes. In these samples, deviant body shape variation could be partly attributed to a reduced condition index, and the distorted genetic composition was most likely due to missing allelic combinations. These results were interpreted as an indication of hybrid breakdown, which likely contributes to post-mating reproductive isolation between the two species.

Keywords: *Solea senegalensis*, *Solea aegyptiaca*, introgressive hybridization, body shape, genetic-phenotypic correlation, transgressive phenotypes

2.1. Introduction

Secondary contacts between closely related species, resulting either from natural processes or anthropogenic activity, often lead to the formation of hybrid zones in which populations with divergent genomes have the potential to exchange genes. These zones have been shown to act as semi-permeable barriers to gene flow, that selectively filter introgression by preventing genomic regions involved in reproductive isolation to be exchanged among species (Barton & Hewitt, 1985; Barton & Gale, 1993; Harrison, 1993; Harrison & Larson, 2014). Depending on the traits considered, introgressive hybridization may led to morphological convergence of parental populations if the underlying divergent genes behave neutrally and readily introgress upon secondary contact (Grant & Grant, 2002). A simple dilution of the parental phenotypic differences whereby hybrids display intermediate phenotypes compared to the two-parental species is often observed in the populations located near the center of a hybrid zone (Mayr, 1963). Indeed, additive traits determined by quantitative traits loci (QTLs) with directional effects (with each QTL having an effect in one parental population but not in the other) should appear intermediate in hybrids compared to their parental populations (Albertson & Kocher, 2005; Rieseberg & Willis, 2007). Nevertheless, transgressive hybrid phenotypes that exceed the range of parental phenotypic variation have also been reported, particularly in the context of hybrid zones (Rieseberg *et al.,* 1999a; Rieseberg *et al.,* 2003). This phenomenon, referred to as transgressive segregation (Rieseberg *et al.,* 1999a ; Bell & Travis, 2005; Stelkens *et al.,* 2009) may be attributed on the contrary to traits encoded by QTLs with antagonistic effects in each parental population (Rieseberg *et al.,* 1999a). Alternatively, morphological differences between species can persist despite genetic introgression if they are themselves involved in reproductive isolation, or if the genes that

control them are tightly linked with genomic regions involved in reproductive isolation (Harrison & Larson, 2014). In this case, a link between morphological trait variation and individual fitness is expected (Arnold & Hodges, 1995). Some incompatible genotypic combinations may be eliminated, thereby preventing hybrids to reach the full range of phenotypic variation that could be potentially generated by recombining the parental genomes. Disentangling the role of the different mechanisms which can be involved in producing morphological variation in hybrids remains fairly poorly studied in many taxa. Yet, it remains an important question for the study of hybrid zones (Gay *et al.,* 2008), since it may help reveal the underlying architecture of reproductive isolation (Rieseberg *et al.,* 1999b) as well as the origin of evolutionary novelties (Parsons *et al.,* 2011; Nichols *et al.,* 2015). Here, as a first step toward understanding the evolutionary constraints limiting gene flow across a species boundary, we explore body shape variation patterns along a natural hybrid zone transect between two species of sole, *Solea senegalensis* and *Solea aegyptiaca*. These two taxa, which geographical distributions partially overlap in the Mediterranean, are recognized as distinct sister species (Borsa & Quignard, 2001; Vachon *et al.,* 2008). Along the North African coast, *S. senegalensis* ranges from Senegal to Tunisia, and *S. aegyptiaca* occurs from Tunisia to Egypt. Few ambiguous morphological features have been described to distinguish them (Quéro *et al.,* 1986), and thus these two taxa can be considered as cryptic species. Hence, the description of their spatial distributions relies primarily on genetics (She *et al.,* 1987; Ouanes *et al.,* 2011). The two-species come into contact in a 100-km wide zone spreading from Bizerte in the north to Cap Bon in the south. This zone encompasses the Gulf of Tunis and the Bizerte Lagoon. Introgressive hybridisation occurs and was evidenced using a few genetic markers in previous studies based on allozymes (She *et al.,* 1987) and intron length polymorphisms (Ouanes *et al.,* 2011). This phenomenon predominantly occurs in the large lagoon of Bizerte that appears to be the main habitat in which hybrids are found. It is however unclear whether this zone is stable

or in the process of widening, because the average hybrid index observed in 2008 (41.2%) (Ouanes *et al.,* 2011) increased by three times the value measured in the same place twenty years earlier (16.1%) by (She *et al.,* 1987). In the present study, we propose to assess morphological variation patterns inside this zone and analyse the co-segregation of body shape and genetic variation. We use deviant segregation patterns at genetic markers as a way to capture possible mechanisms of selection against deleterious allelic combinations in introgressed genotypes. We also use body-shape variation in admixed populations to evaluate the consequence of gene flow on the condition of introgressed individuals.

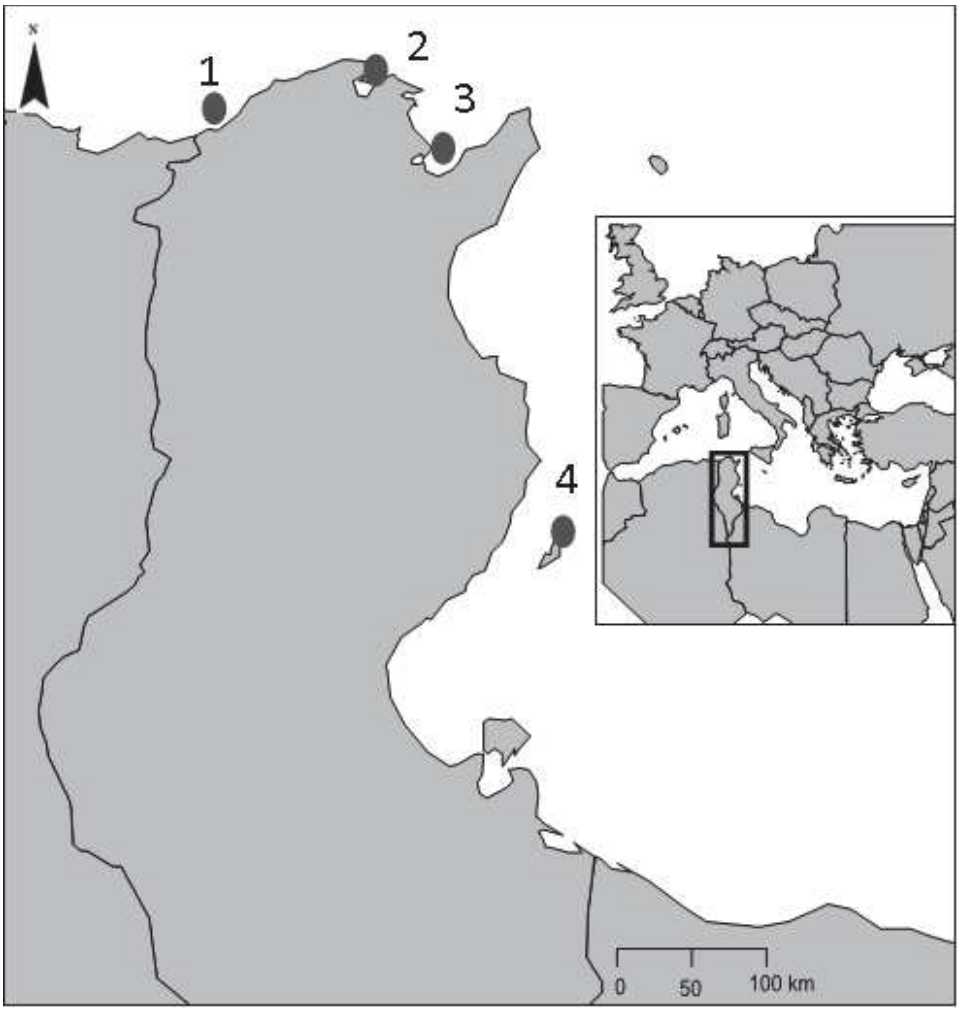


Figure 5: Sampling localities along the Tunisian coast. 1. Tabarka; 2. Bizerte Lagoon; 3. Gulf of Tunis; 4. Kerkennah Island.

2.2. Materials and Methods

a. Sampling

We collected 88 adult individuals of *S. senegalensis* and *S. aegyptiaca* from four localities spanning as much as possible the hybrid zone between the two species along the Tunisian coast: Tabarka (n=8), Bizerte Lagoon (n=47), Gulf of Tunis (n=15) and Kerkennah Islands (n=14) (Figure 5). According to the previous study of (Ouanes *et al.,* 2011), these locations display variable amounts of genetic admixture. Tabarka contains *S. senegalensis* individuals introgressed by ca. 15% of *S. aegyptiaca* alleles. Conversely, on the opposite side the Kerkennah sample contains mostly pure *S. aegyptiaca* individuals. These two locations constitute our peripheral samples. For the two inner samples, Bizerte Lagoon is considered as the nearest to the hybrid zone center and contains *S. senegalensis* genotypes admixed with ca. 41,2% of *S. aegyptiaca* alleles, while Gulf of Tunis contains *S. aegyptiaca* individuals introgressed by ca. 17% of *S. senegalensis* alleles. Individuals were collected from small fishing boats and transported immediately on ice to the lab for morphological and genetic experiments.

b. Genetic analysis

Whole genomic DNA (40ng/µl) was extracted from fin clips using Qiagen Dnaeasy Blood and tissue kit. Four EPIC loci (GH2, Am2B-2, CK6-2 and Met1) that were previously used to distinguish *S. senegalensis* and *S. aegyptiaca* outside the hybrid zone by (Ouanes *et al.,* 2011) were amplified by PCR under the same conditions. PCR products were separated by electrophoresis on a 1% acrylamide gel and individual genotypes were subsequently scored with FMBIO II (Hitachi) using an internal size standard. Genotype scoring was performed twice by two different persons and then checked for consistency. Discriminant Correspondence Analysis (DCA) based on group centroids as implemented in Genetix 4.05.2 (Belkhir, et al. 1996) was used to visualize the partitioning of genetic variation among individuals. This

multivariate analysis is particularly well-suited to describe the genetic composition of individuals in a gradient of admixture, as often encountered in hybrid zones. Genetic differentiation estimated by FST was assessed for each pair of samples and tested using 10 000 permutations in Genetix 4.05.2.

c. Morphometric analysis

Each individual was photographed on the eyed side with a Canon Digit Ixus 95 IS using a 35mm f/2,8 lens with a fixed focal length. All images were digitized in TPSDig 2 (Rohlf 2006) using 21 homolog landmarks to perform geometric morphometrics analysis based on two-dimensional Type I landmarks positioned according to clear homologous anatomical features (Figure 6). In order to remove non-shape variation, we performed a generalized Procrustes analysis using the R package Geomorph 2.1.7 (Adams & Otarola-Castillo, 2013). This transformation consists in standardizing measures to control for differences in individual body size, by translating and rotating the configuration of landmarks to minimize the sum of squared distances between homologue landmarks (Zelditch, 2004). In order to test for remaining allometric effects following Procrustes transformation, we performed a multivariate regression of individual distance to the consensus shape against size using the procD.lm() function in Geomorph. In order to focus on morphological variation between species, we then conducted a canonical discriminant analysis (CDA) to capture the multi-dimensional variation associated with body shape (Albrecht, 1980; Klingenberg *et al.,* 2003; Zelditch, 2004). This method assesses the total amount of variation in body shape among groups of samples, expressed in a n-dimensions space where n is the number of groups minus one. Transformation grids produced with the thin plate spline technique were used to visualize body shape changes among sampled populations. These analyses were performed with the R packages Geomorph (Adams & Otarola-Castillo, 2013) and Morpho (Schlager & Jefferis, 2015).

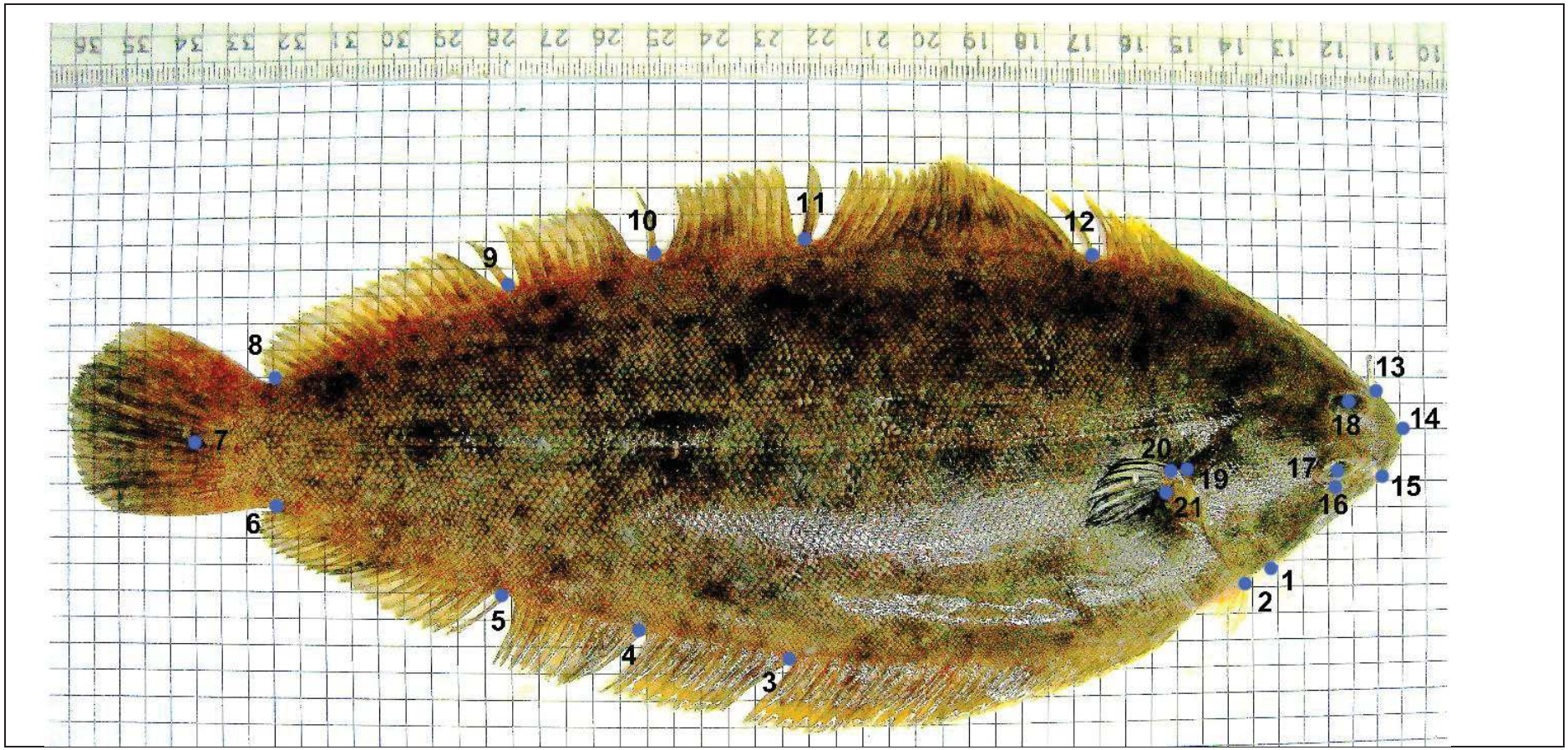


Figure 6: Positions of the digitized Landmarks used for body shape analyses. 1. Insertion of operculum on the ventral profile; 2. Origin of the pelvic fin; 3. Insertion of the 40th anal fin ray (counted from posterior insertion of the anal fin); 4. Insertion of the 30th anal fin ray (counted from posterior insertion of the anal fin); 5. Insertion of the 20th anal fin ray (counted from posterior insertion of the anal fin); 6. Ventral origin of the caudal fin; 7. Caudal end of lateral line; 8. Dorsal origin of the caudal fin; 9. Insertion of the 20th dorsal fin ray (counted from posterior insertion of the dorsal fin); 10. Insertion of the 30th dorsal fin ray (counted from posterior insertion of the dorsal fin); 11. Insertion of the 40th dorsal fin ray (counted from posterior insertion of the dorsal fin); 12. Insertion of the 60th anal fin ray (counted from posterior insertion of the dorsal fin); 13. Anterior origin of the dorsal fin; 14. Snout tip; 15. origin of the mouth (upper jaw); 16. End of the mouth (lower jaw); 17. Center of the left ocular space; 18. Center of the right ocular space; 19. End of operculum; 20. Upper side of the pectoral fin; 21. Lower side of the pectoral fin.

2.3. Results

a. Genetic analyses

As expected from previous studies, discriminant correspondence analysis (DCA) based on genetic data clearly separated the two samples from the periphery of the contact zone on the first axis, which explained most (50.3%) of the variation contained in the dataset. Tabarka specimens were positioned on the positive side of the first axis (*S. senegalensis* side of the hybrid zone) while specimens sampled in Kerkennah Islands were located on the negative part of this axis (*S. aegyptiaca* side of the hybrid zone) (Figure 7a). Fish collected in the inner samples stood in intermediate positions along this axis. Individuals from Bizerte Lagoon were more genetically similar to *S. senegalensis* and Gulf of Tunis samples were more closely related to *S. aegyptiaca* (Table 1). The second axis of the DCA captured 30.7% of the total inertia, and essentially distinguished the Tabarka sample, revealing genetic differentiation between Tabarka and Bizerte samples on the *S. senegalensis* side of the hybrid zone. The third axis explained 19% of total genetic variation, and mostly distinguished the Gulf of Tunis from the other three samples (Figure 7b).

Table 1: Genetic differentiation estimated by FST assessed for each pair of samples. (*** P<0.001).

FST	*Gulf of Tunis*	*Bizerte Lagoon*	*Tabarka*
Kerkennah	*0.358****	*0.498****	*0.732****
Gulf of Tunis		*0.249****	*0.384****
Bizerte Lagoon			*0.328****

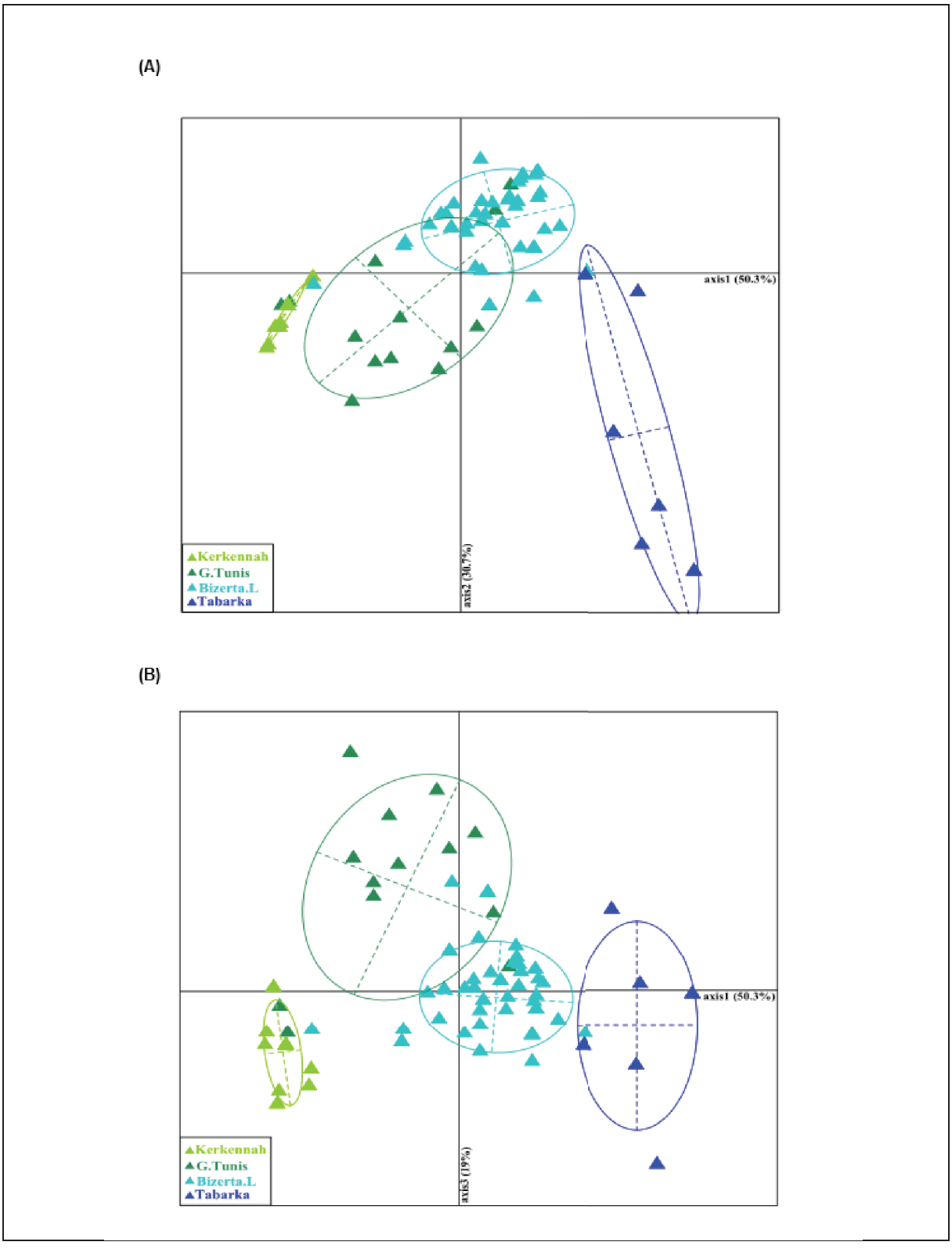


Figure 7: Discriminant correspondence analysis (DCA) on Solea senegalensis and Solea aegyptiaca based on Epic-PCR Intronic markers (Kerkennah Islands, Gulf of Tunis, Bizerte Lagoon and Tabarka). (A) Axis 1 & Axis 2, (B) Axis 1 & Axis3.

b. Geometric morphometrics

Multivariate regression between Procrustes-transformed shape coordinates and size showed no significant association (P = 0.12), indicating the absence of residual allometric effects after Procrustes transformation. The first axis of the CDA explained 60.9 % of body shape variation, and separated the four samples according to their geographic position relative to the center of the hybrid zone (Figure 8). The Tabarka sample located on the *S. senegalensis* side of the hybrid zone was projected in the positive part of axis 1, whereas Kerkennah Island which is located on the *S. aegyptiaca* side of the hybrid zone was projected in the negative part of axis 1. The Gulf of Tunis and Bizerte samples occupied intermediate positions on each side of axis 1. The second and the third axes, which respectively explained 25.9% and 13.2% of shape variation, highlighted more complex patterns. On the one hand, axis 2 separated clearly the two inner samples (Bizerte Lagoon and Gulf of Tunis) from the two peripheral ones (Figure 8a). Hence, along this axis, the inner samples exhibited a morphological variance exceeding that of peripheral (and hence less introgressed) samples, which constitute a transgressive pattern. On the other hand, the third axis principally opposed the Gulf of Tunis and Bizerte samples that were the furthest apart along this axis (Figure 8b). This could be considered as another kind of transgressive pattern whereby the inner samples' morphological variance also exceeds that of peripheral ones, but in opposite directions. The deformation grid obtained from the Procrustes superimposition showed that the principal body shape differences between species were associated with body height, and were more specifically related to a dorso-ventral expansion explained by landmarks 3, 4, 5, 10, 11 and 12. This differentiation follows a gradient along axis 1 in which the *S. senegalensis* individuals from Tabarka exhibit the largest body height while the *S. aegyptiaca* individuals from Kerkennah have the lowest body height (Figure 8c). Deformation grid along axis 2 revealed a deformation component opposing the peripheral from the inner samples, whereby the latter.

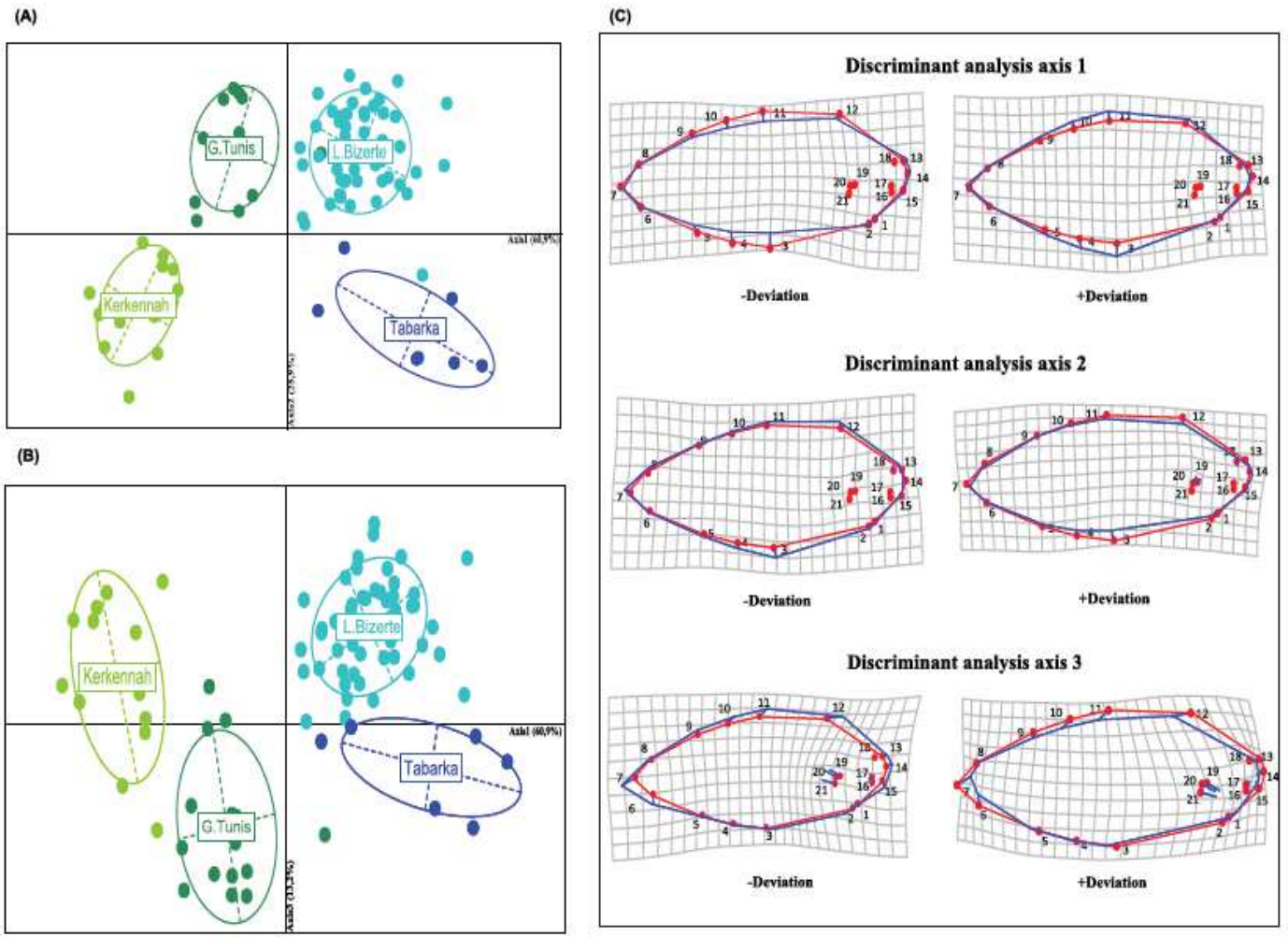


Figure 8: Canonical Discriminant Analysis (CDA) among S. senegalensis and S. aegyptiaca (Kerkennah Islands, Gulf of Tunis, Bizerte Lagoon and Tabarka) based on geometric morphometrics (A) Axis 1 & Axis 2, (B) Axis 1 & Axis3 (C) transformation grids associated to axis 1, axis 2 and 3. These transformation grids show the shape changes from the overall mean associated with each discriminant axis. These transformation grids show as blue hyphens the shape changes from the overall mean (red dots) associated with each discriminant axis. The scale factors for shape changes were set at the program's default values.

have a less convex and more depressed abdomen than the former (landmark 3 and 4). Finally, shape variation along axis 3 was associated with the caudal (landmark 7), pectoral fin (landmark 19, 20 and 21) and head (landmark 14, 15and 16) regions (Figure 8c).

Correlation between morphometry and genetics The Procrustes rotation of the bivariate configurations (i.e. axes 1 and 2) obtained from genetic and morphometric data analysis showed a high correlation between shape and genetic variation ($m2 = 0.67$, $P<0.001$) (Figure 9b). In the new rotated plane, the main axis of morpho-genetic differentiation between species was defined by a line connecting Tabarka and Kerkennah samples (red dashed line, Figure 9a). When projected on this new axis of differentiation between species, the two inner samples of Bizerte Lagoon and Gulf of Tunis occupied intermediate positions, as already observed with morphological and genetic analyses separately (Figure 7a and 8a). Moreover, a second axis of variation perpendicular to the first one (orange dashed line, Figure 9a) revealed a clearly shifted, position of the Bizerte sample, as to a lesser extent for the Gulf of Tunis sample. This significant correlation between genotype and phenotype and the intermediate positions for Bizerte Lagoon and Gulf of Tunis samples were still detected when the groups were not predefined in genetical and morphological analysis ($m^2 = 0.34$, $P<0.001$). Finally, performing the same Procrustes analysis within each sample separately revealed no significant correlation between genotype and phenotype variation (Bizerte Lagoon: $m^2 = 0.18$, $P = 0.8$; Gulf of Tunis: $m^2 = 0.55$, $P = 0.028$) (Figure 10).

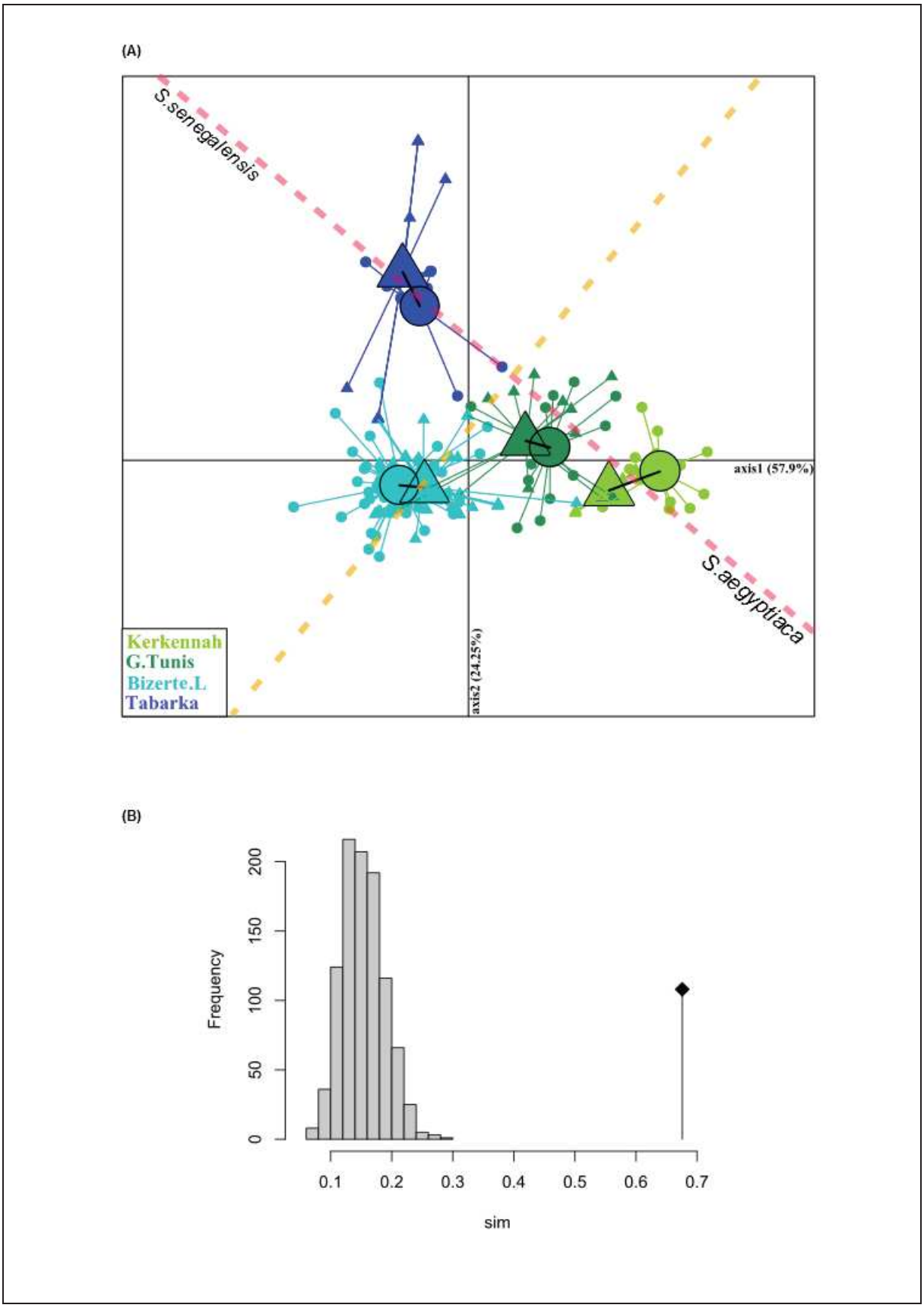


Figure 9: Plot of the genetic and morphometric data after Procruste rotation (A) (genetic data (triangle), morphometric data (circle); Kerkennah Islands, Gulf of Tunis, Bizerte Lagoon and Tabarka). Correlation (m^2) between the genetic and morphometric data after Procruste rotation (B).

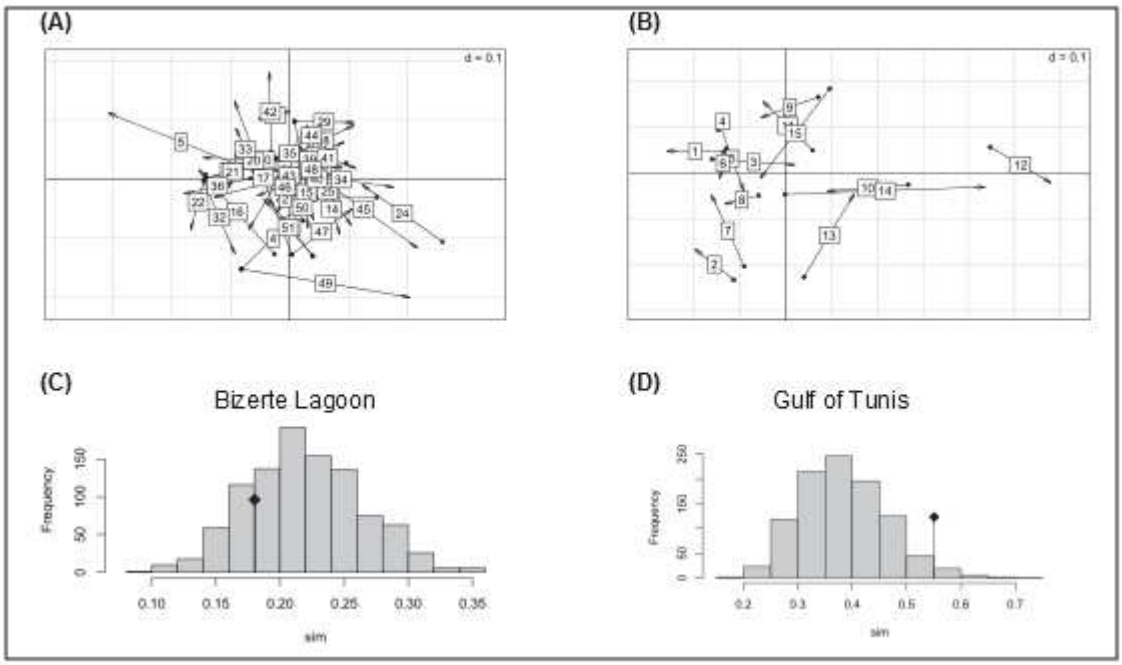


Figure 10: Plot of the genetic and morphometric data after Procruste analysis performed within Bizerte Lagoon samples (A) and within Gulf of Tunis samples (B). The beginning of the arrow is the position of the genotype; the end of the arrow is the position of the phenotype. Correlation (m^2) between the genetic and morphometric data after Procruste rotation performed on Bizerte Lagoon (C) and Gulf of Tunis (D) samples.

2.4. Discussion

Our analysis of genetic variation across the hybrid zone between *S. senegalensis* and *S. aegyptiaca* is in good agreement with previous results by (Ouanes *et al.,* 2011). Here, using new samples, we confirm that Bizerte Lagoon and the adjacent Gulf of Tunis correspond to the area where the strongest admixture signal is found along the Tunisian coast. Moreover, these two locations are distributed on both sides of the hybrid zone, with Bizerte Lagoon being on the *S. senegalensis* side and Gulf of Tunis on the *S. aegyptiaca* side. As for genetic polymorphism, body shape variation also enabled us to discriminate easily between *S. senegalensis* and *S. aegyptiaca* morphotypes outside the zone. Moreover, morphological convergence was observed inside the zone where individuals from inner samples displayed intermediate morphologies compared to samples from parental populations (i.e. in peripheral

samples). The fact that morphological variation closely parallels genetic variation along the first axis in both analyses is expected under a simple dilution of additive genetic effects. Therefore, our data support that at least part of the geographic variation in body shape is explained by the neutral introgression of additive morphological traits along a gradient of admixture. Nevertheless, we found no significant correlation between morphological and genetic variation within samples, including those from the inner part of the hybrid zone. This was expected since most variation occurs between parental populations. Moreover, we only used a few genetic markers which are extremely unlikely to be linked with the loci controlling shape variation. Our analysis reveals a different signal of variation along the second axis of the morphological space. On this axis, the two inner populations are clearly similar to each other while being differentiated from the parental outer samples which are themselves not differentiated along axis 2. This second pattern deviates from the previously mentioned mechanism of morphological convergence since it cannot be explained by additivity of the morphological QTLs alone. Thus, it suggests the existence of more complex epistatic and pleiotropic effects due to hybridisation between divergent genomes. This could be due for instance to a mechanism of hybrid breakdown acting on different types of hybrid pedigree, since the population of Bizerte Lagoon is a *S. senegalensis* population introgressed by *S. aegyptiaca* alleles whereas it is the opposite for the Gulf of Tunis. Phenotypic effects affecting all admixed genotypes in the same direction and contrasting with the absence of such effects in both parental populations could concern some fitness-related trait reflecting the general performance of individuals, such as the condition factor. Interestingly, the deformation grid on the positive part of axis 2 shows a less convex shape of the ventral part in the inner samples which could reflect a lower condition of hybrid individuals compared to those from peripheral samples. In parallel, we also detected a signal of genetic distortion on axis 2 of the genetic correspondence analysis. This suggests the existence of preferential allelic combinations that

cannot simply result from the neutral dilution of the parental genomes in introgressive crosses. Although we only used 4 neutral markers for genotyping, these selective effects can be likely captured due to genome-wide association because of the existence of numerous genetic incompatibilities between highly differentiated (and sometimes coined 'congealed') genomes. The observed genetic distortions possibly reflect selective elimination of certain allelic combinations from the admixed populations, which may due to some form of hybrid breakdown. Such distortions in the wild closely reflect segregation distortions commonly observed in experimental crosses between divergent species (Moyle *et al.,* 2006; Casellas *et al.,* 2012; Gagnaire *et al.,* 2013a). The depressed morphology of introgressed samples along axis 2 of the morphospace may thus constitute a phenotypic translation of this segregation load. Therefore, the morpho-genetic correlation detected along axis 2 may reflect the fact that only hybrids combine at the same time admixed genotypes and lower condition. Finally, a more classical pattern consistent with phenotypic transgression due to QTLs with antagonistic effects is evidenced along axis 3 of the morphospace. This effect is expected to occur in opposite directions on both sides of a hybrid zone as exemplified in Table 2.

In one case the predominant introgression of some aegyptiaca alleles inside the senegalensis background generates a positive transgression in the Bizerte sample and the opposite is true for the Gulf of Tunis sample. To conclude, our study of the relationship between genetic composition and morphological shape highlighted mainly two key findings. First, when considering all four population samples, a significant correlation between genetic and phenotypic data was observed. This is expected because of the global linkage generated inside a tension zone maintained by equilibrium between counter-selection of recombinant genotypes and gene flow from parental species. Second, deviant segregation of genotypes and phenotypes evidenced by transgressive positions of the most introgressed inner samples along axes 2 and 3 of the multivariate analyses could be interpreted both in terms of signature of hybrid breakdown

and recombination of QTLs with antagonistic effects. Altogether, these results call for a genome-wide assessment of the architecture of gene exchange between these two hybridising species of sole. Such approach would be useful to specify what proportion of the genome can still be exchanged neutrally between species, and what are the fitness consequences of interspecific gene flow in recombinant genotypes.

Table 2: Hypothetical examples of transgressive segregation patterns involving six biallelic loci in two parental populations and their hybrid descendants.

	Differentiated traits							
	Directional effects QTL				Antagonistic effects QTL			
	Sp1	Sp2	Introgressed Sp1	Introgressed Sp2	Sp1	Sp2	Introgressed Sp1	Introgressed Sp2
A	+1	0	+1	+1	+1	-1	+1	-1
B	+1	0	+1	0	+1	-1	+1	-1
C	+1	0	0	0	+1	-1	+1	-1
D	0	-1	0	-1	+1	-1	+1	-1
E	0	-1	-1	0	-1	+1	+1	+1
F	0	-1	0	0	-1	+1	-1	-1
∑	+3	-3	+1	0	+2	-2	+4	-4

Chapitre II : Acquisition d'un jeu de données de SNPs haute densité

1. Echantillonnage, Construction des librairies RAD et Séquençage

1.1. Echantillonnage

Afin de suivre la direction et l'intensité du flux génique entre *S. senegalensis* et *S. aegyptiaca* le long des côtes nord-africaines, le choix de la stratégie d'échantillonnage est primordial. Dans cette optique, nous avons pu échantillonner 190 individus qui proviennent de 10 stations différentes réparties le long d'un transect géographique passant par la zone d'hybridation (Figure 11) :

- Trois stations autour de la zone d'hybridation décrite par (Ouanes et al., 2011) : la lagune de Bizerte, le golfe de Tunis, Tabarka.
- Quatre stations plus au moins éloignées de cette zone : Deux stations dans le Golfe de Gabes (îles Kerkennah et lagune d'El Bibane) et deux autres en Algérie la Lagune de Mellah et Annaba
- Trois stations très éloignées de la zone : Cadix (Espagne) et Dakar (Sénégal) à l'ouest et la lagune de Bardawil en Egypte à l'est.

Cette stratégie avec une densité d'échantillonnage plus forte au cœur de la zone d'hybridation a été adoptée car nous nous attendions à un fort taux d'introgression au niveau de la lagune de Bizerte, diminuant au fur et à mesure qu'on s'éloigne de la zone de contact, avec une introgression de gènes type senegalensis dans le pool génique de type aegyptiaca à l'est et

réciproquement à l'ouest. Ceci devrait nous permettre également de vérifier si la zone hybride s'est déplacée par rapport aux précédentes études, comme il l'a été suggéré par Ouanes et al en 2011. A ces deux espèces nous avons séquencé aussi deux individus provenant de deux espèces différentes, *Solea solea* et *Pegusa impar,* afin de pouvoir les utiliser comme groupe externe (ou « outgroup ») dans les méthodes d'inférences démographiques.

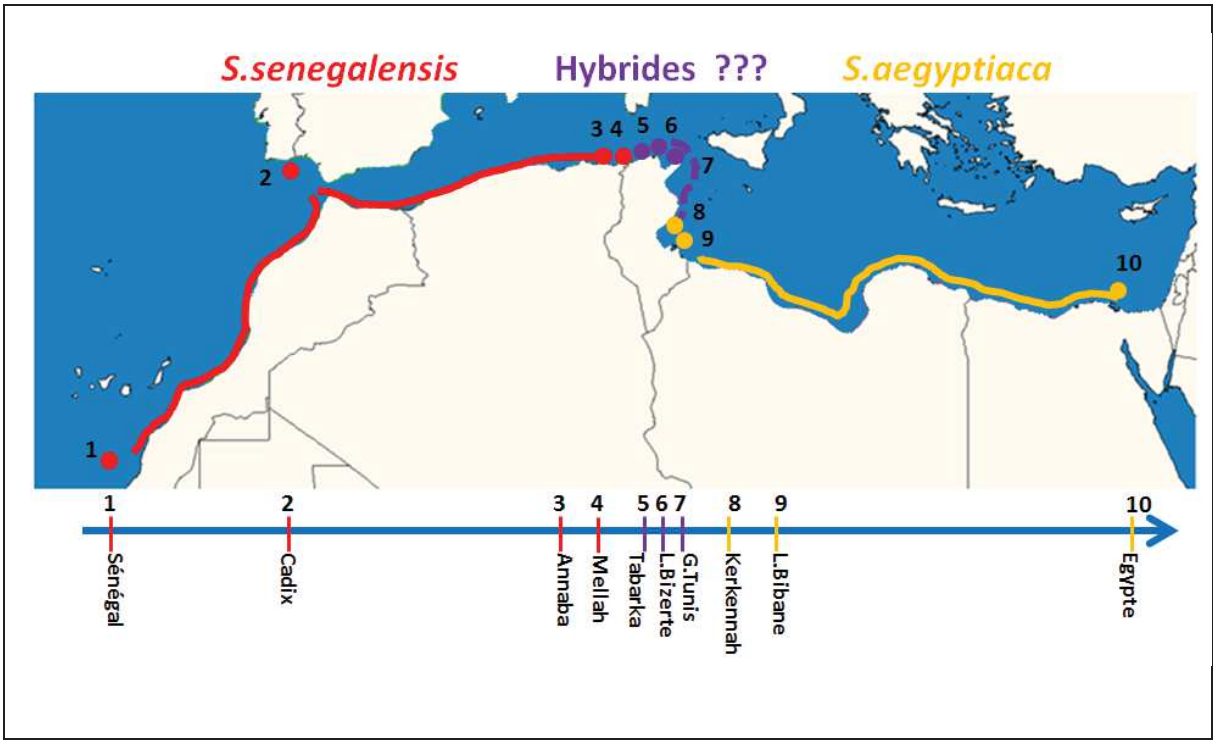


Figure 11 : Sites d'échantillonnage des différentes populations.

1.2. Construction des librairies et séquençage

Pour étudier la diversité génétique des populations de soles à travers les localités échantillonnées nous avons opté pour une approche de génotypage par séquençage à haut débit, le séquençage de fragments d'ADN associés à des sites de restrictions dit « RAD sequencing » (Miller *et al.,* 2007; Baird *et al.,* 2008). Cette technique basée sur le séquençage de petits fragments d'ADN adjacents aux sites de coupure d'une enzyme de restriction, permet une

réduction de la complexité du génome tout en conservant une très grande couverture. Elle permet ainsi, en fonction de l'enzyme de restriction choisie, de générer un marqueur génétique tous les 5 à 100 kb en moyenne.

Tout d'abord nous avons extrait, à partir de la nageoire ou du muscle, l'ADN total de chaque individu à l'aide du kit d'extraction DNeasy Blood & Tissue kit (Qiagen). La qualité des extractions a ensuite été vérifiée par migration sur gel d'agarose et la concentration d'ADN a été mesurée au fluorimètre pour être standardisée à 25 $ng.\mu L^{-1}$ .Ensuite nous avons préparé six librairies RAD contenant chacune les ADNs de 32 individus mélangés en proportion stœchiométriques. Une version modifiée du protocole décrit dans (Baird *et al.,* 2008) a été suivie pour la construction de ces librairies, en utilisant comme enzyme de restriction l'enzyme Sbf1 dont le site de coupure correspond à une séquence de 8 paires de bases (5'-CCTGCAGG-3'). Chaque individu a été marqué par un code-barre moléculaire spécifique de 5 à 6 nucléotides intégré dans l'adaptateur de séquençage et le site de coupure, afin d'attribuer chaque séquence à un individu après le séquençage. Ces librairies ont ensuite été séquencés à l'Institut de Biologie de Lille (France) par la plateforme de séquençage du laboratoire « Génomique Intégrative et Modélisation des Maladies Métaboliques » (UMR 8199), sur un séquenceur Illumina Hi-Seq 2500 produisant des séquences courtes de 101 pb, de part et d'autre de chaque site de coupure.

2. Préparation des données et analyses bio-informatiques

Le traitement des données brutes de séquençage a été réalisé avec la suite de programmes contenue dans le pipeline Stacks (Catchen *et al.,* 2011) (Figure 12). Tout d'abord, les séquences obtenues ont été réassignées aux individus en utilisant l'information portée par les codes-barres moléculaires. Par la suite nous avons utilisé une version non encore totalement assemblée du génome de *Solea senegalensis* (en préparation, 98590 scaffolds) afin d'aligner

nos séquences et identifier les polymorphismes nucléotidiques de type SNP (single nucleotide polymorphism). Pour cette étape, nous avons utilisé le logiciel bowtie 2 (v.2.1.0) (Langmead et Salzberg 2012) en choisissant l'option « very high sensitivity ». Afin de prendre en compte le phénomène d'introgression ainsi que la divergence entre S. senegalensis et S. aegyptiaca lors de l'alignement, nous avons autorisé un nombre total de 7 discordances (mismatchs) (m=7) entre une séquence et le génome de référence. Ce nombre a été déterminé empiriquement en utilisant un sous-échantillon d'individus. Ensuite, dans pstacks nous n'avons conservé que les séquences ayant une couverture minimale de 5X sur lesquelles nous avons appliqué le model « SNP » autorisant un taux d'erreur de séquençage inférieur à 2.5% afin d'en extraire les vraies positions hétérozygotes. Ce filtre consiste à comparer pour les positions variables la vraisemblance d'un modèle hétérozygote dans lequel la proportion attendue des deux allèles échantillonnés suit une loi binomiale d'espérance 0.5 avec un modèle incluant uniquement les erreurs de séquençage pouvant atteindre un taux maximal de 2.5%. Par la suite, nous avons utilisé cstacks pour construire un catalogue regroupant les séquences de locus RAD homologues à partir de leurs positions génomiques (option –g dans cstacks) tout en appliquant le module rxstacks et l'option (--prune_haplo) afin d'éliminer les locus ayant un log de vraisemblance (log-likelihood) supérieur à -300, et de retirer les haplotypes artefactuels. Enfin chaque individu a été génotypé au niveau de toutes les positions variables contenues dans le catalogue avec sstacks et les données de génotypes ont été exportées au format VCF avec le module populations.

Avant de procéder aux analyses statistiques, le jeu de données précédemment obtenu avec Stacks a été filtré à l'aide du logiciel VCFtools (Danecek *et al.,* 2011) sur la base de critères de qualité ainsi que certains paramètres de génétique des populations comme le filtre d'écart significatif à l'équilibre d'Hardy-Weinberg (--hwe) qui nous permet d'exclure les locus

présentant d'éventuels cas de paralogie qui provoqueraient des excès d'hétérozygotie et un polymorphisme artéfactuel.

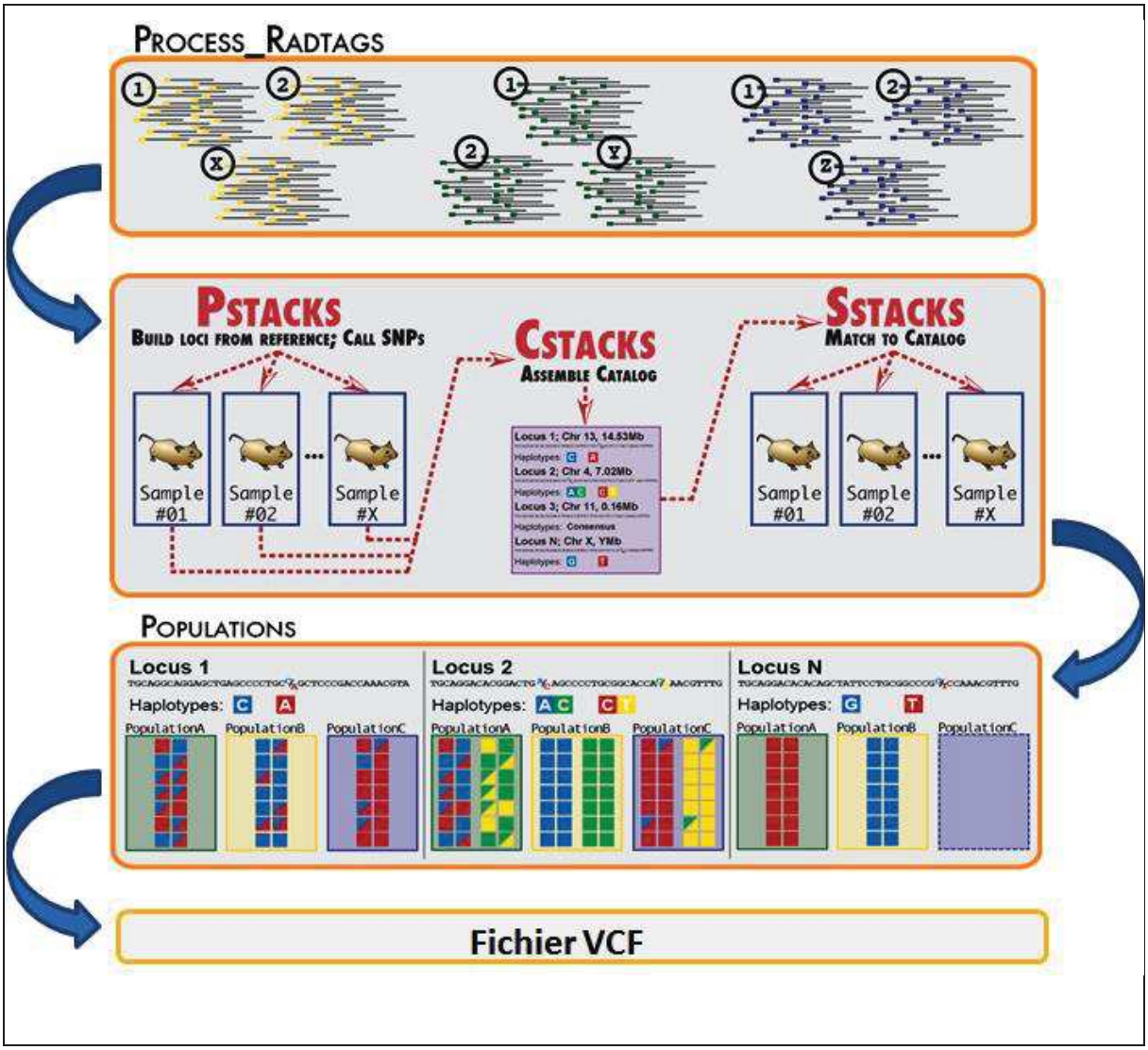


Figure 12 : Schéma d'analyse des données de séquençage via la suite des programmes contenus dans le pipeline Stacks.

3. Analyse descriptive du polymorphisme

Afin de décrire la distribution de la variabilité génétique entre individus, nous avons utilisé l'analyse en composantes principales (ACP) qui permet de résumer l'information contenue dans le jeu de données global dans un nombre de dimensions restreint. Cette analyse a été conduite avec le package R adegenet (Jombart, 2008). Les données de génotypes manquantes ont été remplacées par la fréquence moyenne de l'allèle le plus fréquent dans la population à laquelle l'individu appartient.

En plus de l'analyse en composante principale nous avons eu recours à une méthode d'assignation non spatialisé implémenté dans le programme fastStructure (Raj *et al.,* 2014) afin d'inférer la structure entre les populations échantillonnée, d'assigner les individus à ces populations et de révéler l'existence éventuelle de l'hybridation ou de l'introgression. Cette méthode ne nécessite pas d'*a priori* et fait l'hypothèse que les individus appartiennent à des populations correctement échantillonnées, et que ces populations sont à l'équilibre d'Hardy-Weinberg ainsi qu'à l'équilibre de liaison.

4. Résultats

4.1. Obtention du jeu de données final

Suite au séquençage, l'assignation des séquences obtenues à leur individu d'origine à l'aide des barcodes moléculaire a permis de retenir 750 107 245 séquences, avec une moyenne de 3 906 809 par individu (figure). Le pourcentage de séquences retenues par librairie varie entre 80.9% et 90.8% (figure), attestant du succès de la construction et du séquençage des librairies.

L'analyse de ces séquences à l'aide du logiciel Stacks a permis de générer un fichier des génotypes individuels au format VCF, comportant 174 490 positions polymorphes parmi les 192 individus. Pour la suite des analyses, nous avons retiré de ce jeu de données 6 individus

ayant plus de 60% de données génotypiques manquantes. Avec le programme VCFtools, nous avons ensuite subdivisé notre fichier VCF global comportant 186 individus retenus en 5 jeux de données afin de séparer les individus en fonction de leur appartenance aux différentes espèces échantillonnées et de leur provenance géographique (*P. impar*, *S. solea*, *S. senegalensis*, *S. aegyptiaca*, zone de contact). Les individus de la zone de contact ont été considérés comme un groupe à part, ceci dans le but d'appliquer des filtres spécifiques à chaque espèce sans y inclure des individus hybrides potentiels. Cette subdivision a permis de récupérer 158 994 SNPs polymorphes pour le jeu de données *S. senegalensis* (Sénégal, Cadix, Annaba et lagune de Mellah, 49 individus), 127 912 SNPs pour celui de *S. aegyptiaca* (Kerkennah, lagune El Bibane et Egypte, 48 individus), 140 738 SNPs pour le jeu de données comportant les individus de la zone de contact (Tabarka, lagune de Bizerte et Golfe de Tunis, 79 individus) et finalement 138 104 et 81 342 SNPs polymorphes respectivement pour *S. solea* et *Pegusa impar* (9 et 1 individus respectivement). En dernière étape, nous avons regroupé ces 5 jeux de données pour obtenir un nouveau fichier VCF catalogue composé de 186 individus et 171 991 SNP ayant passé les filtres appliqués au sein des espèces, et qu'on utilisera par la suite pour créer de nouveaux jeux de données spécifiques à d'autres analyses postérieures.

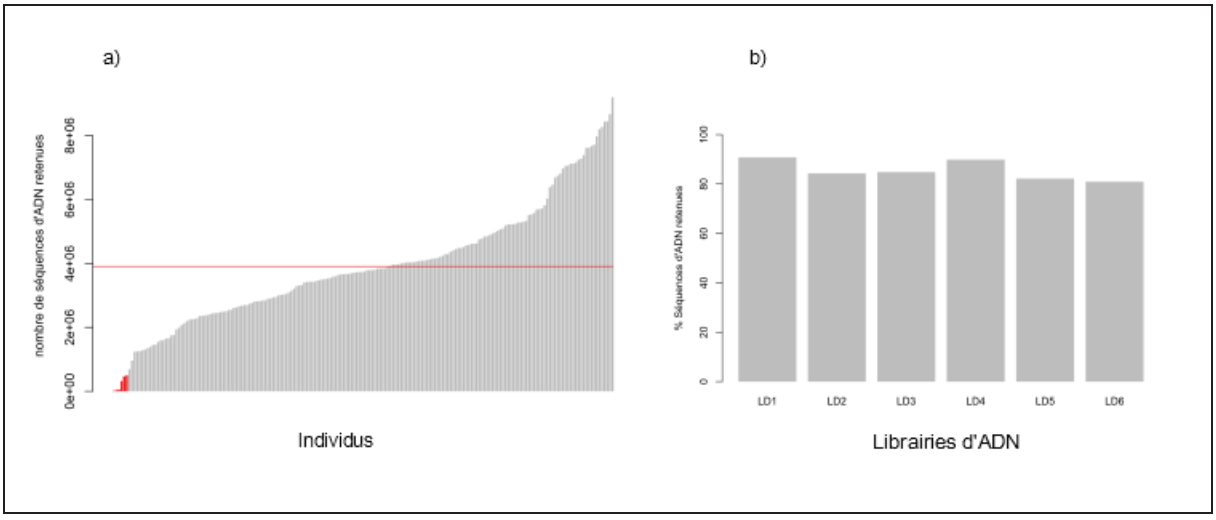


Figure 13 : Résultats du succès de séquençage. (a) Nombre de séquences retenues par individu, (b) Pourcentage de séquences retenues par librairie.

4.2. Distribution de la variabilité génétique

L'ACP réalisée sur le jeu de données global (avant l'application des filtres) composé de 174 490 SNPs nous a permis de distinguer les quatre différentes espèces *S. solea*, *S. senegalensis*, *S. aegyptiaca* et *P. impar*. Le premier axe de l'ACP explique 56.67% de l'inertie génétique totale entre individus et discrimine aisément les individus appartenant aux espèces *S. senegalensis* et *S. aegyptiaca*. Les deuxième et troisième axe de l'ACP qui représentent 13 % de la variation totale distinguent simultanément les individus *S. solea* et *P. impar* des deux précédentes espèces. Alors que l'axe trois ne met en évidence qu'un seul individu *S. impar* le deuxième axe quant à lui nous révèle que le nombre total d'individus *S. solea* que comporte notre jeu de donnée est de 9 individus au lieu d'un seul, indiquant des problèmes d'identification morphologiques. A cause de la variation apportée par les axes 2 et 3 de cette ACP qui ont tendance à regrouper les différentes populations de *S. senegalensis* et *S. aegyptiaca* en deux grands groupes, la structure génétique inter populations au sein de ces derniers est moins évidente à discerner. Une deuxième ACP a donc été réalisé à partir du jeu de données filtré (jeu de données final utilisé dans les analyses postérieures du chapitre 3) tout en tenant compte cette fois ci d'éliminer tous les individus de *S. solea* et *P. impar* qui tiraient l'axe 2 et 3 de la précédente analyse pour s'intéresser uniquement à nos deux espèces modèles.

L'axe 1 de cette deuxième ACP absorbe 74.2% de la variation génique totale et permet de distinguer clairement deux groupes. Au sein de ces derniers les populations échantillonnées suivent un ordre géographique allant du Sénégal jusqu'à la lagune de Bizerte (Groupe *S. senegalensis*) et de L'Egypte vers le Golfe de Tunis (Groupe *S. aegyptiaca*) et ceci selon que l'on se rapproche du centre de l'ACP. L'axe 2 quant à lui explique 0.73% de la variation génétique totale et permet de distinguer les individus issus de l'Egypte tandis que l'axe 3 avec 0.66% de la variation génétique totale discrimine légèrement les populations échantillonnées au Sénégal et à Cadix.

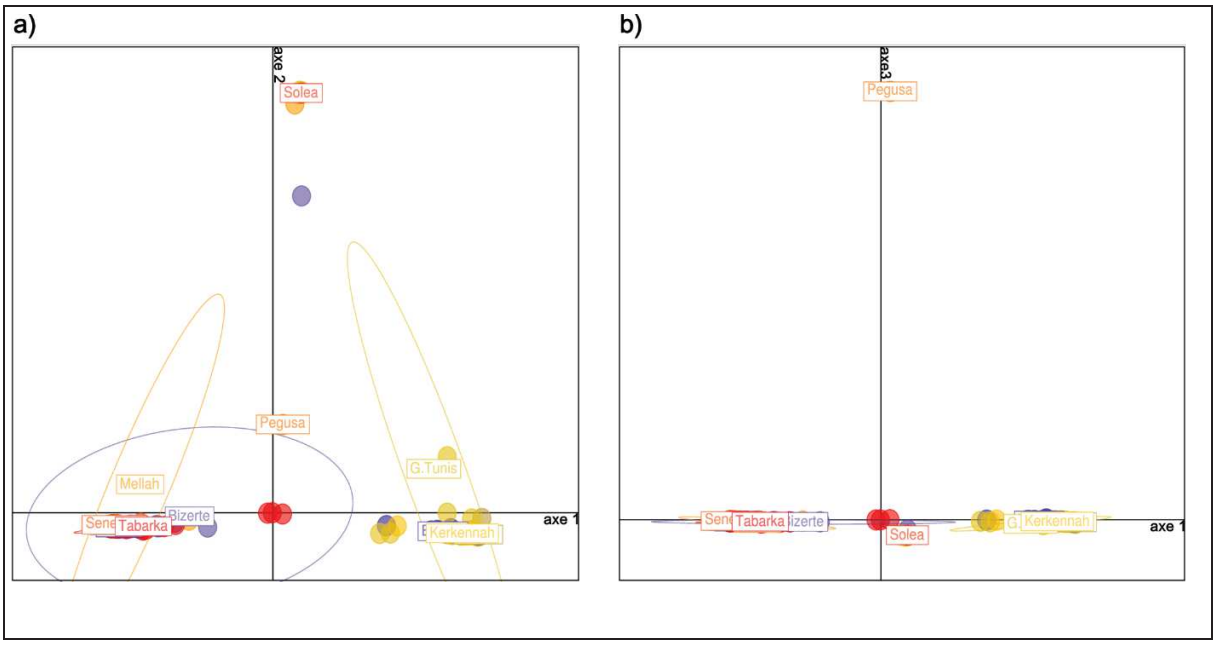


Figure 14 : Analyse en composantes principales des données non filtrées. a) Axe 1& 2. b) Axe 1 & 3.

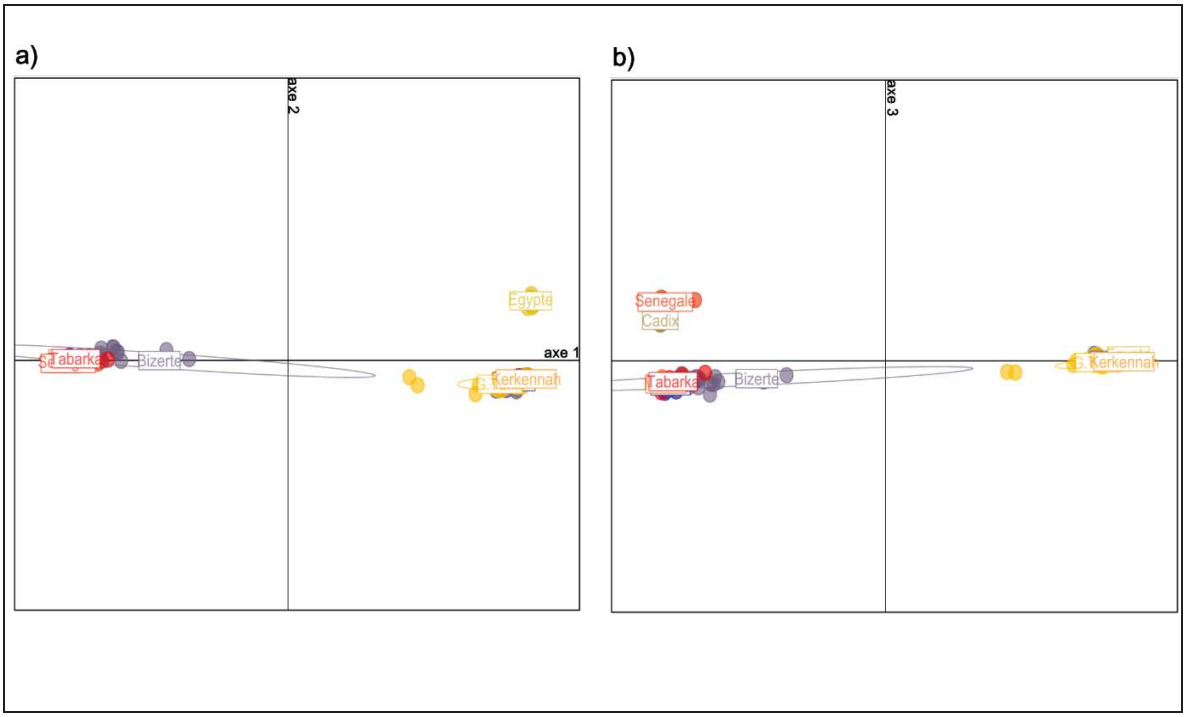


Figure 15 : Analyse en composantes principales sur le jeu de données final utilisé dans les analyses postérieures du chapitre 3 (161 individus & 10 756 locus). a) Axe 1& 2. b) Axe 1 & 3.

L'analyse réalisée avec fastStructure sur les données non filtrées avec un nombre de populations fixé à priori K=4 nous a permis de conforter les résultats obtenus par les différentes ACP. En effet grâce à cette méthode d'assignation nous avons pu confirmer l'existence d'individus mal identifiés comme étant des *S. senegalensis* et ou *S. aegyptiaca* dans différentes localités (Anaba, Lagune de Bizerte et Golfe de Tunis).Aussi nous avons constaté qu'en éliminant les 6 individus ayant plus de 60% de données manquantes nous éliminons aussi les individus qui semblaient présenter des taux de mélanges élevés à Tabarka (mêmes individus qui occupent le centre de l'ACP) mettant ainsi en évidence l'impact des données manquantes sur le taux de l'introgression (Figure16a&b).Enfin l'analyse des données ayant passées tous les filtres dans fastStructure avec un a priori K =2 nous a permis de confirmer l'existence d'un phénomène d'introgression qui touche essentiellement la lagune de Bizerte et le golfe de Tunis et avec un degrés moindre quelques individus de Tabarka (Figure 16c).

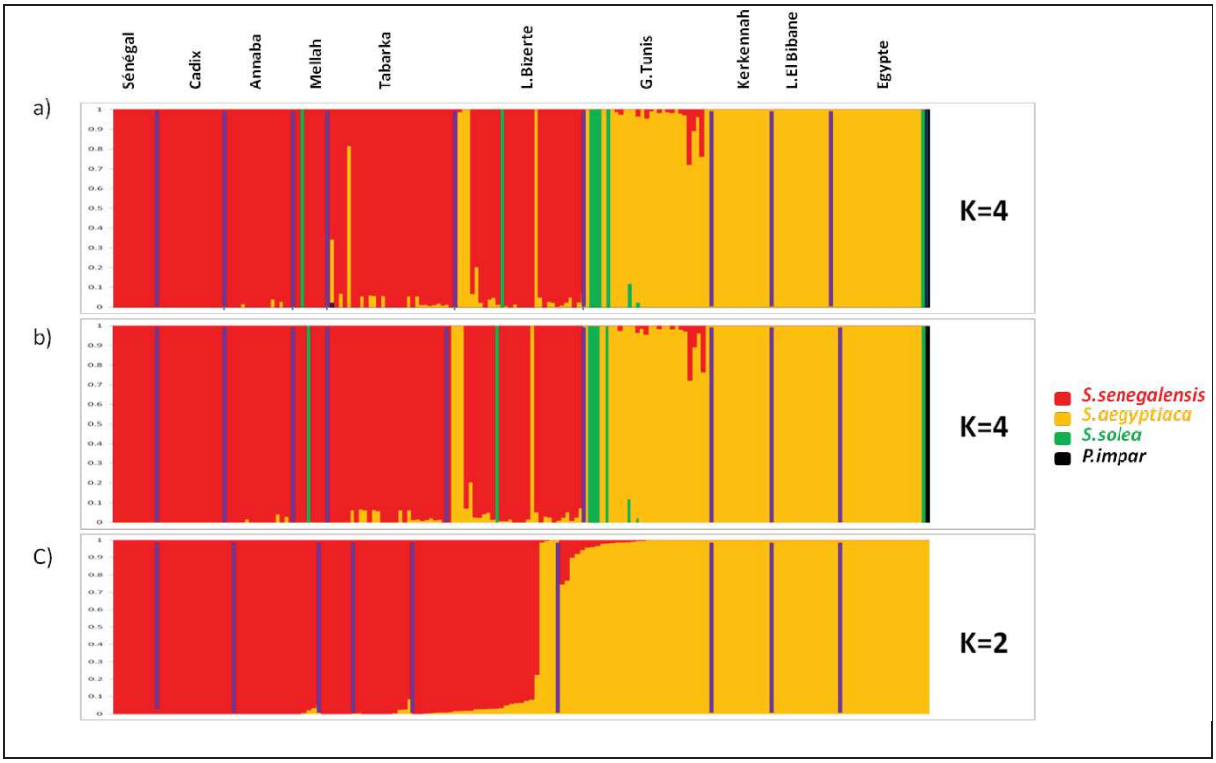


Figure 16 : Résultats de l'analyse avec fastStructure pour différentes valeurs de K.(a) Données brutes non filtrées (b) Données brutes sans données manquantes. (c) Données finales utilisées dans les analyses du chapitre 3 (161 individu & 10756 locus).

5. Discussion

A l'issu du séquençage et des analyses bio-informatiques nous avons obtenu un catalogue de 171 991 SNPs provenant de quatre espèces différentes. Etant donné que ces SNPs sont échantillonnés aléatoirement sur le génome, ce résultat nous procure une puissance statistique considérable pour la discrimination de ces 4 espèces. En effet, ceci a bien été traduit dans notre étude, par la découverte de 8 individus supplémentaires de l'espèce *S. solea* que nous avions considéré au départ comme des *S. senegalensis* et/ou *S. aegyptiaca* en se basant sur une analyse sommaire du phénotype. Cependant l'utilisation d'un nombre aussi important de marqueurs pour l'étude de la génétique des populations suscite d'autant plus de précautions par rapport à l'hétérogénéité de séquençage ainsi que dans le choix de paramètres de filtrage des données qui peuvent impacter les méthodes statistiques sensibles aux données manquantes ou artéfactuelles et influencer l'interprétation biologique des résultats. L'analyse de la distribution de la variabilité génétique entre populations de *S. senegalensis* et *S. aegyptiaca* montre que le gradient de différentiation génétique diminue selon que l'on se rapproche géographiquement de la zone de contact précédemment décrite en 1987 et 2011 (She *et al.,* 1987; Ouanes *et al.,* 2011). Ce résultat semble donc indiquer une introgression plus forte au niveau des localités proches de la zone de contact, et ce chez les deux espèces. Au-delà de cette évidence indirecte d'échanges génétiques interspécifiques, nos données actuelles ne semblent cependant pas contenir d'hybrides de première génération (F1), qui occuperaient une position centrale par rapport à l'axe 1 de l'ACP. Cette observation est néanmoins conforme à celle généralement faite que les hybrides F1 sont habituels rares au sein des zones d'hybridation étudiées chez les autres espèces. Elle suggère donc que nous serions plutôt dans une situation où les deux espèces échangent des gènes par introgression par l'intermédiaire des rares génotypes hybrides produits. L'analyse détaillée de ces échanges génétiques à travers la zone de contact entre *S. senegalensis* et *S. aegyptiaca* fera l'objet du chapitre suivant.

Chapitre III : Histoire du flux génique et signature génomique des échanges génétiques à travers la zone d'hybridation semi-perméable

Introduction

Pour ce dernier chapitre, notre objectif est d'utiliser les données de polymorphisme génomique générées précédemment afin de reconstituer l'histoire de la divergence entre *S. senegalensis* et *S. aegyptiaca* et de caractériser les échanges génétiques qui ont eu lieu entre ces deux espèces à travers leur zone d'hybridation. Nous allons chercher en particulier à caractériser les éventuelles variations d'intensité du flux génique le long du génome, prédites par le modèle de barrière semi-perméable. Notre démarche fait intervenir une combinaison de trois approches différentes. En premier lieu, nous nous sommes intéressés à reconstituer l'histoire du flux génique entre *S. senegalensis* et *S. aegyptiaca* en utilisant une méthode d'inférence démo-génétique basée sur l'analyse de la différenciation génétique entre espèces. Cette approche n'est pas locus spécifique et donne donc une vision globale des flux de gènes sans nous renseigner sur l'intensité de l'introgression à l'échelle d'un locus particulier. Nous avons donc utilisé en second lieu une méthode se basant sur l'étude du comportement individuel des locus en fonction de l'indice hybride génomique. Cette méthode compare le degré d'introgression de chaque locus à son attendu par rapport à la moyenne de tous les autres locus. Elle permet donc d'identifier des locus avec des patrons d'introgression exceptionnels, qui introgressent plus fortement dans l'une ou l'autre des espèces ou qui au contraire résistent fortement à l'introgression. Enfin, dans le but de caractériser l'introgression dans l'espace et de pouvoir détecter des locus montrant des patrons de fréquence allèliques indiquant une introgression forte, nous avons utilisé la méthode

des clines géographiques qui permet d'analyser le flux de gènes le long d'un transect à travers la zone d'hybridation.

Un avantage à combiner ces trois approches est qu'elles s'appuient sur des aspects différents des données (temporel, génomique, spatial), tels qu'ils peuvent être analysés et prédits par les modèles appropriés de la génétique des populations. L'étude de la congruence des différents résultats nous permettra donc d'étudier ces différents aspects de l'introgression entre *S. senegalensis* et *S. aegyptiaca* de manière conjointe. Ainsi, cette démarche intégrative combinant ces trois dimensionnalités de l'introgression nous apportera plus de précision sur le processus de la spéciation entre ces deux espèces.

1. Analyses démo-génétiques

1.1. Spectre joint des fréquences allèliques

Afin de caractériser le processus de divergence qui a pu conduire au paysage actuel de la différenciation génétique entre *S. senegalensis* et *S. aegyptiaca*, nous avons utilisé des modèles offrant une représentation simplifiée de la réalité, que l'on a comparés aux données observées. L'objectif ici est donc d'identifier lequel de ces modèles procure le meilleur pouvoir explicatif des données observées. Pour résumer l'information contenue dans les données de polymorphisme, nous avons utilisé une méthode basée sur l'information contenue dans le spectre joint des fréquences alléliques (ou JAFS pour joint allele frequency spectrum). Cette méthode est implémentée dans une version modifié du programme δaδi (Gutenkunst *et al.,* 2009). Le spectre joint est une statistique résumée de la différentiation génétique sur l'ensemble des marqueurs considérés. Cette statistique se présente sous la forme d'une représentation en deux dimensions de la fréquence d'occurrence d'une mutation dérivée présente en xi copies dans une espèce 1 (*S. senegalensis*) et yi copies dans une espèce 2 (*S. aegyptiaca*). Pour un nombre N d'individus diploïdes échantillonnés dans chaque espèce, le spectre joint représente

donc la fréquence d'apparition de cette mutation pour xi et yi variant entre 1 et 2N copies dans chaque espèce.

Pour construire le spectre joint des fréquences allèliques, nous avons tout d'abord essayé d'éliminer toute information redondante portée par les locus se trouvant en déséquilibre de liaison physique, en choisissant aléatoirement un SNP par paire de locus RAD associés à un même site de restriction. Pour cela, nous utilisons la position connue des locus RAD sur le génome de référence. Comme nous avons séquencé 100pb de chaque côté de chaque site de restriction, nous choisissons aléatoirement un SNP par fenêtre de 200 pb le long du génome. Les sites de restriction étant la plupart du temps espacés de plusieurs milliers de paires de bases, cela nous permet d'éliminer une grande partie du déséquilibre de liaison physique contenu dans les données initiales. Cette première étape a pour objectif de réduire le jeu de données à une collection de SNPs considérés comme indépendants les uns des autres. Afin de pouvoir distinguer l'allèle dérivé de l'allèle ancestral au niveau de chaque SNP, nous avons utilisé des séquences RAD obtenues chez S. solea comme groupe externe. Cette seconde étape nous permet d'orienter les mutations lors de la construction du spectre pour mieux décrire la magnitude des échanges géniques entre *S. senegalensis* et *S. aegyptiaca* (Gutenkunst *et al.,* 2009). Enfin nous avons fait le choix de regrouper pour chaque espèce différentes populations éloignées de la zone d'hybridation pour obtenir une bonne précision d'estimation des fréquences allèliques dans chaque espèce, et ainsi avoir un spectre joint de fréquence allèliques le plus précis possible. Ce choix a été fait en prenant en considération les résultats issus des précédentes analyses (ACP et Structure). Ainsi, nous avons obtenu un jeu de données composé de 10 756 SNPs chez 88 individus (44 *S. senegalensis*, 44 *S. aegyptiaca*). Pour ce jeu de données, nous avons autorisé au maximum 4 génotypes manquants par espèce pour chaque locus. Théoriquement, il est donc possible de construire un spectre joint de dimension 80x80 (40

génotypes diploïdes dans chaque espèce), que nous avons réduit à 40x40 pour lisser l'information contenue dans le spectre.

1.2. δaδi et ajustement des modèles démographiques

Dans le but de reconstituer l'histoire démo-génétique à l'origine du paysage actuel de différentiation génétique entre *S. senegalensis* et *S. aegyptiaca*, nous avons utilisé une version modifiée du programme δaδi pour ajuster des modèles simplifiés de la réalité à nos données observées décrites par le spectre joint des fréquences allèliques (JAFS). Ce programme se base sur des équations de diffusion pour générer une approximation analytique du spectre joint et ceci pour toute combinaison de valeurs de paramètres associées à un modèle donné de divergence. A chaque vecteur de paramètres testé lui est associée une valeur de vraisemblance des données observées selon le modèle considéré. L'ajustement des valeurs des paramètres se fait alors par maximisation de la vraisemblance en explorant l'espace des paramètres par une approche statistique. Contrairement aux versions initiales du programme, nous avons utilisé la méthode d'optimisation dite de recuit simulé (simulated annealing) combinée avec une méthode quasi-Newtonienne nommée BFGC, telle qu'implémentée dans (Tine *et al.,* 2014). Cette méthode permet de se déplacer plus efficacement dans l'espace des paramètres en autorisant des franchissements de vallées de vraisemblance.

Dans le cadre de notre étude, nous avons comparé 7 modèles représentant les processus démographiques et de sélection qui ont modelé le paysage génomique de la divergence durant le processus de la spéciation (Tine *et al.,* 2014; Le Moan *et al.,* 2016). Le premier et le plus simple des modèles décrit une situation d'isolement strict (SI, strict isolation). Ce modèle correspond à un scénario de divergence allopatrique où l'on considère une population ancestrale de taille efficace NA qui se divise en deux populations filles de taille N1 (*S. senegalensis*) et N2 (*S. aegyptiaca*) qui évoluent ensuite séparément pendant un nombre de générations Ts Sans échanger de migrant. Trois autres modèles intégrant du flux génique lors de la divergence ont

été testés par la suite. Le premier est caractérisé par une migration continue, sans interruption du flux génique durant tout le temps de la divergence (IM, isolation with migration). Le deuxième modèle correspond à un cas d'une migration ancienne (AM, ancient migration) où le flux de gène ne peut avoir lieu que pendant une certaine période de TAM générations directement après la divergence et avant un temps TS d'isolement strict. Par contraste à cette situation de migration ancienne, le troisième modèle que nous avons testé est celui d'un contact secondaire (SC, secondary contact) qui intègre le flux génique uniquement pendant un temps de TSC générations après une divergence allopatrique de TS générations. Dans cette catégorie de modèles d'isolement, le flux génique entre les deux espèces est considéré comme homogène, c'est à dire que le paramètre de migration (m12 de *S. aegyptiaca* vers *S. senegalensis* et m21 dans l'autre direction) présente la même valeur pour tous les locus dans chaque direction. Cependant comme les taux de migration réciproques entre *S. senegalensis* et *S. aegyptiaca* peuvent être variables à travers le génome, nous avons ajouté une seconde catégorie de paramètres de migration (m'12 et m'21) qui s'applique à un sous ensemble de locus. Ainsi nous construisons des modèles dis à migration hétérogène qui permettent de prendre en considération deux catégories de locus à travers le génome. La première catégorie (taux m12 et m21 à migration réduite) serait étroitement liée aux régions génomiques qui contribuent au maintien de la barrière d'isolement entre *Solea senegalensis* et *Solea aegyptiaca*. L'autre catégorie, avec des taux de migration m'12 et m'21 plus élevés serait le résultat d'une diffusion de locus neutres qui sont suffisamment éloignés de ces régions contre-sélectionnées. Nous avons alors testé trois nouveaux modèles (IM2m, AM2m et SC2m) qui intègrent l'hétérogénéité du flux génique via de simples extensions implémentées à partir des modèles IM, AM et SC. En effet cette dernière catégorie de modèles considère deux catégories de locus présents en proportion P et 1-P dans le génome selon qu'ils s'échangent avec des taux de migration faibles ou fort. Cette extension

offre ainsi la possibilité de capturer l'effet de la sélection qui réduit localement la migration efficace pour générer des patrons hétérogènes de divergence à travers le génome.

Etant donné que la vraisemblance d'un modèle est dépendante du nombre de paramètres utilisés, la comparaison des modèles entre eux est basée sur le critère d'Akaiké (AIC) qui pénalise la vraisemblance d'un modèle en fonction de son nombre de paramètres. L'AIC est calculé à partir de la formule suivante :

$$AIC = 2k - 2ln(L)$$

Où k est le nombre de paramètre du modèle considéré et ln(L) le logarithme de la vraisemblance. Ainsi le scénario évolutif retenu comme le plus vraisemblable correspond au modèle ayant la valeur d'AIC la plus faible.

2 Clines génomiques

Pour évaluer l'introgression entre *S. senegalensis* et *S. aegyptiaca*, nous avons utilisé la méthode des clines génomiques implémentée dans programme BGC de (Gompert & Buerkle, 2011; Gompert & Buerkle, 2012a). Le principe de cette méthode consiste à décrire la variation de la proportion héritée d'une population parentale pour un locus donné, ou probabilité d'ascendance (probability of ancestry), en fonction d'un indice hybride moyen calculé sous un modèle nul pour l'ensemble des locus considérés (Gompert & Buerkle, 2011). Ainsi la probabilité qu'un locus soit hérité de l'espèce parentale 1 (*S. senegalensis*) ou l'espèce parentale 2 (*S. aegyptiaca*) est reliée à l'indice hybride d'un individu. Cette méthode utilise une approche bayésienne pour estimer les deux principaux paramètres qui décrivent la forme des clines génomiques. Ces deux paramètres sont (α) et (β) qui représentent respectivement le centre et la pente d'un cline génomique. Le premier paramètre (α) mesure le décalage éventuel ($\alpha \neq 0$) de la probabilité d'ascendance pour ce locus relativement à l'attendu neutre basé sur l'indice hybride moyen de l'individu et le second (β) décrit la vitesse de transition d'une ascendance

parentale à une autre quand l'indice hybride varie. Dans la présente étude nous nous appuyons sur ces deux paramètres pour examiner la direction que prend l'introgression ainsi que pour quantifier l'intensité du flux de gènes entre les deux espèces. En effet, une valeur négative ou positive de (α) dénote respectivement une diminution ou une hausse de la probabilité qu'un locus soit issu de S. aegyptiaca tandis qu'une valeur négative ou positive de (β) décrit une transition lente (clines génomiques plats) ou rapide (clines génomiques abruptes) d'un locus d'un pool génique à un autre en fonction de l'indice hybride entre 0 et 1 (où les génotypes parentaux *S. senegalensis* ont un indice hybride de 0 et les génotypes parentaux *S. aegyptiaca* ont un indice hybride de 1).

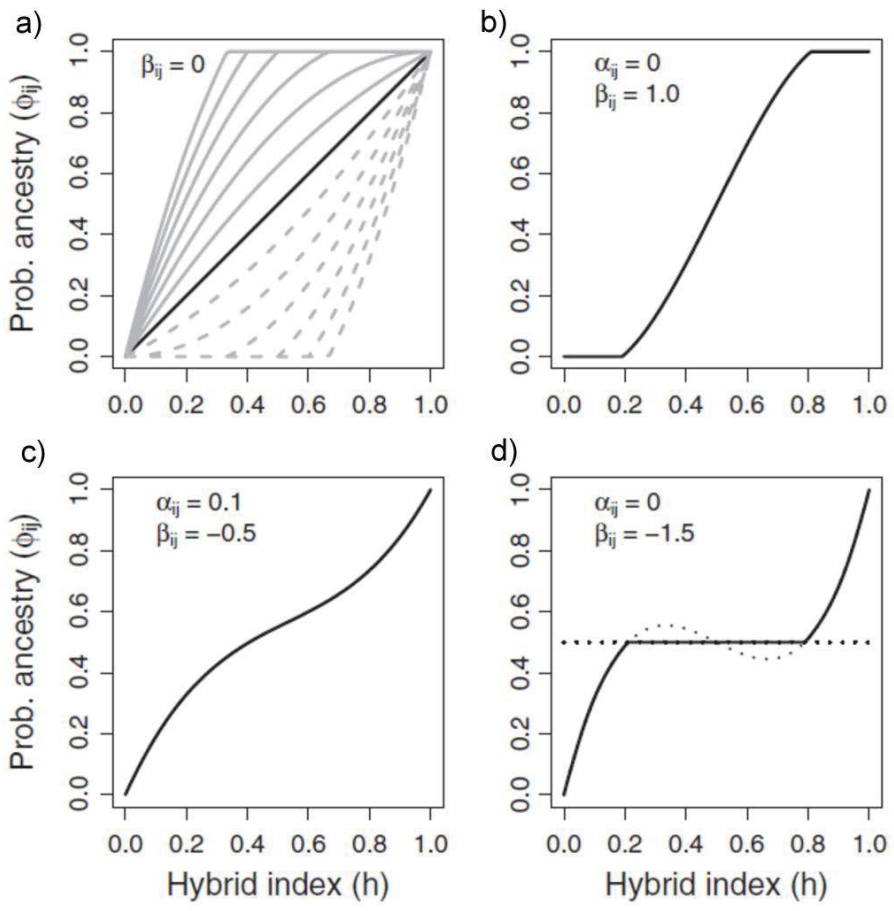


Figure 17 : Exemples de clines génomiques avec différentes valeurs de (αij) et (βij) mettant en évidence les effets de (α) & (β). (a) Pour β= 0 la ligne noire décrit l'aspect d'un cline génomique lorsque α=0 et les lignes grise en pointillés décrivent des clines pour différentes valeurs négatives de α comprises entre -0.25 et 0.25 quant aux lignes grises pleines décrivent des clines pour des valeurs de α positives comprise entre 0.25 et 1.5;(b) aspect d'un cline pour une valeurs positive de β ; (c) et (d) aspect des clines pour des valeurs de β négatives (Gompert & Buerkle, 2011).

Afin de bien estimer ces paramètres nous avons utilisé deux chaînes de simulations indépendantes dont on a vérifié par la suite la convergence vers des valeurs similaires. Chaque chaîne se compose de 150 000 itérations dont on a éliminé les 140 000 premières itérations (burn-in) pour s'assurer de la convergence, et récupéré les 10 000 dernières itérations dans la zone de convergence.

3 Clines géographiques

La mise en place d'une zone de tension sous les effets antagonistes de la migration en provenance des populations parentales et de la contre-sélection de certains génotypes hybrides peut être modélisée sous forme de clines géographiques. Ces derniers permettent de prédire la manière dont changent les fréquences alléliques des locus diagnostiques entre deux taxons le long d'un transect géographique qui traverse la zone hybride (Barton & Hewitt, 1985). Ces clines sont modélisés par une partie centrale sigmoïde, flanquée de deux queues d'introgression exponentielles. La partie centrale relativement abrupte est synonyme d'une forte contre-sélection des hybrides en raison du déséquilibre de liaison entre génotype parentaux. Par contre dans les queues d'introgression, sous l'influence de la recombinaison qui va casser les associations parentales et diminuer ainsi le déséquilibre de liaison, la forme des clines reflète la valeur sélective individuelle des allèles allo-parentaux. Afin d'étudier la géographie de l'introgression, nous avons utilisé la librairie R « Hzar » (Derryberry *et al.,* 2014) pour ajuster des clines géographiques qui modélisent la variation de la fréquence allélique pour un locus donné le long de notre transect d'échantillonnage.

Pour suivre ainsi les caractéristiques géographiques de l'introgression, nous avons utilisé un modèle où les fréquences alléliques minimales et maximales sont estimées dans le cas de locus incomplètement fixés chez les populations parentales et où l'ajustement des queues du cline se fait de façon indépendante de chaque côté (Derryberry *et al.,* 2014). Les valeurs des paramètres

de centre (c) de cline (coordonnée géographique du point d'inflexion) de largeur (w), ainsi que les fréquences alléliques minimales (Pmin) et maximales (Pmax) sur les queues des clines ajustés ont été estimées pour chaque locus. Ces valeurs de paramètres nous ont permis de calculer par la suite la pente (S) de chaque cline par la relation : S= (Δ P⁄w) (Figure 18) (Stankowski *et al.,* 2016).

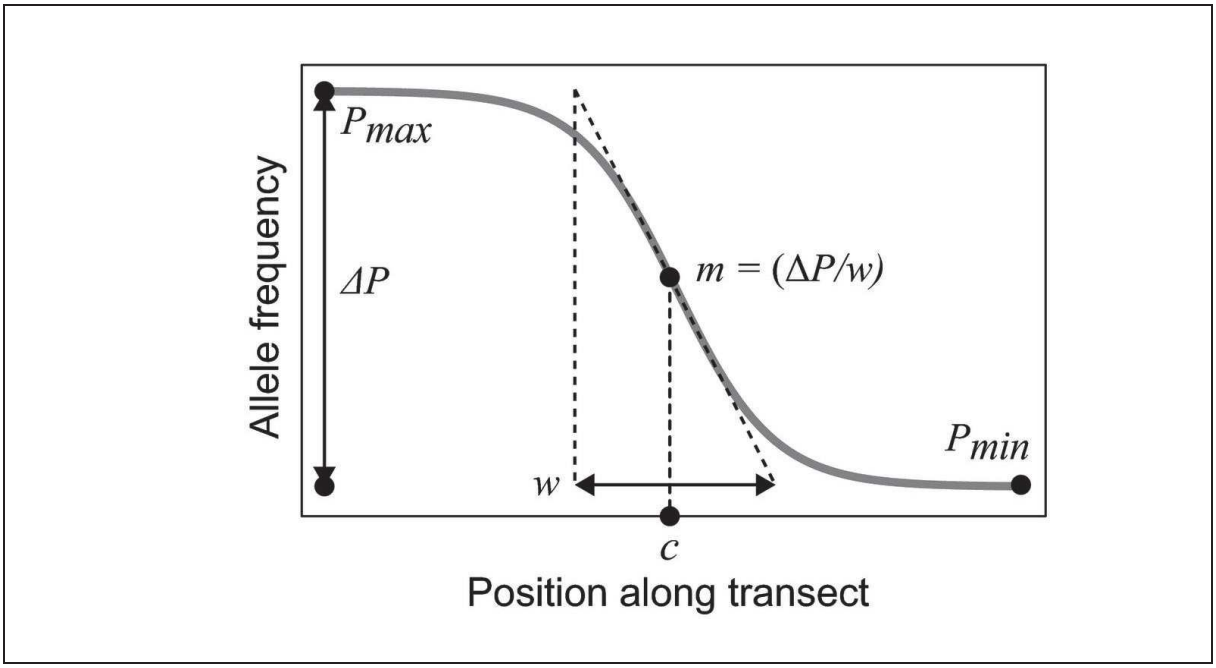


Figure 18 : Modèle sigmoïde d'un cline géographique (Stankowski *et al.,* 2016).

4 Résultats

4.1. Ajustement du meilleur modèle démo-génétique de divergence

Afin d'avoir une meilleure résolution du spectre joint, ce dernier a été construit sur 20 individus de chaque espèce (total de 40 allèles par espèce) échantillonnés parmi les 44 individus du départ. Cette projection permet ainsi d'augmenter la densité de SNPs présents dans chaque case (celle-ci diminuant de manière quadratique avec le nombre de cases du spectre). Ce spectre

montre que la majorité du polymorphisme se situe au niveau du cadre du spectre et que relativement peu de SNPs sont présents sur la diagonale ou au centre de ce dernier (Figure 19). Nous constatons aussi qu'une grande partie des allèles dérivés est fixé différentiellement entre les deux espèces (coins en bas à droite et coins en haut à gauche). Ces données observées décrites par le spectre des fréquences allèliques nous révèlent que la quasi-totalité de la variation du génome est sous la forme de variant soit privés à l'une des deux espèces, soit différentiellement fixés entre *S. senegalensis* et *S. aegyptiaca*, et que seule une petite partie des polymorphismes a éventuellement diffusé par introgression d'une espèce à l'autre.

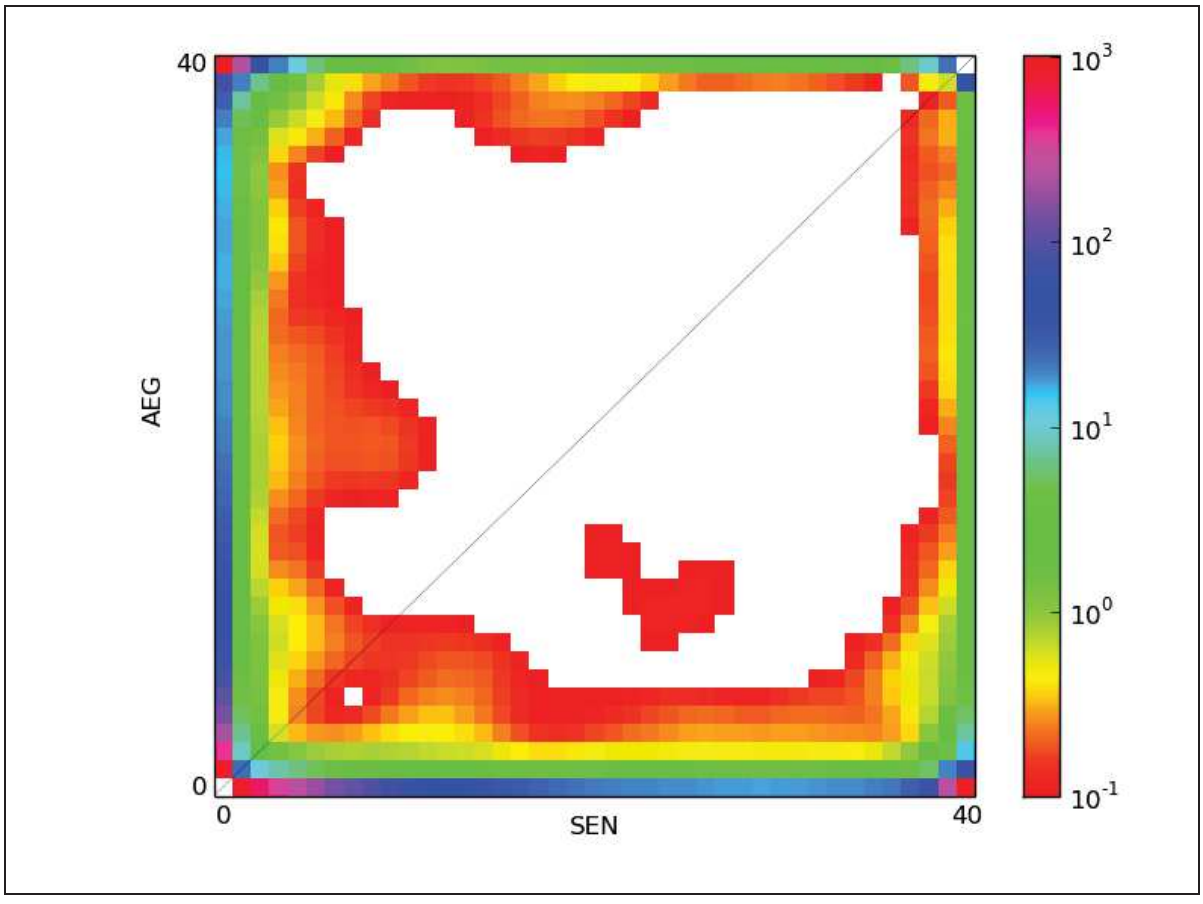


Figure 19 : Spectre joint des fréquences alléliques ajusté aux données observées (40x40).

Le spectre joint des fréquences alléliques construit à partir des données observées a été utilisé pour ajuster différents modèles évolutifs représentant divers scenarios d'isolement avec ou sans flux génique (§ ci-dessus). Le premier modèle de divergence ajusté au spectre observé est celui d'une situation d'isolement strict (SI pour Strict Isolation) où le de flux de gène entre les deux espèces est nul. Les résultats obtenus avec ce modèle sont loin de la réalité observée et décrite par le spectre joint (Figure 20).

En effet le scenario d'un isolement strict, s'il permet d'expliquer les marqueurs présents sur les bords du spectre observé, prédit en revanche mal autres catégories de locus qui occupent la partie centrale du spectre. La seconde catégorie de modèles démographiques testée intègre des paramètres de migration (m12 et m21) décrivant un flux génique pouvant être soit continu (IM pour Isolation with Migration) ou ancien (AM pour Ancient Migration) ou récent (SC pour Secondary Contact) dans le temps. Cette catégorie de modèles de divergence avec flux génique permet de mieux prédire l'existence de marqueurs localisés à l'intérieur du spectre joint. Cependant, ces modèles sous-prédisent encore la densité des SNPs situés loin du cadre dans la partie centrale du spectre. Finalement la troisième catégorie de scénarios de divergence, qui intègre un flux génique hétérogène à travers le génome s'ajuste mieux aux résultats observés (tableau 3) et permet notamment de prédire à la fois la présence des locus très différenciés et des locus localisés dans la partie centre du spectre (Figure 20).

En effet, en ajoutant une seconde catégorie de paramètres de migration (m'12 et m'21) aux modèles précédents, on prend en considération deux grandes catégories de locus qui sont caractérisés par des niveaux de flux génique potentiellement très différents. La première catégorie de locus caractérisée par un taux de migration réduite serait ainsi étroitement liée aux régions génomiques hautement différenciés qui contribuent au maintien de la barrière d'isolement entre *Solea senegalensis* et *Solea aegyptiaca*. Par contre, seconde catégorie de locus avec un taux de migration plus élevé refléterait la diffusion de locus neutres qui sont

suffisamment éloignés des zones génomiques impliqués dans l'isolement reproductif, ou encore correspondrait à des locus qui sont en liaison avec des mutations avantageuses.

Tableau 3 : Détail des inférences démographique ainsi que les estimations des différents paramètres des modèles (***n*** : taille efficace ; ***m*** : taux de migration neutre ; ***m'*:** migration efficace réduite ; T_s : temps de séparation ; T_{am} : temps de migration ancestrale ; T_{sc} : temps de contact secondaire ; ***P*** : proportion de locus à migration réduite ; ***O*** : probabilité de bonne orientation du spectre).

model	*AIC*	*Log-likelihood*	*Theta*	n_1	n_2	m_{12}	m_{21}	m'_{12}	m'_{21}	T_s	T_{am}	T_{sc}	*P*	*O*
SC2M	2424.46	-1202.23	747.20	0.818	1.136	4.607	0.381	0.057	0.153	4.80		0.081	0.05	0.971
SC	2564.84	-1275.42	2160.72	0.291	0.354	0.215	0.521			1.01	0.0	0.023		0.981
AM2M	3041.03	-1510.51	1207.61	0.477	0.649	0.846	0.308	0.033	0.058	3.1			0.051	0.982
IM2M	3041.33	-1511.66	1491.59	0.395	0.525	0.110	6.635	0.046	0.074	2.41			0.054	0.982
AM	3048.24	-1517.12	514.41	1.141	1.533	0.017	0.026			8.55	0.0			0.977
IM	3049.93	-1518.96	1018.56	0.572	0.768	0.035	0.053			3.80				0.981
SI	4611.53	-2301.76	2569.18	0.226	0.287	0.571								0.988

Les inférences démo-génétiques réalisées entre *S. senegalensis* et *S. aegyptiaca* permettent de sélectionner de façon nette le modèle du contact secondaire avec migration hétérogène (SC2m). Ce modèle possède la valeur d'AIC la plus faible (2424,4), et donc l'écart avec le second modèle le plus vraisemblable est supérieur à 10, considéré comme un support statistique très fort en faveur du meilleur modèle. L'analyse des valeurs des paramètre de modèle SC2m révèle que le temps de divergence entre les deux espèces est près de 60 fois plus élevé que le temps écoulé depuis la remise en contact. Ce modèle nous indique donc que le contact secondaire est un événement relativement récent dans l'histoire de la divergence entre les deux espèces de soles. Il révèle aussi que seulement 5% du génome est susceptible de s'échanger entre ces deux

espèces plus facilement que la vaste majorité du génome. Ainsi, près de 95 % du génome présenterait une migration efficace très faible, et se faisant essentiellement de *S. aegyptiaca* vers *S. senegalensis* (2.6 fois plus élevée dans ce sens). Les 5 % restant du génome présenterait une migration efficace plus élevée, mais se faisant essentiellement de *S. senegalensis* vers *S. aegyptiaca* (12 fois plus élevée dans ce sens). L'asymétrie de migration détectée par notre modèle de contact secondaire avec introgression variable montre donc des directions d'introgression préférentielles opposées entre : (i) la majorité (95%) du génome qui s'échange peu entre les deux espèces mais plus facilement dans le sens d'une introgression chez *S. senegalensis*, et (ii) une minorité (5%) du génome qui introgresse avec un taux nettement plus élevé, mais essentiellement dans le sens d'une introgression chez *S. aegyptiaca*

4.2. Probabilité de l'introgression

Afin d'évaluer le comportement individuel des locus, nous avons développé une statistique nous permettant d'estimer la probabilité qu'un locus localisé dans une case donnée du spectre joint appartienne à la catégorie des locus qui introgressent le plus. Pour ce faire, nous utilisons les paramètres du meilleur modèle ajusté, celui d'un contact secondaire hétérogène avec deux taux de migration dans chaque direction (SC2m). Ces deux taux caractérisent le flux génique au sein de deux compartiments du génome : les zones sous sélection où la migration efficace est faible (95 % du génome avec des taux m12, m21) et les zones évoluant neutralement (5 % du génome avec des taux m'12, m'21). Ainsi, notre modèle SC2m peut être décrit comme la combinaison linéaire de deux modèles simples de contact secondaire (SC) décrivant chacun les flux géniques propres à ces deux types de régions du génome. Nous avons donc utilisé ces deux modèles simples de contact secondaire pour générer des jeux de données théoriques par simulations en prenant en considération les paramètres de migration inférés par notre modèle

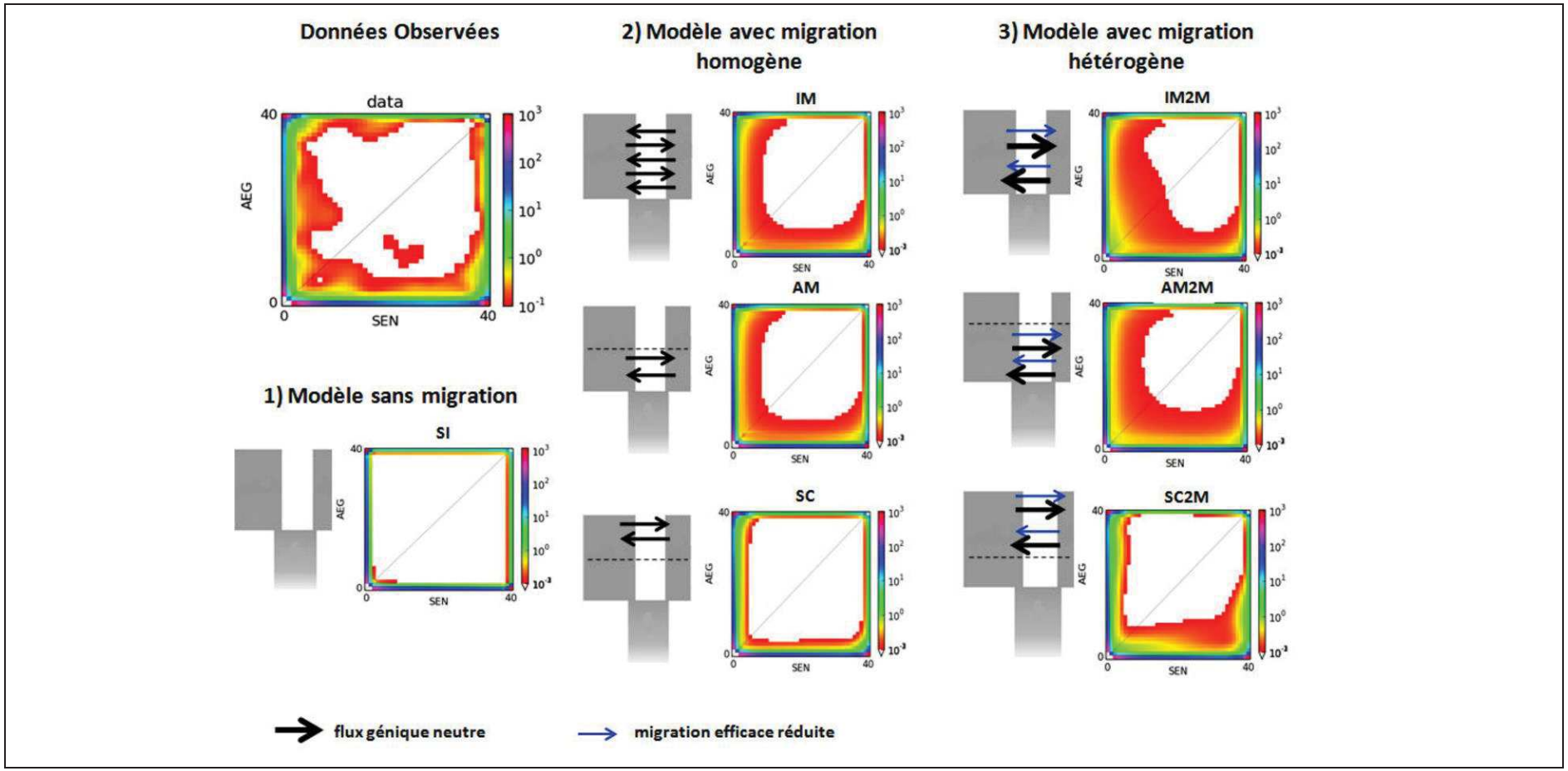


Figure 20 : Résultats des inférences démographiques des différents scénarios évolutifs. (**SI**) Isolement stricte, (**IM**) Isolement avec migration, (**AM**) Ancienne migration, (**SC**) Contact secondaire, (**IM2M**) Isolement avec migration hétérogène, (**AM2M**) Ancienne migration hétérogène, (**SC2M**) Contact secondaire avec migration hétérogène.

SC2m. Nous avons réalisé 1000 simulations sous chacun des deux modèles de contact secondaire (SC) avec le programme MSMS (Ewing & Hermisson, 2010), et construit deux spectres joints moyennés à partir des données simulées. A partir de ces deux spectres, nous avons par la suite généré un nouveau spectre de probabilité que nous avons défini pour chaque case comme la probabilité qu'un locus appartienne au compartiment à migration efficace réduite comme suit :

$$Ps=((95*SCm12m21))/(((95*SCm12m21+5*SCm'12m'21)))$$

Où SCm12m21 est la matrice des valeurs prédites par δaδi dans un spectre joint s'il n'y avait qu'un seul taux de migration de valeurs m12 et m21 et SCm'12m'21la même chose relative à la migration neutre avec taux m'12 et m'21. Réciproquement, la probabilité Pn qu'un locus présent dans une case donnée du spectre observé résulte d'une introgression neutre (libre) s'écrira $Pn=1-Ps$. **0.1**

4.3. Clines génomiques

L'analyse des clines génomique réalisée avec le programme BGC nous a révélé différents types de clines chez différents locus. En effet, suivant les valeurs de α estimées nous avons pu distinguer des locus qui ont une probabilité significativement élevée d'ascendance *S. senegalensis* (α négatif) et également d'autres qui ont une probabilité forte d'ascendance *S. aegyptiaca* (α positif). Nous avons constaté aussi la présence de locus qui se caractérisent soit par l'augmentation ou la diminution du taux d'introgression qui se reflètent respectivement par des clines à pente raide quand ($\beta > 0$) et vice versa quand ($\beta < 0$). Dans notre étude nous avons considéré deux seuils de significativité statistique pour identifier les locus présentant des comportements d'introgression atypique, c'est à dire déviants de la majorité des autres locus du génome. Le premier seuil se base sur l'intervalle de confiance bilatéral à 95% de la distribution de α et de β. Ainsi on dit qu'un locus présente un excès d'ascendance (excess ancestry) par

rapport à l'attendu quand l'intervalle de confiance bilatéral à 95% de la distribution de α ne comporte pas 0.

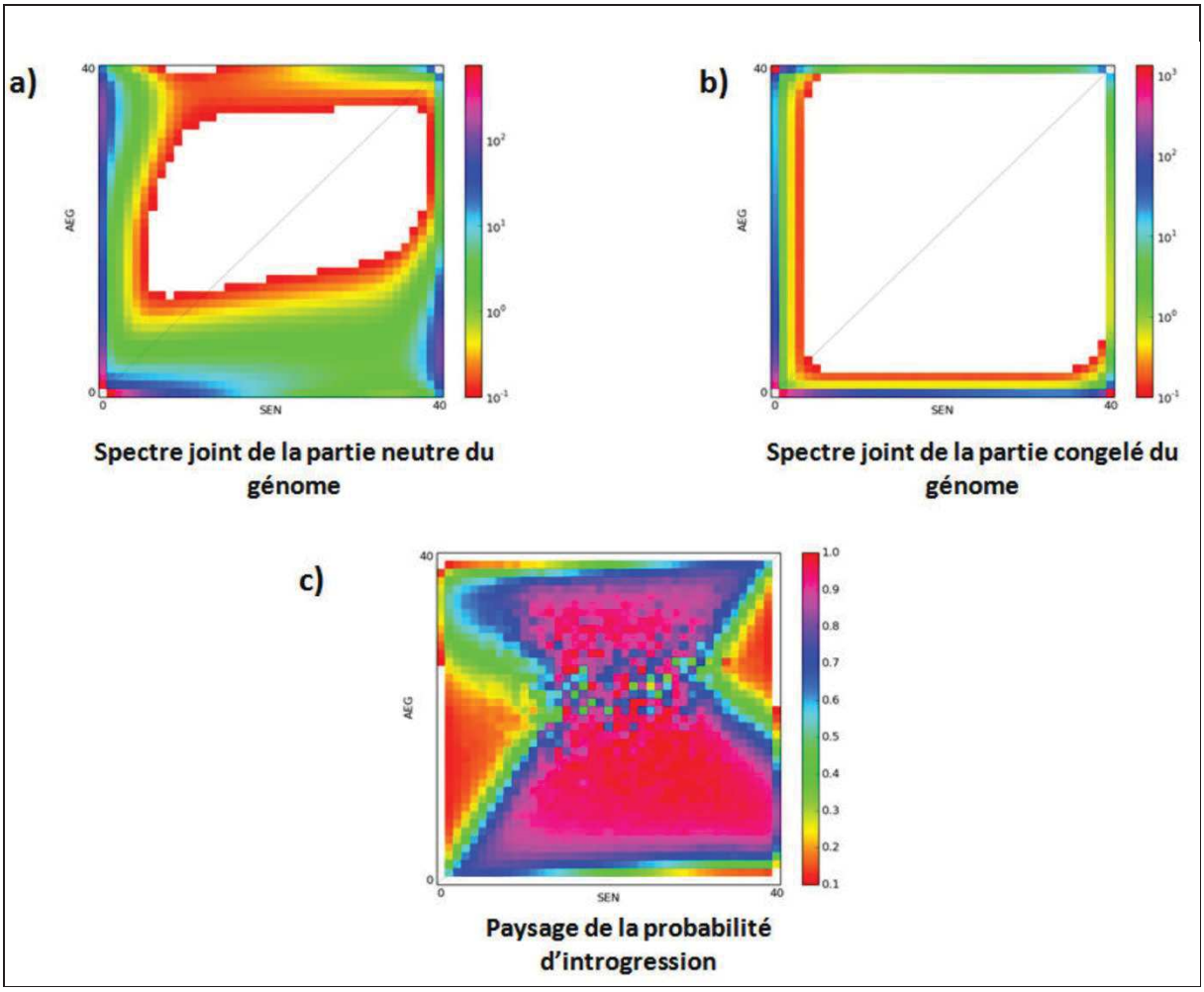


Figure 21 : Spectres joint de fréquences alléliques à partir de données simulées & probabilité d'introgression. (a) Spectre joint des 5% du génome à forte migration efficace, (b) Spectre joint des 95% du génome à faible migration efficace, (c) Paysage de la probabilité d'introgression à travers le génome.

De la même manière, un locus présente un taux d'introgression accru ($\beta < 0$) ou réduit ($\beta > 0$) quand l'intervalle de confiance bilatéral à 95% de la distribution de β ne comporte pas 0. Le deuxième seuil de significativité se base sur l'intervalle défini par les quantiles 0.025 et 0.975 d'une distribution normale de moyenne 0 et de variance estimée à partir du paramètre de précision du cline tau-alpha et tau-beta (inverse des variances de α et β). Ainsi, on qualifie un locus d' « outlier » ou d' « aberrant » quand la probabilité postérieure de la médiane de α ou de β n'est pas incluse dans cet intervalle (Gompert & Buerkle, 2012a; Grossen *et al.,* 2016).

Grâce à cette analyse nous avons obtenu des estimations de paramètres qui varient entre -17.34 et 30.6 pour α et entre -7.8 et 7.09 pour β. Sur le total de 10 756 locus analysés, 5778 locus présentent une valeur négative de α et 4978 locus présentent une valeur positive. Sur ces locus, une majorité montre un excès d'ascendance pour *S. senegalensis* (5190 outliers, dont 4657 détectés avec la méthode des quantiles), tandis qu'une minorité montre un excès d'ascendance pour *S. aegyptiaca* (3222 outliers, dont 1889 détectés avec la méthode des quantiles). Parmi les 10 756 locus analysés, 5550 locus présentent un taux de transition très faible ($\beta < 0$), 5206 un taux très fort ($\beta > 0$), et 5675 locus présentent un taux d'introgression considéré comme aberrant ($\beta < 0$ et $\beta > 0$ cumulés), c.à.d. que la forme de leur cline génomique ne se conforme pas au modèle d'introgression moyen du génome. Ces résultats montrent ainsi une introgression hétérogène entre les locus qui se traduisent par différent types de clines (Figure 22). Pour la suite des analyses (recherches de corrélations avec les autres approches) nous nous sommes intéressés uniquement aux locus détectés comme outliers.

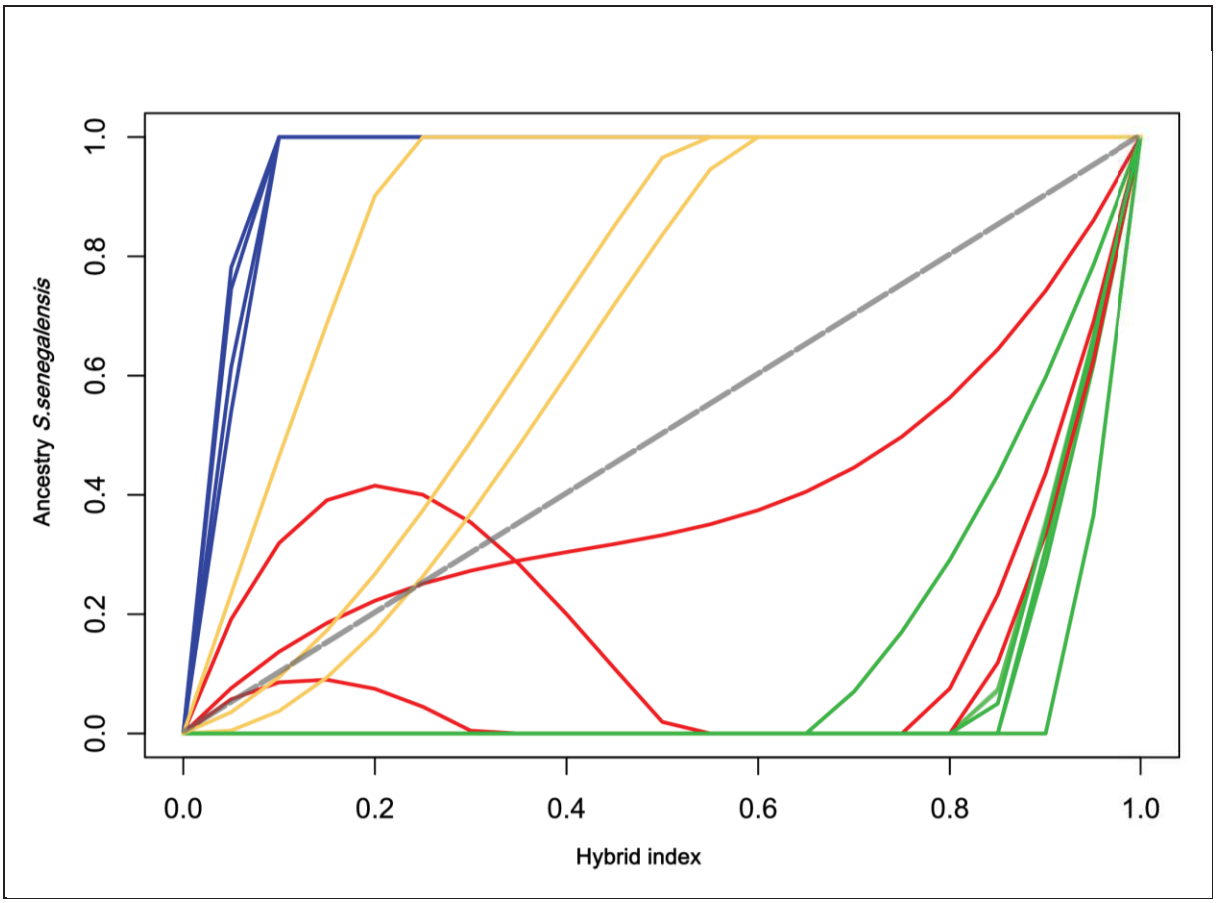


Figure 22 : Détail de clines génomiques de 17 locus à fort taux d'introgression.

4.4. Géographie de l'introgression

En analysant les comportements des différents locus le long de la zone hybride, nous observons deux principaux résultats. Le premier concerne l'estimation du centre moyen des clines qui présente un décalage de 12 km environs vers l'est par rapport à la lagune de Bizerte. Le second résultat issu de cette approche nous indique que les locus ne suivent pas tous le même patron dans la transition de leurs fréquences alléliques (Figure 23). En effet nous avons constaté que plus on s'éloigne du centre estimé (x=12.42km) de la zone hybride plus les valeurs de pente décroissent, c'est à dire que les clines deviennent de moins en moins abruptes (Figure 24).

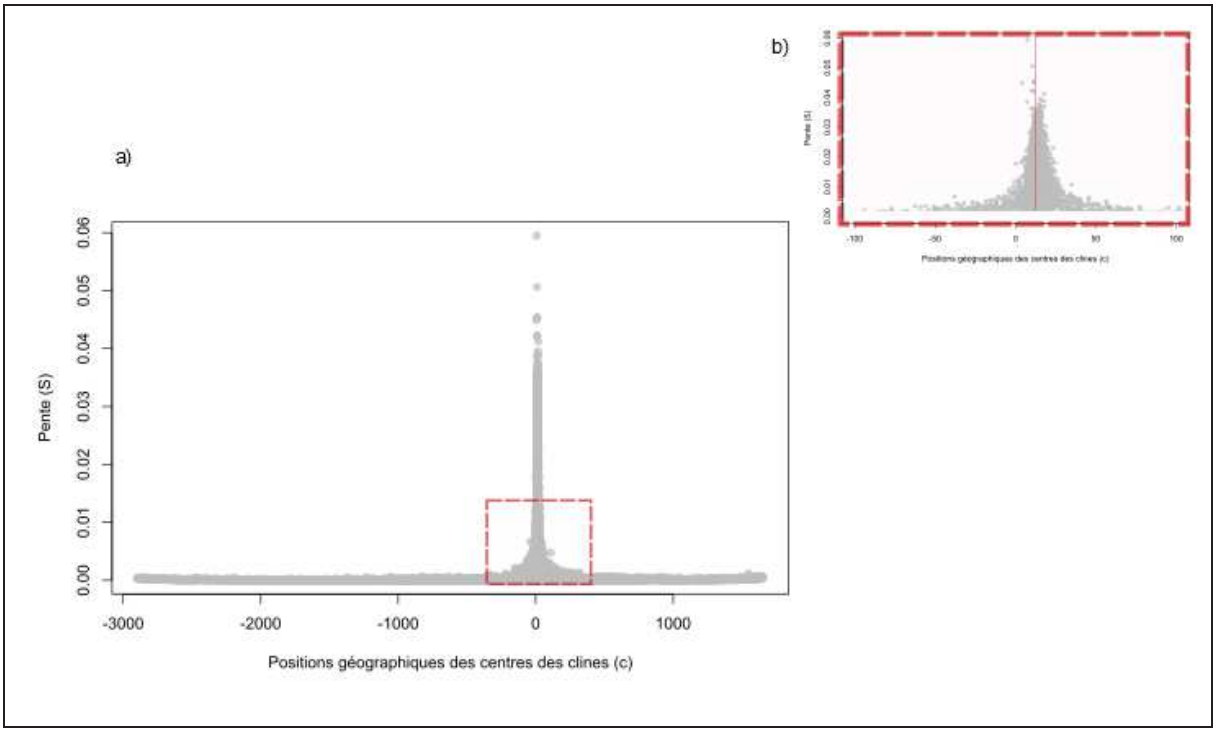


Figure 23 : La pente (S) en fonction des positions des centres des clines géographique (a) & Zoom sur la zone de contact entre S. senegalensis et S. aegyptiaca (b).

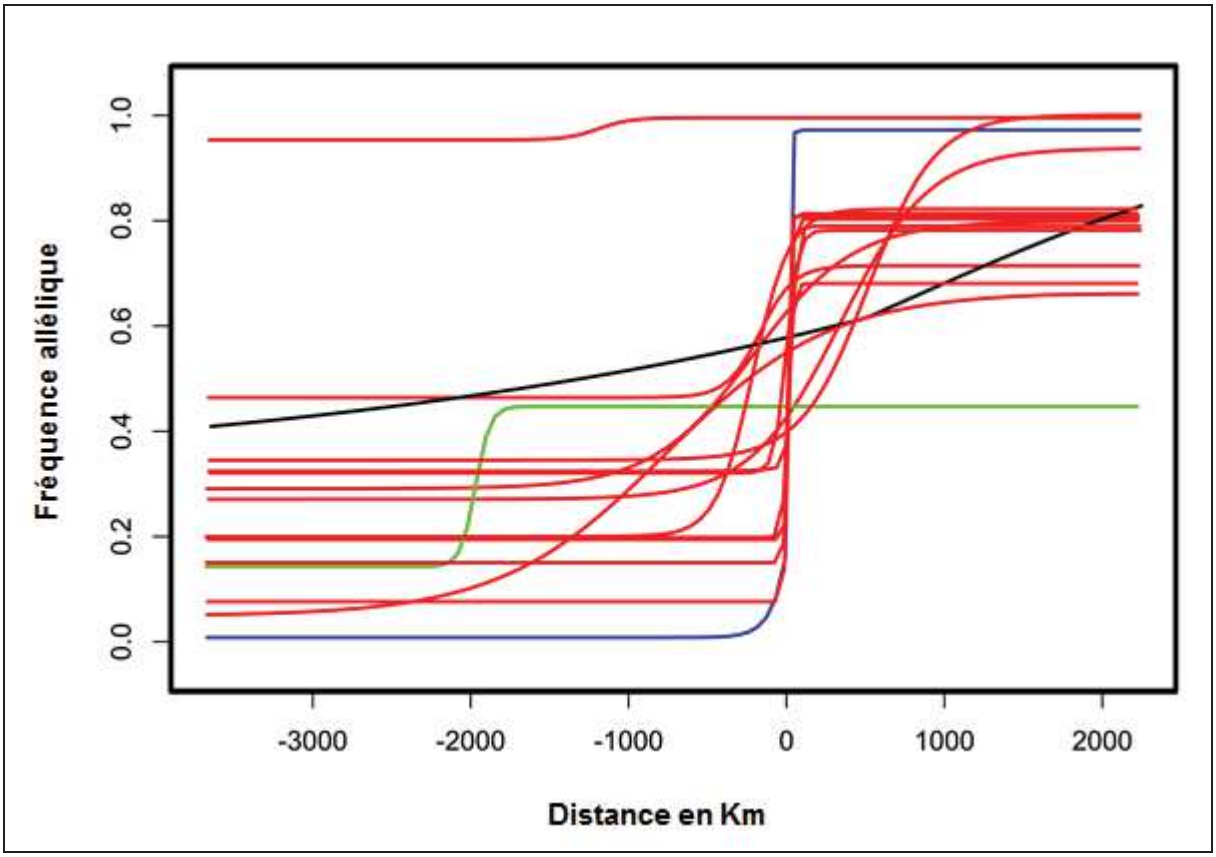


Figure 24 : Détail des clines géographiques de 17 locus à fort taux d'introgression, la ligne bleue correspond au cline ajusté sur la fréquence allélique moyenne des 10 756 SNP.

4.5. Analyse comparative de l'introgression

D'après l'analyse géographique de l'introgression, les locus présentant une forte pente ont des centres localisés essentiellement dans un périmètre de 10km autour du centre de la zone hybride. Pour ces locus, on s'attendrait alors à des profils de transition génomique rapide d'une espèce à une autre (donc une introgression réduite). Afin de vérifier cette prédiction à partir d'une estimation indépendante, nous avons testé la corrélation entre la pente S et les valeurs outlier positives de β (synonyme d'un taux de transition de fréquences allèliques élevé en fonction de l'indice hybride). Nous détectons en effet une corrélation positive entre ces deux paramètres ($R^2_{S\text{-}\beta}$=0.27, $P<10^{-10}$), qui confirme que les locus à clines géographiques plus abruptes ont tendance à être caractérisés par un β. plus fort, et donc une introgression plus faible (Figure 25).

Par la suite nous avons essayé de faire le lien entre l'approche des clines géographiques et l'approche démo-génétique, en se basant sur la pente et la probabilité d'introgression Pn estimée à l'aide de l'approche des spectres joints.

Nous constatons ici que la probabilité d'introgression décroit avec l'augmentation de la pente (Figure 26). Ainsi pour des pentes très fortes (S=0.06) la probabilité d'introgression inférée avec les spectres est nulle (Pn=0), alors que les plus fortes probabilités d'introgression correspondent à des locus à pente faible. Afin de caractériser les clines géographiques et génomiques des locus ayant une forte probabilité d'introgression dans l'approche des spectres joints ($0.95 \leq Pn \leq 1$), nous avons regardé les distributions des paramètres c et α de ces locus. Les résultats obtenus nous ont révélé que la majorité de ces locus ont des centres géographiques situés de part et d'autre de la zone de tension, et dont le décalage spatial est essentiellement marqué du côté aegyptiaca (plus de valeurs de c>12.42) (Figure 27). L'analyse de la distribution du paramètre α révèle également qu'une majorité de ces locus présentent un excès d'ascendance aegyptiaca ($\alpha > 0$) (Figure 28).

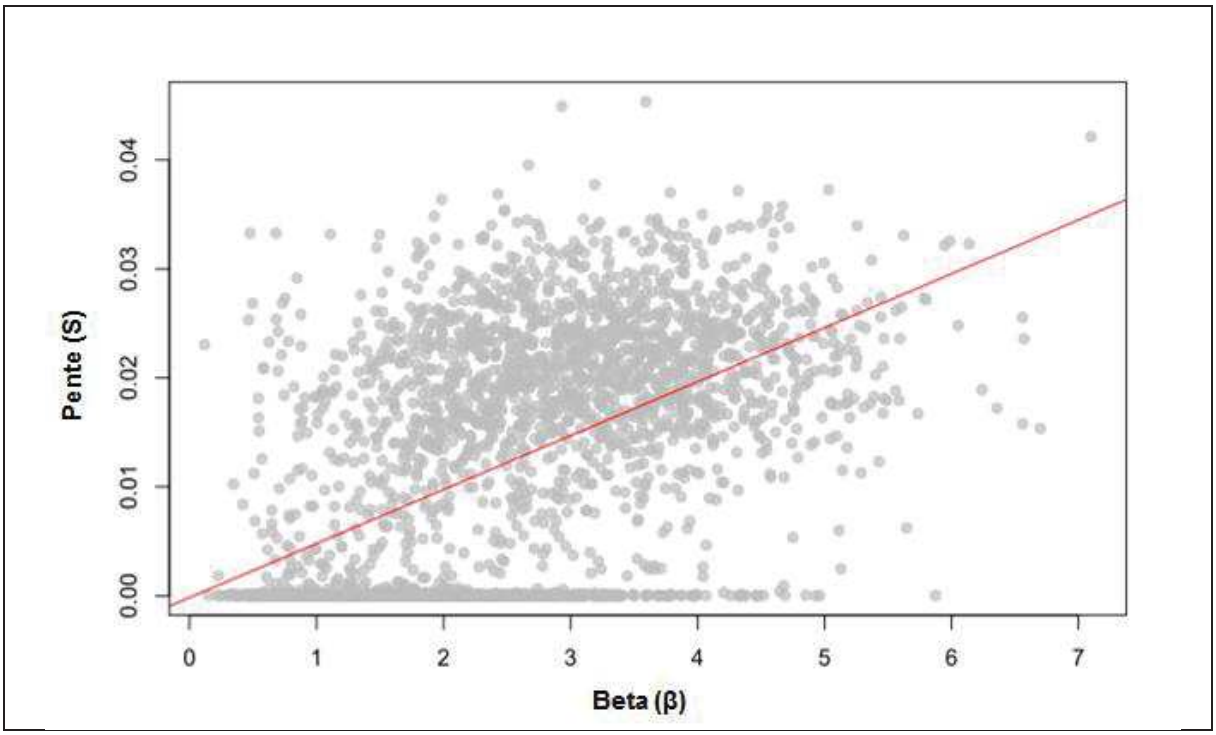


Figure 25 : Corrélation entre le paramètre β et la pente des clines géographiques (S).

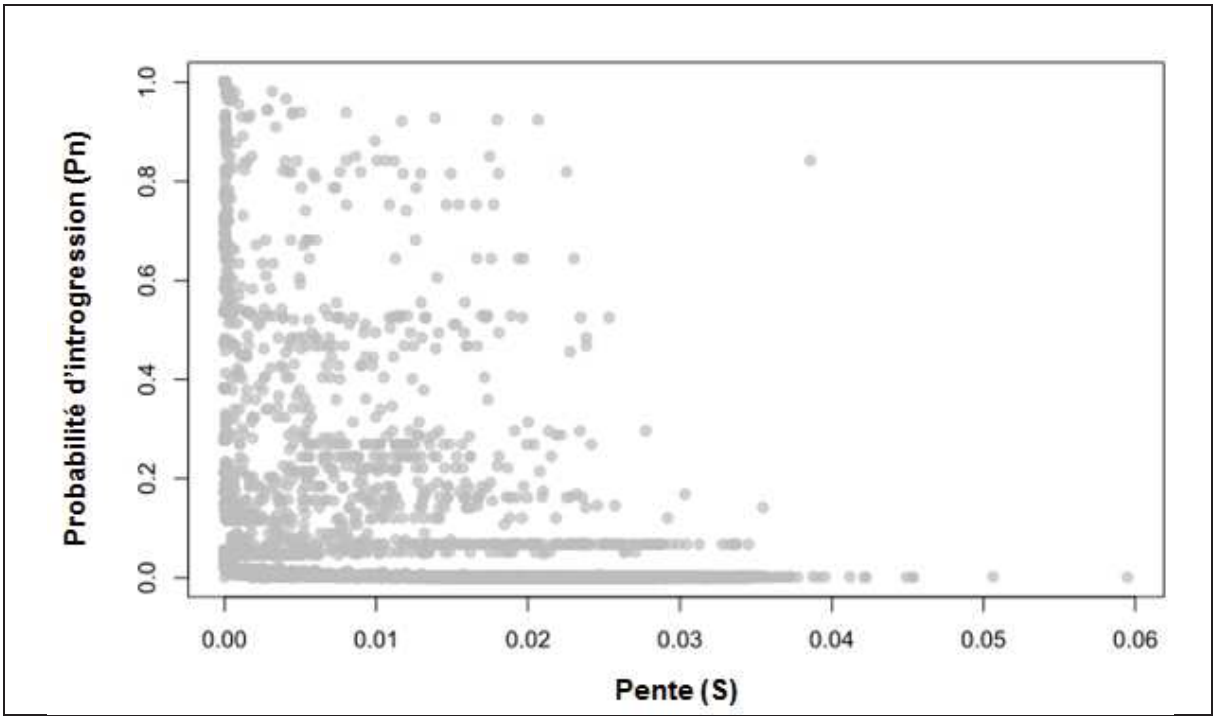


Figure 26 : Corrélation entre la pente des clines géographiques (S) et la probabilité d'introgression (Pn).

Nous évaluons enfin les conséquences de cette asymétrie d'introgression sur la distribution conjointe entre α et c de chaque côté de la zone d'hybridation. Les tests effectués à partir des locus à forte probabilité d'introgression révèlent une corrélation positive 12 fois plus forte entre α et le centre géographique c du côté *S. aegyptiaca* ($R^2_{c-\alpha}$=0.274, $P<10^{-10}$) que du côté *S. senegalensis* ($R^2_{c-\alpha}$=0.023, $P<10^{-10}$) (Figure 29).

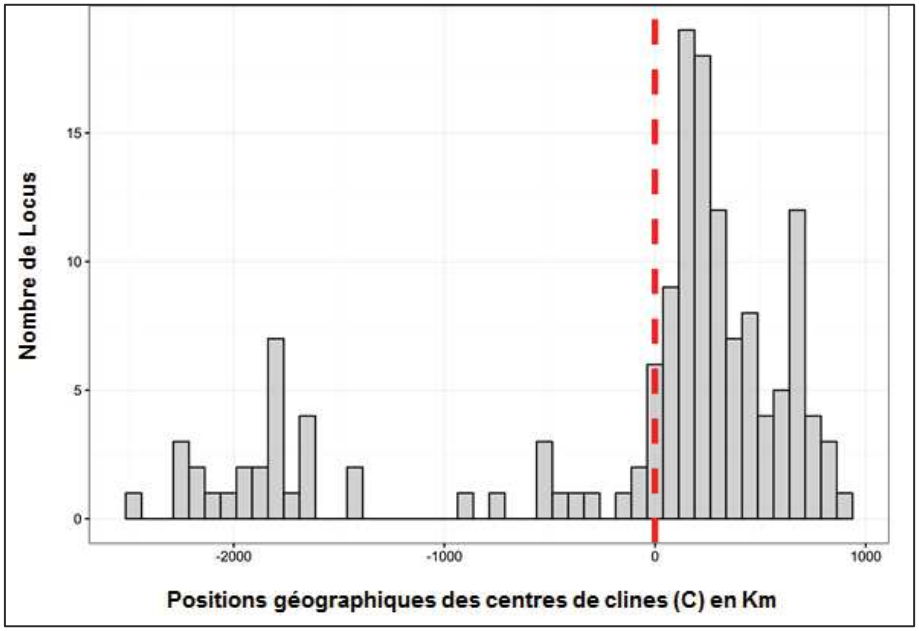


Figure 27 : Distribution des centres des clines géographiques (**c**) pour les locus dont la probabilité d'introgression est très forte (**Pn**>95).

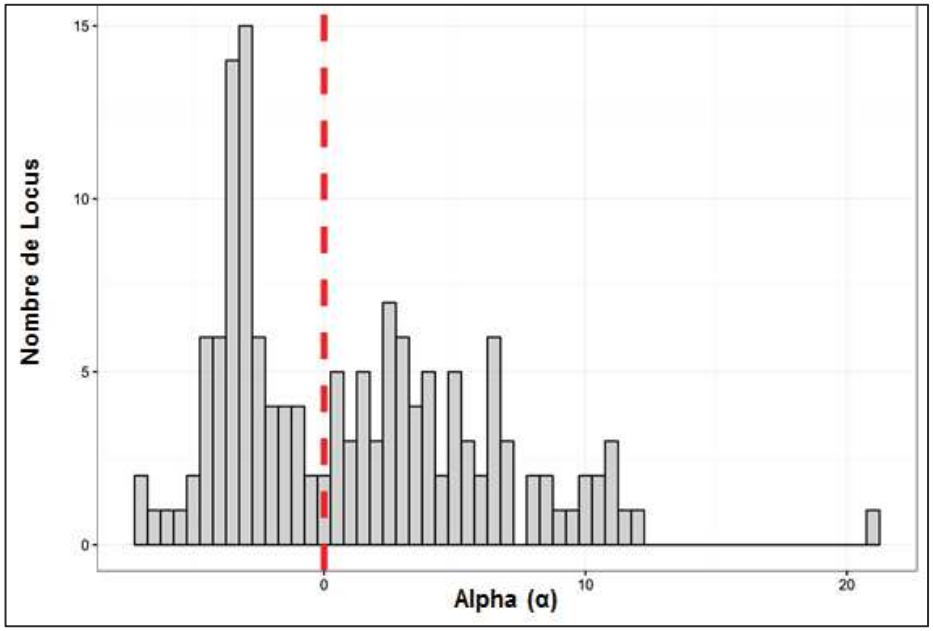


Figure 28 : Distribution du paramètre **α** pour les locus dont la probabilité d'introgression est très forte (**Pn**>95).

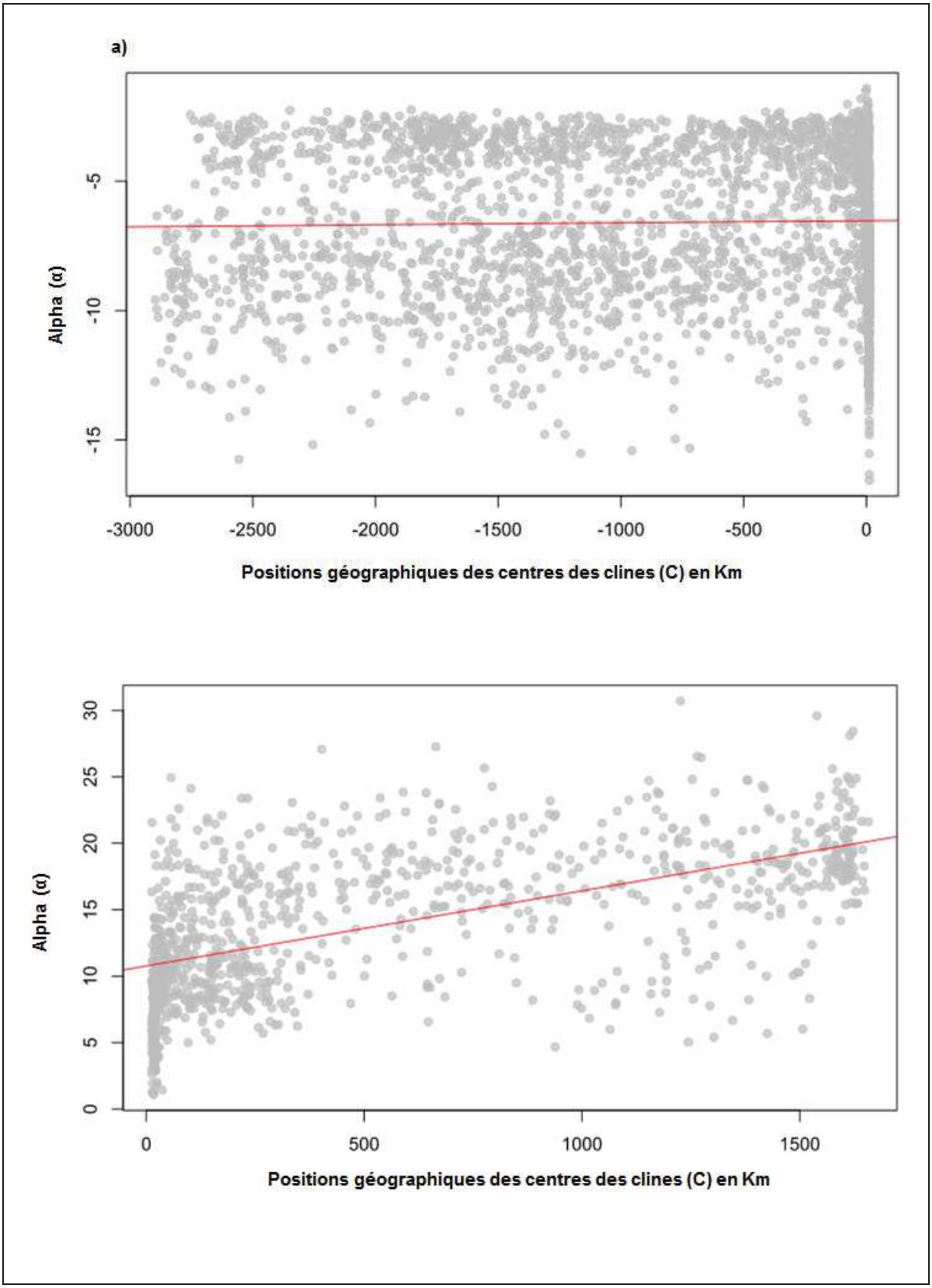


Figure 29 : Corrélation entre le paramètre α et la position en Km des centres (c) des clines géographiques pris de part et d'autre de la zone hybride. (a) Fond génétique *S. senegalensis* (α négatif) & aire de répartition géographique *S. senegalensis,* (b) fond génétique *S. aegyptiaca* (α positif) & aire de répartition géographique *S. aegyptiaca.*

5 Discussion

Dans notre présente étude décrivant la zone hybride entre *S. senegalensis* et *S. aegyptiaca* le premier résultat marquant est l'absence de génotypes hybrides de première génération Fig. En effet, les génotypes F1 sont théoriquement caractérisés par un indice hybride de 0,5 et une hétérozygotie interspécifique forte (égale à 1 pour des marqueurs diagnostiques). Or ici, aucun individu présentant un indice hybride proche de 0.5 n'a été détecté. En revanche, une 6 individus dans nos populations échantillonnées s'avèrent être compatibles avec des génotypes de type backcross. Ces individus dont l'indice hybride est proche de 0.875 correspondent très probablement à des backcrosses de seconde génération dans le fond génétique *S. aegyptiaca*, et ont été trouvés dans le Golfe de Tunis. En dehors de ces hybrides de générations précoces, nos résultats indiquent également la présence d'individus introgressés, correspondant des génotypes hybrides de générations tardives (backcrosses de nièmes générations). Ces résultats nous informent sur la présence d'un flux génique entre espèces qui se réalise par suite à de rares événements d'hybridation. Cette situation de déficit en hybrides associée à la présence d'individus introgressés dans chaque fond génétique est en accord avec le modèle classique des zones de tensions, où le flux génétique (et donc l'introgression) entre espèce est assuré via l'intermédiaire des croisements retours (backcross) entre les rares hybrides F1 et les génotypes parentaux de part et d'autre de la zone de contact. Nos résultats diffèrent donc de ce qui a était rapporté par Ouanes et al en 2011. En effet ces auteurs, en plus de relever un taux moyen d'hybridation élevé qui avoisine les 40% dans la lagune de Bizerte, ont observé un vrai continuum de génotypes hybrides qui s'étend de la lagune de Bizerte jusqu'au Golfe de Tunis, alors que dans la présente étude qu'aucun continuum spatial dans la distribution des indice hybrides n'a été observé (on observe plutôt une transition abrupte d'une catégorie de génotypes parentaux vers l'autre, entre Bizerte et Golfe de Tunis). Cette différence entre ces deux études pourrait être liée aux différentes approches génétiques utilisées. En effet,

contrairement au précédent travail de (Ouanes *et al.,* 2011), où l'analyse de la zone hybride ne portait que sur quelques marqueurs diagnostiques (2 locus allozymes et 4 locus nucléaires), la précédente étude repose sur plusieurs milliers de locus diagnostiques et quasi-diagnostiques. Or seule une approche génomique reposant sur un nombre suffisant de marqueurs génétiques permet de distinguer sur la base des seules fréquences alléliques l'hybridation de l'introgression. Ceci est dû (i) à la nécessité de mesurer l'indice hybride (ou la valeur d'admixture calculée par structure) à l'aide de nombreux locus pour obtenir une estimation statistiquement robuste, (ii) au fait que les locus identifiés comme diagnostiques entre les populations parentales ne sont pas forcément liés aux locus d'isolement reproductif (ils peuvent donc potentiellement introgresser). Pour illustrer ce second point, considérons un individu non-hybride de l'espèce *S. aegyptiaca* mais introgressé sur 50 % de son génome tout en étant homozygote pour des allèles aegyptiaca aux locus d'isolement. Un tel individu présenterait un indice hybride de 0,5 (il serait donc potentiellement confondable avec un hybride F1) s'il était calculé avec des locus susceptibles d'introgresser, mais un indice hybride de 1 s'il était calculé à l'aide de marqueurs liés aux gènes d'isolement reproductif. En effet l'analyse d'une zone hybride dépend de l'état de la variation génétique examinée à l'égard de l'isolement reproductif. Ainsi seuls les gènes d'isolement reflètent réellement l'hybridation alors que le suivi des allèles susceptibles d'introgresser permet de décrire plutôt l'introgression (Bierne *et al.,* 2003). Dans la présente étude, une fraction bien plus importante du génome a pu être analysée par rapport à l'étude précédente. Nous sommes donc en mesure d'inclure dans le calcul de l'indice hybride individuel une certaine proportion de locus liés aux gènes d'isolement, qui sont les seuls capables de nous renseigner sur le statut hybride ou introgressé d'un individu. Or d'après les analyses δaδi, le génome des soles est composé d'une faible proportion de locus qui sont susceptibles d'introgresser (environ 5 %), contre une majorité (95 %) de locus fixés différentiellement entre les deux espèces. Ceci suggère qu'on nombre relativement limité de

locus diagnostiques serait suffisant pour distinguer les génotypes hybrides et introgressés chez les soles. Cependant, le nombre de locus utilisés par (Ouanes *et al.,* 2011) n'est peut-être pas suffisant pour permettre cette distinction, d'où les incohérences apparentes avec notre étude. Ainsi, les précédents résultats obtenus avec un faible nombre de marqueurs restent cohérents avec une faible fréquence d'individus hybrides et une forte fréquence d'individus introgressés, confondus sous le terme générique d'hybrides dans la précédente étude mais clairement distingués ici.

Ces résultats, ainsi que ceux de l'analyse des clines géographiques qui montrent que la plupart des clines à pente forte possèdent des centres superposés au même endroit, témoignent que nous sommes en présence d'une zone hybride conforme au modèle de zone de tension (Barton & Hewitt, 1985). De plus, nos analyses de l'histoire démo-génétique confirment ce qui avait été suggéré dans la précédente étude sur ces deux espèces, que cette zone de tension résulte de la confrontation lors d'un contact secondaire de deux pools géniques qui ont divergé en allopatrie. L'accumulation de barrières d'isolement pendant la phase d'allopatrie expliquerait alors le maintien actuel de la zone par un équilibre entre la contre-sélection des hybrides et la dispersion des génotypes parentaux dans la zone de contact. Nous avons en effet relevé la présence de quatre individus ayant des génotypes *S. aegyptiaca* quasiment purs dans la lagune de Bizerte, individus qui sont probablement des migrants parentaux. Par ailleurs, les résultats de l'analyse démo-génétique nous indiquent que la durée de la séparation allopatrique entre les deux espèces a été beaucoup plus longue que la durée du contact secondaire à l'origine de cette zone de tension, permettant ainsi à différentes incompatibilités génétiques de se développer entre *S. senegalensis* et *S. aegyptiaca*. Ces incompatibilités créeraient alors des barrières au flux génique depuis la remise en contact des deux espèces. Cette zone d'hybridation qui semble être caractérisée par une forte contre-sélection des génotypes recombinants ressemble à d'autres cas rapportés chez plusieurs organismes comme les souris (Teeter *et al.,* 2010), les pics (Grossen

et al., 2016), les peupliers (Stölting *et al.,* 2013; Christe *et al.,* 2016), les moules (Bierne *et al.,* 2003) et les ciones (Roux *et al.,* 2013). Ces différents modèles d'étude sont caractérisés par différents niveaux d'isolement reproductif se traduisant par une perméabilité plus ou moins forte des barrières au flux génique. Considérant le fait que nous estimons ici qu'une majeure partie du génome présente un taux d'introgression très réduit (environ 95%), le cas des soles semble indiquer une perméabilité relativement faible au flux génique. Malgré cette forte imperméabilité, notre reconstruction démo-génique nous indique d'une faible proportion de locus (5%) parviennent à traverser la barrière entre espèces. Cette situation semble être très similaire au cas de la zone hybride des ciones. En effet, (Roux *et al.,* 2013) décrivent la zone de contact entre *Ciona intestinalis* et *C. robusta* comme étant le résultat d'un contact secondaire récent entre deux espèces très divergentes. Depuis le début du contact, les taux d'introgression sont très hétérogènes entre locus et seules de rares régions du génome parviennent à traverser la barrière entre espèces. Dans ce type de situation, il est alors intéressant d'étudier en particulier le comportement individuel des quelques locus qui parviennent à traverser la barrière entre espèces, afin de mieux comprendre les raisons de cette introgression.

Afin de mieux décrire le fonctionnement de la zone de tension entre *S. senegalensis* et *S. aegyptiaca*, nous avons adopté une approche intégrative combinant modélisation démo-génétique, analyse des clines de fréquence géographique et des clines génomiques. Chacune de ces méthodes se focalise sur un aspect différent des données. En les combinant, nous avons donc pu caractériser des aspects complémentaires de la dynamique des flux de gènes entre *S. senegalensis* et *S. aegyptiaca*, au centre et en périphérie de la zone de contact. Au cœur d'une zone de tension, à cause de la contre-sélection des génotypes hybrides, les transitions d'un pool génique à un autre se font rapidement en engendrant un fort déséquilibre de liaison et en donnant un aspect abrupt aux clines de fréquences alléliques. Ces clines à fortes pentes forment alors une barrière peu perméable au flux génique s'ils coïncident au même endroit. En effet dans ce

cas de figure chaque locus cumule l'effet sélectif des autres locus en plus de son propre effet sélectif, et ce de façon proportionnelle au déséquilibre de liaison. Ainsi le génome en entier se comporterait dans les cas les plus extrêmes comme un seul locus sous sélection, on parle alors dans ce cas de génome congelé (Turner, 1967; Kruuk *et al.,* 1999). Dans la présente étude, les locus à cline abrupte, dont le patron spatial reflète l'effet d'une barrière au flux génique, font probablement partie des 95% du génome à taux d'introgression faible prédits par l'approche démo-génétique. En effet nous avons constaté qu'une majorité de clines géographiques à pente forte et à faible probabilité d'introgression se localisent dans un périmètre d'une vingtaine de kilomètres aux alentours de la zone de contact. Pour ces clines, nous avons également détecté une corrélation positive entre la pente et les valeurs positives outlier de β, le paramètre qui décrit l'effet barrière au flux génique entre espèces. Des valeurs positives du paramètre β sont généralement attendues soit dans le cas d'une sous-dominance, indiquant alors une contre-sélection des génotypes recombinants, soit dans le cas de populations structurées au sein de la zone hybride ou encore à cause de l'homogamie résultant d'un possible choix de partenaire (Gompert *et al.,* 2012b). Contrairement à ce qui a été observé chez les manakins (petit passereau) (Parchman *et al.,* 2013), où les fortes valeurs de β seraient dues à la structure des populations dans la zone hybride, nous sommes dans un cas qui se rapproche plus de celui des mésanges nord-américaines (Taylor *et al.,* 2014), où la dispersion est forte et la contre sélection des hybrides serait probablement le mécanisme responsable des valeurs fortes de β. Cependant l'hypothèse de l'homogamie reste aussi probable. Ainsi, la corrélation positive entre les valeurs extrêmes positives de β et la pente des clines géographiques indique probablement des niveaux de contre sélection plus ou moins forts entre différentes régions du génome. Cette variance peut être attribuée soit aux effets individuels variables des locus d'isolement, soit au déséquilibre de liaison plus ou moins fort qui existe entre eux. Par exemple, la densité locale de locus impliqués

dans l'isolement au sein d'une région génomique donnée pourrait influencer à la fois la pente des clines géographiques et le paramètre β des clines génomiques.

Parallèlement à ce qui a été observé au cœur de la zone de contact, les paramètres des clines en périphérie de cette zone nous informent sur des aspects complémentaires du fonctionnement de la zone de tension. Par exemple, les clines géographiques dont les centres sont décalés, soit du côté *S. senegalensis* soit du côté *S. aegyptiaca*, sont de moins en moins abruptes en fonction de l'éloignement de leur centre par rapport au cœur de la zone. Cette observation est cohérente avec la prédiction qu'un locus dont le centre est décalé ne bénéficie pas du cumul des effets sélectifs des autres locus transmis via le déséquilibre de liaison. Sa pente est donc logiquement plus faible. De plus, nos résultats montrent un lien entre la position du centre des clines et la probabilité d'introgression déterminée via l'approche démo-génétique. En effet, les locus à forte probabilité d'introgression qui sont donc inclus dans les 5 % des locus qui introgressent le plus présentent majoritairement des centres décalés. La distribution des valeurs des centres pour ces locus montre en outre un décalage plus marqué du coté de *S. aegyptiaca*. Ces mêmes locus montrent une distribution bimodale du paramètre α des clines génomique, avec des valeurs souvent différentes de zéro et montrant donc un excès d'ascendance pour l'une ou l'autre des espèces. Les approches des clines géographiques et génomiques détectent donc toutes les deux des valeurs singulières des paramètres de clines pour les locus à forte probabilité d'introgression. Le décalage spatial des centres des clines est positivement corrélé à un excès d'ascendance de chaque côté de la zone de contact. Cependant, ici encore cette corrélation est bien plus marquée pour les locus dont l'introgression est plus forte dans le sens *S. senegalensis* vers *S. aegyptiaca* que l'inverse. Ces résultats confirment donc les prédictions faites par δaδi, qui suggéraient que le flux de gène dans les régions introgressées est douze fois plus fort dans cette direction. Ces résultats obtenus avec différentes approches nous indiquent donc que malgré l'existence d'un isolement reproductif fort, des flux géniques persistent entre *S.*

senegalensis et *S. aegyptiaca* dans certaines régions génomiques, et ce avec une direction d'introgression préférentielle qui est opposée à la direction principale détectée sur la majeure partie du génome.

Les causes évolutives de cette introgression et de l'asymétrie qui la caractérise peuvent être multiples. Les patrons spatiaux des flux géniques détectés en dehors et au cœur de la zone de tension peuvent alternativement résulter soit d'un contact très récent soit d'une dislocation en cours de la barrière, soit d'un possible mouvement de la zone d'hybridation. Par exemple, chez les crevettes du genre *Paratya*, il a été documenté un mouvement de la zone hybride de quelques kilomètres qui serait probablement lié soit à l'asymétrie de l'introgression soit à la translocation de la barrière chez *Paratya australiensis* (Wilson *et al.,* 2016).

Conclusion

Contrairement à la plupart des zones d'hybridation terrestres où les patrons spatiaux de fréquence allélique sont étudiés à l'endroit même où se produisent la reproduction et la dispersion, comme dans le cas de la zone hybride entre les souris *Mus musculus* et *Mus domesticus*, la zone hybride étudiée ici entre *Solea senegalensis* et *Solea aegyptiaca* correspond plutôt à une transposition spatio-temporelle de ce qui se passe lors de la reproduction. En effet, les individus ne peuvent être échantillonnés aisément que sur les sites de nourrissage que sont le milieu côtier et les lagunes, et la localisation des zones de reproduction est largement inconnue. Ce cas de figure n'est pas unique en milieu marin, et rajoute une couche de complexité supplémentaire chez les espèces migratrices, qu'il s'agisse de migrations saisonnières comme chez la sole ou de ces plus extrêmes comme chez les anguilles atlantiques. A la fois chez l'anguille américaine (*Anguilla rostrata*) et l'anguille européenne (*Anguilla anguilla*), la reproduction n'a lieu qu'à un seul endroit, la Mer des Sargasses. Le partage de cette zone de reproduction par les deux espèces et le recoupement partiel de leurs périodes de reproduction est à l'origine de leur hybridation (Als *et al.,* 2011). Pourtant, à l'issue de leur longue phase de dispersion larvaire qui emprunte les courants du Gulf Stream, les civelles (nom donné aux post-larves d'anguilles) d'*Anguilla rostrata* recrutent en Amérique du nord, alors que celles de l'espèce *Anguilla anguilla* recrutent en Europe, et que les individus hybrides dont le timing de métamorphose est intermédiaire recrutent préférentiellement en Islande. La distribution spatiale des stades juvéniles et adultes de ces deux espèces et de leurs hybrides dans les eaux continentales américaines et européenne correspond donc à la transposition spatiale des croisements intra et interspécifiques qui se produisent dans la Mer des Sargasses. La ségrégation spatiale des génotypes lors de leur recrutement s'explique probablement par des

durées de migration larvaire qui diffèrent fortement entre l'anguille américaine et européenne, et dont le déterminisme en partie génétique serait à l'origine d'une durée de vie larvaire intermédiaire chez les hybrides.

Chez la sole, du fait de la courte période de la phase larvaire (entre 17 et 19 jours par exemple chez *Solea senegalensis*) (Bedoui, 1995; Yúfera *et al.,* 1999) ainsi que du mode de vie benthique des adultes, il est fortement improbable qu'une situation semblable à l'anguille existe. Cela supposerait l'existence d'une seule et unique zone de reproduction à partir de laquelle les deux espèces disperseraient à travers la méditerranée, alors que les hybrides ne recruteraient qu'au large des côtes nord Tunisiennes. Il est donc plus probable que, comme chez plusieurs espèces de poissons marins, il existerait pour *S. senegalensis* et *S. aegyptiaca* une succession de sites de ponte le long de leurs aires de distribution. Ainsi en Méditerranée ouest, territoire de *S. senegalensis*, il existe probablement plusieurs aires de ponte *S. senegalensis*, alors qu'en Méditerranée Est, plusieurs aires de ponte existent également pour *S. aegyptiaca*. En ce qui concerne les individus hybrides que nous observons dans la lagune de Bizerte et le golf de Tunis, il est difficile de départager l'hypothèse d'une seule zone de reproduction locale qui alimenterait les nurseries côtières de cette région, de l'hypothèse de deux zones de reproduction différentes pour chaque espèce mais qui se chevaucheraient spatialement ou temporellement. Comme la fréquence des hybrides observés est faible, le cas d'une seule aire de ponte régionale commune avec appariement aléatoire des adultes est peu probable car on s'attend dans ce cas à un fort taux d'hybridation. Même si dans notre cas le faible nombre d'individus hybrides observés pourrait être la conséquence d'un fort effet délétère, il est difficile d'envisager comme un tel système générateur d'hybrides non viables pourrait être maintenu en raison de son coût évolutif. Alternativement, la rareté des hybrides pourrait être simplement la conséquence aussi bien d'un choix de partenaire que d'un décalage temporel de la ponte, qui maintiendrait ainsi un isolement prézygotique entre les deux espèces. Concernant la deuxième hypothèse de deux

aires de ponte séparées (ou se chevauchent partiellement) au large de la lagune de Bizerte, elle ne peut être rejetée vu la fréquence des cas de philopatrie reproductive décris chez les poissons.

Malgré les limites de nos connaissances sur les composantes spatiales des zones de reproduction et de recrutement de *S. senegalensis* et *S. aegyptiaca*, nous avons essayé dans ce présent travail d'étudier les relations entre génotype et phénotype au sein de cette zone hybride, ainsi que de reconstituer l'histoire du contact entre ces deux espèces et la dynamique des flux génique à travers la barrière d'espèces. Ainsi, dans la première partie de cette thèse, nous avons évalué les conséquences des échanges génétiques interspécifiques sur la forme des poissons, et ce en fonction des combinaisons allèliques observées au centre de la zone hybride. En plus d'un effet de dilution de certains traits divergents dans la zone de contact, nous avons pu mettre en évidence la présence de transgressions phénotypiques sur d'autres traits chez les individus les plus introgressés, qui serait en étroite relation avec des effets épistatiques et pléiotropies des gènes impliqués dans la construction de ces traits. D'autres conséquences morphologiques de l'hybridation ont été détectées chez les hybrides, et semblent indiquer une condition plus faible de ces génotypes par rapport aux génotypes parentaux, reflétant ainsi possiblement les effets délétères de l'hybridation entre des génomes divergents. Dans la deuxième partie de cette thèse, nous nous sommes intéressés à caractériser l'histoire des échanges génétiques au travers de la zone hybride, qui s'est révélée être en lien avec un contact secondaire relativement récent par rapport au temps de divergence entre les deux espèces. Ainsi les soles constituent un nouveau cas de spéciation original qui viens s'ajouter à différentes autres espèces. Ces organismes, pour la plupart se caractérisent soit par un contact récent après une divergence faible comme chez le bar et les corégones soit par un contact ancien qui suit une divergence plus ancienne chez les moules par exemple (Figure 30).

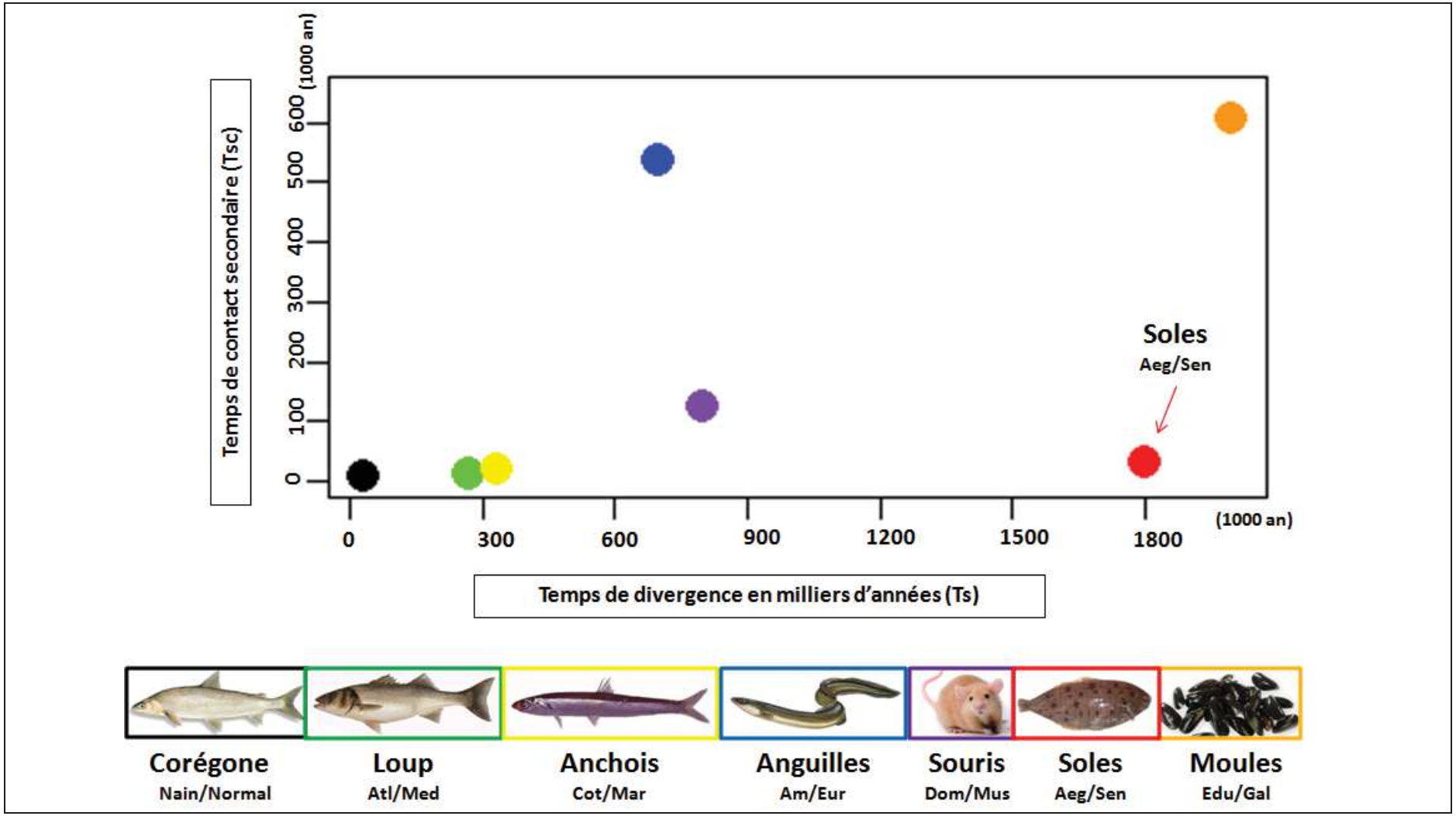


Figure 30 : Temps de divergence et temps de remise en contact en milliers d'années de certaines paires d'espèce étudiées au sein du laboratoire.

Le contact secondaire récent entre *S. senegalensis* et *S. aegyptiaca* apporte de nouvelles perspectives dans l'étude de la spéciation. En effet, la plupart des études qui s'intéressent à la génomique de la spéciation, comparent des organismes partiellement isolés à la recherche de régions génomiques étanches au flux génétique. Ces zones particulières du génome sont susceptibles de cacher les gènes de spéciation responsables des prémices de l'isolement reproductif. Cependant, comme ces régions sont généralement larges au niveau du génome, ces gènes restent dans la plupart des cas introuvables. Alors que chez ces deux espèces de soles dont la divergence longue est synonyme d'une spéciation presque finie (95% du génome est congelé), un contact secondaire récent représente une opportunité d'entreprendre les questions liées à la génomique de la spéciation avec un nouveau regard, c'est-à-dire en se focalisant non pas sur les gènes de spéciations mais plutôt sur les régions qui introgressent. Ainsi en présence d'un génome bien assemblé il serait intéressant d'aller chercher ces gènes de connaitre leur nature et leur distribution spatiale au niveau du génome.

Dans la continuité des travaux entrepris par She en 1987 et Ouanes en 2011 sur la zone hybride des soles, notre présente étude avait pour objectif de revisiter les précédents résultats en utilisant une nouvelle stratégie d'échantillonnage plus focalisés sur la zone de contact et de nouvelles techniques moléculaires de génotypage-par-séquençage. Nos résultats ont permis d'éclaircir quelques zones d'ombres liées à la biologie et à l'histoire de la spéciation entre *S. senegalensis* et *S. aegyptiaca*, et ont montré l'originalité de modèle d'étude dont le contact est très récent par rapport à la durée de la divergence. Cependant, plusieurs autres questions relatives à cette zone hybride demeurent encore sans réponse notamment où se situent les zones de reproduction de ces deux espèces ? S'agit-il d'une zone commune ou de plusieurs zones de reproduction ? Comment sont organisées les zones à fort taux d'introgression dans le génome ? Quelle est la nature des gènes partagés entre ces deux espèces et dans quelles fonctions sont-ils impliqués ?

Pourquoi ils ont une tendance à introgresser dans une direction opposée à l'ensemble du génome ? comment évoluerait cette zone dans les décennies à venir ?

Bibliographie

Abbott, R., D. Albach, S. Ansell, J. W. Arntzen, S. J. Baird, N. Bierne, J. Boughman, et al. 2013. «Hybridization and speciation.» *J Evol Biol* 26: 229-46.

Adams, D. C., et E. Otarola-Castillo. 2013. «geomorph: an r package for the collection and analysis of geometric morphometric shape data.» *Methods Ecol. Evol.* 4: 393-399.

Albertson, R. C., et T. D. Kocher. 2005. «Genetic architecture sets limits on transgressive segregation in hybrid cichlid fishes.» *Evolution* 59: 686-90.

Albrecht, G. H. 1980. «Multivariate-Analysis and the Study of Form, with Special Reference to Canonical Variate Analysis.» *Am. Zool.* 20: 679-693.

Als, T. D., M. M. Hansen, G. E. Maes, M. Castonguay, L. Riemann, K. Aarestrup, P. Munk, H. Sparholt, R. Hanel, et L. Bernatchez. 2011. «All roads lead to home: panmixia of European eel in the Sargasso Sea.» *Molecular Ecology* 20: 1333-1346.

Arnold, M. L., et S. A. Hodges. 1995. «Are Natural Hybrids Fit or Unfit Relative to Their Parents.» *Trends in Ecology \& Evolution* 10: 67-71.

Backstrom, N., G. P. Saetre, et H. Ellegren. 2013. «Inferring the demographic history of European Ficedula flycatcher populations.» *BMC Evol. Biol.* 13: 2. http://www.ncbi.nlm.nih.gov/pubmed/23282063.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver, Z. A. Lewis, E. U. Selker, W. A. Cresko, et E. A. Johnson. 2008. «Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers.» *Plos One* 3: e3376.

Barton, N. H. 1979. «The dynamics of hybrid zones.» *Heredity* 43: 341-359.

Barton, N. H. 1980. «The Hybrid Sink Effect.» *Heredity* 44: 277-278.

Barton, N. H., et G. M. Hewitt. 1985. «Analysis of Hybrid Zones.» *Annual Review of Ecology and Systematics* 16: 113-148.

Barton, N. H., et K. S. Gale. 1993. «Genetic Analysis of Hybrid Zones.» Dans *Hybrid Zones and the Evolutionary Process*, édité par Richard G. Harrison, 14-45. New York, New York, USA: Oxford University Press.

Barton, N., et B. O. Bengtsson. 1986. «The Barrier to Genetic Exchange between Hybridizing Populations.» *Heredity* 57: 357-376.

Beaumont, M. A., W. Zhang, et D. J. Balding. 2002. «Approximate Bayesian computation in population genetics.» *Genetics* 162: 2025-35.

Bedoui, R. 1995. «Elevage de Solea senegalensis (Kaup, 1958) en Tunisie.» *Marine aquaculture finfish species diversification. Proceedings of the seminar of the CIHEAM Network on Technology of Aquaculture in the Mediterranean (TECAM), Nicosia (Cyprus).* 14-17.

Belkhir, K., P. Borsa, L. Chikhi, N. Raufaste, et F. Bonhomme. 1996. «GENETIX 4.05, Logiciel Sous Windows TM Pour Al Genetique De Populations.»

Bell, M. A., et M. P. Travis. 2005. «Hybridization, transgressive segregation, genetic covariation, and adaptive radiation.» *Trends Ecol. Evol.* 20: 358-61.

Bierne, N., J. Welch, E. Loire, F. Bonhomme, et P. David. 2011. «The coupling hypothesis: why genome scans may fail to map local adaptation genes.» *Mol. Ecol.* 20: 2044-72. http://www.ncbi.nlm.nih.gov/pubmed/21476991.

Bierne, N., P. A. Gagnaire, et 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.

Bierne, N., P. Borsa, C. Daguin, D. Jollivet, F. Viard, F. Bonhomme, et P. David. 2003. «Introgression patterns in the mosaic hybrid zone between Mytilus edulis and M-galloprovincialis.» *Molecular Ecology* 12: 447-461.

Bierne, Nicolas, Thomas Lenormand, François Bonhomme, et Patrice David. 2002. «Deleterious mutations in a hybrid zone: can mutational load decrease the barrier to gene flow?» *Genet. Res.* 80: 197-204.

Borsa, P., et J. P. Quignard. 2001. «Systematics of the Atlantic-Mediterranean soles Pegusa impar, P-lascaris, Solea aegyptiaca, S-senegalensis, and S-solea (Pleuronectiformes : Soleidae).» *Canadian Journal of Zoology-Revue Canadienne De Zoologie* 79: 2297-2302.

Boutier, B., J. F. Chiffoleau, J. L. Gonzalez, P. Lazure, D. Auger, et I. Truquet. 2000. «Influence of the Gironde estuary outputs on cadmium concentrations in the coastal waters: consequences on the Marennes-Oleron bay (France).» *Oceanologica Acta* 23: 745-757.

Casellas, J., R. J. Gularte, C. R. Farber, L. Varona, M. Mehrabian, E. E. Schadt, A. J. Lusis, A. D. Attie, B. S. Yandell, et J. F. Medrano. 2012. «Genome scans for transmission ratio distortion regions in mice.» *Genetics* 191: 247-59.

Catchen, J. M., A. Amores, P. Hohenlohe, W. Cresko, et J. H. Postlethwait. 2011. «Stacks: building and genotyping Loci de novo from short-read sequences.» *G3 (Bethesda)* 1: 171-82. http://www.ncbi.nlm.nih.gov/pubmed/22384329.

Charlesworth, B., M. T. Morgan, et D. Charlesworth. 1993. «The effect of deleterious mutations on neutral molecular variation.» *Genetics* 134: 1289-303. http://www.ncbi.nlm.nih.gov/pubmed/8375663.

Christe, C., K. N. Stolting, L. Bresadola, B. Fussi, B. Heinze, D. Wegmann, et C. Lexer. 2016. «Selection against recombinant hybrids maintains reproductive isolation in hybridizing Populus species despite F1 fertility and recurrent gene flow.» *Mol Ecol* 25: 2482-98. http://www.ncbi.nlm.nih.gov/pubmed/26880192.

Coyne, Jerry A., et H. Allen Orr. 2004. *Speciation.* Vol. 37. Sinauer Associates Sunderland, MA.

Cruickshank, Tami E., et Matthew W. Hahn. 2014. «Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow.» *Mol. Ecol.* 23: 3133-3157.

Danecek, P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, et al. 2011. «The variant call format and VCFtools.» *Bioinformatics* 27: 2156-8. http://www.ncbi.nlm.nih.gov/pubmed/21653522.

Debes, P. V., F. E. Zachos, et R. Hanel. 2008. «Mitochondrial phylogeography of the European sprat (Sprattus sprattus L., Clupeidae) reveals isolated climatically vulnerable populations in the Mediterranean Sea and range expansion in the northeast Atlantic.» *Mol. Ecol.* 17: 3873-88. http://www.ncbi.nlm.nih.gov/pubmed/18643878.

Derryberry, E. P., G. E. Derryberry, J. M. Maley, et R. T. Brumfield. 2014. «hzar: hybrid zone analysis using an R software package.» *Molecular Ecology Resources* 14: 652-663.

Desoutter Meniger, Martine. 1997. «Révision systématique des genres de la famille des Soleidae presents sur les côtes de l'Est-Atlantique et de la Méditerranée.» Ph.D. dissertation.

Digby, P. G. N., et Rodney Alistair Kempton. 1987. *Multivariate analysis of ecological communities.* Springer Science & Business Media.

Dobzhansky, T. 1937. «Genetic nature of species differences.» *Am. Nat.* 71: 404-420.

Dray, S., D. Chessel, et J. Thioulouse. 2003. «Procrustean co-inertia analysis for the linking of multivariate datasets.» *Ecoscience* 10: 110-119.

Dray, S., et A. B. Dufour. 2007. «The ade4 package: Implementing the duality diagram for ecologists.» *Journal of Statistical Software* 22: 1-20.

Edmands, S. 1999. «Heterosis and outbreeding depression in interpopulation crosses spanning a wide range of divergence.» *Evolution* 53: 1757-1768.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backstrom, T. Kawakami, A. Kunstner, et al. 2012. «The genomic landscape of species divergence in Ficedula flycatchers.» *Nature* 491: 756-60. http://www.ncbi.nlm.nih.gov/pubmed/23103876.

Ewing, Gregory, et Joachim Hermisson. 2010. «MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus.» *Bioinformatics* 26: 2064-2065.

Feder, J. L., et P. Nosil. 2010. «The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation.» *Evolution* 64: 1729-47. http://www.ncbi.nlm.nih.gov/pubmed/20624183.

Feder, J. L., S. P. Egan, et P. Nosil. 2012. «The genomics of speciation-with-gene-flow.» *Trends Genet.* 28: 342-50. http://www.ncbi.nlm.nih.gov/pubmed/22520730.

Gagnaire, P. A., E. Normandeau, S. A. Pavey, et L. Bernatchez. 2013a. «Mapping phenotypic, expression and transmission ratio distortion QTL using RAD markers in the Lake Whitefish (Coregonus clupeaformis).» *Mol. Ecol.* 22: 3036-3048.

Gagnaire, P. A., S. A. Pavey, E. Normandeau, et L. Bernatchez. 2013b. «The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing.» *Evolution* 67: 2483-97. http://www.ncbi.nlm.nih.gov/pubmed/24033162.

Gay, L., P. A. Crochet, D. A. Bell, et T. Lenormand. 2008. «Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models.» *Evolution* 62: 2789-806.

Gompert, Z., et C. A. Buerkle. 2011. «Bayesian estimation of genomic clines.» *Mol Ecol* 20: 2111-27. http://www.ncbi.nlm.nih.gov/pubmed/21453352.

Gompert, Z., et C. A. Buerkle. 2012a. «bgc: Software for Bayesian estimation of genomic clines.» *Mol Ecol Resour* 12: 1168-76. http://www.ncbi.nlm.nih.gov/pubmed/22978657.

Gompert, Z., T. L. Parchman, et C. A. Buerkle. 2012b. «Genomics of isolation in hybrids.» *Philos Trans R Soc Lond B Biol Sci* 367: 439-50. http://www.ncbi.nlm.nih.gov/pubmed/22201173.

Grant, P. R., et B. R. Grant. 2002. «Unpredictable evolution in a 30-year study of Darwin's finches.» *Science* 296: 707-11.

Grossen, Christine, Sampath S. Seneviratne, Daniel Croll, et Darren E. Irwin. 2016. «Strong reproductive isolation and narrow genomic tracts of differentiation among three woodpecker species in secondary contact.» *Molecular Ecology* 25: 4247-4266.

Gutenkunst, R. N., R. D. Hernandez, S. H. Williamson, et C. D. Bustamante. 2009. «Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data.» *PLos Genet.* 5: e1000695. http://www.ncbi.nlm.nih.gov/pubmed/19851460.

Gyllensten, U. 1985. «The genetic structure of fish: differences in the intraspecific distribution of biochemical genetic variation between marine, anadromous, and freshwater species.» *Journal of Fish Biology* 26: 691-699.

Harr, B. 2006. «Genomic islands of differentiation between house mouse subspecies.» *Genome Res.* 16: 730-7. http://www.ncbi.nlm.nih.gov/pubmed/16687734.

Harrison, R. G. 1986. «Pattern and Process in a Narrow Hybrid Zone.» *Heredity* 56: 337-349.

Harrison, R. G., et E. L. Larson. 2014. «Hybridization, introgression, and the nature of species boundaries.» *J. Hered.* 105 Suppl 1: 795-809.

Harrison, Richard Gerald. 1993. *Hybrid zones and the evolutionary process.* Oxford University Press on Demand.

Hedgecock, D., et A. I. Pudovkin. 2011. «Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary.» *Bulletin of Marine Science* 87: 971-1002.

Herrig, D. K., A. J. Modrick, E. Brud, et A. Llopart. 2014. «Introgression in the Drosophila subobscura--D. Madeirensis sister species: evidence of gene flow in nuclear genes despite mitochondrial differentiation.» *Evolution* 68: 705-19. http://www.ncbi.nlm.nih.gov/pubmed/24152112.

Hewitt, G. 2000. «The genetic legacy of the Quaternary ice ages.» *Nature* 405: 907-13. http://www.ncbi.nlm.nih.gov/pubmed/10879524.

Hey, J., et R. Nielsen. 2007. «Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics.» *Proc. Natl. Acad. Sci. U.S.A.* 104: 2785-90. http://www.ncbi.nlm.nih.gov/pubmed/17301231.

Jackson, D. A. 1995. «Protest - a Procrustean Randomization Test of Community Environment Concordance.» *Ecoscience* 2: 297-303.

Jombart, T. 2008. «adegenet: a R package for the multivariate analysis of genetic markers.» *Bioinformatics* 24: 1403-5. http://www.ncbi.nlm.nih.gov/pubmed/18397895.

Klingenberg, C. P., M. Barluenga, et A. Meyer. 2003. «Body shape variation in cichlid fishes of the Amphilophus citrinellus species complex.» *Biol. J. Linn. Soc.* 80: 397-408.

Kondrashov, Alexey S., et Mikhail V. Mina. 1986. «Sympatric speciation: when is it possible?» *Biol. J. Linn. Soc.* 27: 201-223.

Kotoulas, G., F. Bonhomme, et P. Borsa. 1995. «Genetic-Structure of the Common Sole Solea-Vulgaris at Different Geographic Scales.» *Marine Biology* 122: 361-375.

Kruuk, L. E., S. J. Baird, K. S. Gale, et N. H. Barton. 1999. «A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids.» *Genetics* 153: 1959-71. http://www.ncbi.nlm.nih.gov/pubmed/10581299.

Langmead, B., et S. L. Salzberg. 2012. «Fast gapped-read alignment with Bowtie 2.» *Nat Methods* 9: 357-9. http://www.ncbi.nlm.nih.gov/pubmed/22388286.

Le Moan, A., P-A. Gagnaire, et F. Bonhomme. 2016. «Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact.» *Molecular ecology.*

Martin, S. H., K. K. Dasmahapatra, N. J. Nadeau, C. Salazar, J. R. Walters, F. Simpson, M. Blaxter, A. Manica, J. Mallet, et C. D. Jiggins. 2013. «Genome-wide evidence for speciation with gene flow in Heliconius butterflies.» *Genome Res.* 23: 1817-28. http://www.ncbi.nlm.nih.gov/pubmed/24045163.

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

Mayr, Ernst. 1963. *Animal species and evolution.* Cambridge, Massachusett: Belknap Press of Harvard University Press.

Miller, M. R., J. P. Dunham, A. Amores, W. A. Cresko, et E. A. Johnson. 2007. «Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers.» *Genome Research* 17: 240-248.

Moyle, L. C., E. B. Graham, et Smbe Tri-National Young Investigators. 2006. «Proceedings of the SMBE Tri-National Young Investigators' Workshop 2005. Genome-wide associations between hybrid sterility QTL and marker transmission ratio distortion.» *Mol. Biol. Evol.* 23: 973-80.

Muller, Herman J. 1942. «Isolating mechanisms, evolution and temperature.» *Biol. Symp.* 71-125.

Nachman, M. W., et B. A. Payseur. 2012. «Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice.» *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 367: 409-21. http://www.ncbi.nlm.nih.gov/pubmed/22201170.

Nichols, P., M. J. Genner, C. van Oosterhout, A. Smith, P. Parsons, H. Sungani, J. Swanstrom, et D. A. Joyce. 2015. «Secondary contact seeds phenotypic novelty in cichlid fishes.» *Proc Biol Sci* 282: 20142272.

Nielsen, Einar E., Peter H. Nielsen, Dorte Meldrup, et Michael M. Hansen. 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.

Nielsen, R., et J. Wakeley. 2001. «Distinguishing migration from isolation: a Markov chain Monte Carlo approach.» *Genetics* 158: 885-96. http://www.ncbi.nlm.nih.gov/pubmed/11404349.

Noor, M. A., et S. M. Bennett. 2009. «Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species.» *Heredity (Edinb)* 103: 439-44. http://www.ncbi.nlm.nih.gov/pubmed/19920849.

Nosil, P., D. J. Funk, et D. Ortiz-Barrientos. 2009. «Divergent selection and heterogeneous genomic divergence.» *Mol. Ecol.* 18: 375-402. http://www.ncbi.nlm.nih.gov/pubmed/19143936.

Orr, H. A. 1995. «The population genetics of speciation: the evolution of hybrid incompatibilities.» *Genetics* 139: 1805-13. http://www.ncbi.nlm.nih.gov/pubmed/7789779.

Ouanes, K., L. Bahri-Sfar, O. K. Ben Hassine, et F. Bonhomme. 2011. «Expanding hybrid zone between Solea aegyptiaca and Solea senegalensis: genetic evidence over two decades.» *Mol. Ecol.* 20: 1717-28.

Palumbi, S. R. 1994. «Genetic-Divergence, Reproductive Isolation, and Marine Speciation.» *Annual Review of Ecology and Systematics* 25: 547-572.

Parchman, T. L., Z. Gompert, M. J. Braun, R. T. Brumfield, D. B. McDonald, J. A. Uy, G. Zhang, E. D. Jarvis, B. A. Schlinger, et C. A. Buerkle. 2013. «The genomic consequences of adaptive divergence and reproductive isolation between species of manakins.» *Mol. Ecol.* 22: 3304-17. http://www.ncbi.nlm.nih.gov/pubmed/23441849.

Parsons, K. J., Y. H. Son, et R. C. Albertson. 2011. «Hybridization Promotes Evolvability in African Cichlids: Connections Between Transgressive Segregation and Phenotypic Integration.» *Evolutionary Biology* 38: 306-315.

Patarnello, Tomaso, Filip A. M. J. Volckaert, et Rita Castilho. 2007. «Pillars of Hercules: is the Atlantic–Mediterranean transition a phylogeographical break?» *Molecular ecology* 16: 4426-4444.

Peres-Neto, Pedro R., et Donald A. Jackson. 2001. «How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test.» *Oecologia* 129: 169-178.

Pialek, J., et N. H. Barton. 1997. «The spread of an advantageous allele across a barrier: the effects of random drift and selection against heterozygotes.» *Genetics* 145: 493-504. http://www.ncbi.nlm.nih.gov/pubmed/9071602.

Presgraves, D. C. 2010. «The molecular evolutionary basis of species formation.» *Nat Rev Genet* 11: 175-80. http://www.ncbi.nlm.nih.gov/pubmed/20051985.

Quéro, J. C., M. Desoutter, et F. Lagardère. 1986. «Soleidae.» *Fishes of the North-eastern Atlantic and the Mediterranean* 3: 1308-1324.

Raj, A., M. Stephens, et J. K. Pritchard. 2014. «fastSTRUCTURE: variational inference of population structure in large SNP data sets.» *Genetics* 197: 573-89. http://www.ncbi.nlm.nih.gov/pubmed/24700103.

Ravigné, V., A. Barberousse, N. Bierne, J. Britton-Davidian, P. Capy, Y. Desdevises, T. Giraud, E. Jousselin, C. Moulia, et C. Smadja. 2010. «La spéciation.»

Rieseberg, L. H. 2001. «Chromosomal rearrangements and speciation.» *Trends Ecol. Evol.* 16: 351-358. http://www.ncbi.nlm.nih.gov/pubmed/11403867.

Rieseberg, L. H., A. Widmer, A. M. Arntz, et J. M. Burke. 2003. «The genetic architecture necessary for transgressive segregation is common in both natural and domesticated populations.» *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 1141-7.

Rieseberg, L. H., et J. H. Willis. 2007. «Plant speciation.» *Science* 317: 910-4.

Rieseberg, L. H., J. Whitton, et K. Gardner. 1999. «Hybrid zones and the genetic architecture of a barrier to gene flow between two sunflower species.» *Genetics* 152: 713-727.

Rieseberg, L. H., M. A. Archer, et R. K. Wayne. 1999. «Transgressive segregation, adaptation and speciation.» *Heredity (Edinb)* 83 (Pt 4): 363-72.

Roessig, J. M., C. M. Woodley, J. J. Cech, et L. J. Hansen. 2004. «Effects of global climate change on marine and estuarine fishes and fisheries.» *Reviews in Fish Biology and Fisheries* 14: 251-275.

Rohlf, F. J. 2006. «tpsDig, version 2.10.» *Department of Ecology and Evolution, State University of New York, Stony Brook.*

Rolland, J. L., F. Bonhomme, F. Lagardere, M. Hassan, et B. Guinand. 2007. «Population structure of the common sole (Solea solea) in the Northeastern Atlantic and the Mediterranean Sea: revisiting the divide with EPIC markers.» *Marine Biology* 151: 327-341.

Roux, C., G. Tsagkogeorga, N. Bierne, et N. Galtier. 2013. «Crossing the species barrier: genomic hotspots of introgression between two highly divergent Ciona intestinalis species.» *Mol. Biol. Evol.* 30: 1574-87. http://www.ncbi.nlm.nih.gov/pubmed/23564941.

Schlager, Stefan, et Gregory Jefferis. 2015. «Package 'Morpho'.»

She, J. X., M. Autem, G. Kotulas, N. Pasteur, et F. Bonhomme. 1987. «Multivariate-Analysis of Genetic Exchanges between Solea-Aegyptiaca and Solea-Senegalensis (Teleosts, Soleidae).» *Biol. J. Linn. Soc.* 32: 357-371.

Smadja, C., J. Galindo, et R. Butlin. 2008. «Hitching a lift on the road to speciation.» *Mol. Ecol.* 17: 4177-80. http://www.ncbi.nlm.nih.gov/pubmed/19378398.

Smith, J. M., et J. Haigh. 1974. «The hitch-hiking effect of a favourable gene.» *Genet. Res.* 23: 23-35. http://www.ncbi.nlm.nih.gov/pubmed/4407212.

Sobel, J. M.; Chen, G. F.; Watt, L. R.; Schemske, D. W. 2010. «The biology of speciation.» *Evolution* 64: 295-315. http://www.ncbi.nlm.nih.gov/pubmed/19891628.

Sousa, V., et J. Hey. 2013. «Understanding the origin of species with genome-scale data: modelling gene flow.» *Nat. Rev. Genet.* 14: 404-14. http://www.ncbi.nlm.nih.gov/pubmed/23657479.

Stankowski, Sean, James M. Sobel, et Matthew A. Streisfeld. 2016. «Geographic cline analysis as a tool for studying genome-wide variation: a case study of pollinator-mediated divergence in a monkeyflower.» *Molecular ecology.*

Stelkens, R. B., C. Schmid, O. Selz, et O. Seehausen. 2009. «Phenotypic novelty in experimental hybrids is predicted by the genetic distance between species of cichlid fish.» *BMC Evol. Biol.* 9: 283.

Stölting, Kai N., Rick Nipper, Dorothea Lindtke, Celine Caseys, Stephan Waeber, Stefano Castiglione, et Christian Lexer. 2013. «Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species.» *Molecular ecology* 22: 842-855.

Taylor, S. A., R. L. Curry, T. A. White, V. Ferretti, et I. Lovette. 2014. «Spatiotemporally consistent genomic signatures of reproductive isolation in a moving hybrid zone.» *Evolution* 68: 3066-81. http://www.ncbi.nlm.nih.gov/pubmed/25138643.

Teeter, K. C., L. M. Thibodeau, Z. Gompert, C. A. Buerkle, M. W. Nachman, et P. K. Tucker. 2010. «The Variable Genomic Architecture of Isolation between Hybridizing Species of House Mice.» *Evolution* 64: 472-485.

Tine, M., H. Kuhl, P. A. Gagnaire, B. Louro, E. Desmarais, R. S. Martins, J. Hecht, et al. 2014. «European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation.» *Nat. Commun.* 5: 5770. http://www.ncbi.nlm.nih.gov/pubmed/25534655.

Turelli, M., et H. A. Orr. 1995. «The dominance theory of Haldane's rule.» *Genetics* 140: 389-402. http://www.ncbi.nlm.nih.gov/pubmed/7635302.

Turner, John R. G. 1967. «Why does the genotype not congeal?» *Evolution* 645-656.

Vachon, Josiane, Francois Chapleau, et Martine Desoutter-Meniger. 2008. «Révision taxinomique du genre Solea et réhabilitation du genre Barnardichthys (Soleidae; Pleuronectiformes).» *Cybium* 32: 9-26.

Via, S., et J. West. 2008. «The genetic mosaic suggests a new role for hitchhiking in ecological speciation.» *Mol. Ecol.* 17: 4334-45. http://www.ncbi.nlm.nih.gov/pubmed/18986504.

Wakeley, J., et J. Hey. 1997. «Estimating ancestral population parameters.» *Genetics* 145: 847-55. http://www.ncbi.nlm.nih.gov/pubmed/9055093.

Waples, Robin S. 1998. «Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species.» *Journal of Heredity* 89: 438-450.

Ward, R. D., M. Woodwark, et D. O. F. Skibinski. 1994. «A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes.» *Journal of fish biology* 44: 213-232.

Whitehead, P. J. P., M. L. Bauchot, J. C. Hureau, J. Nielson, et E. Tortonese. 1986. «Fishes of the North-eastern Atlantic and the Mediterranean. Vol. I, II & III. Paris: United Nations Educational, Scientific and Cultural Organisation.»

Wilson, J. D., D. J. Schmidt, et J. M. Hughes. 2016. «Movement of a Hybrid Zone Between Lineages of the Australian Glass Shrimp (Paratya australiensis).» *Journal of Heredity* 107: 413-422.

Yúfera, M., G. Parra, R. Santiago, et M. Carrascosa. 1999. «Growth, carbon, nitrogen and caloric content of Solea senegalensis (Pisces: Soleidae) from egg fertilization to metamorphosis.» *Marine Biology* 134: 43-49.

Zelditch, Miriam. 2004. *Geometric morphometrics for biologists : a primer.* Amsterdam ; London: Elsevier Academic Press.

Evolutionary Applications
Open Access
Evolutionary Applications ISSN 1752-4571

REVIEWS AND SYNTHESIS

Using neutral, selected, and hitchhiker loci to assess connectivity of marine populations in the genomic era

Pierre-Alexandre Gagnaire,[1,2] Thomas Broquet,[3,4] Didier Aurelle,[5] Frédérique Viard,[3,4] Ahmed Souissi,[1] François Bonhomme,[1,2] Sophie Arnaud-Haond[1,6] and Nicolas Bierne[1,2]

1 Université de Montpellier, Montpellier, France
2 CNRS – Institut des Sciences de l'Evolution, UMR 5554 UM-CNRS-IRD-EPHE, Station Méditerranéenne de l'Environnement Littoral, Sète, France
3 CNRS team Diversity and connectivity of coastal marine landscapes, Station Biologique de Roscoff, Roscoff, France
4 Sorbonne Universités, UPMC Université Paris 06, UMR 7144, Station Biologique de Roscoff, Roscoff, France
5 Aix Marseille Université, CNRS-IRD-Avignon Université, IMBE UMR 7263, Marseille, France
6 Ifremer, UMR "Ecosystèmes Marins Exploités", Sète, France

Keywords
connectivity, gene flow, marine conservation, population genomics, population structure.

Correspondence
Pierre-Alexandre Gagnaire, Université de Montpellier, 34095 Montpellier, France.
Tel.: +33-4-67463375;
fax: +33-4-67463399;
e-mail: pagagnai@univ-montp2.fr

Received: 5 March 2015
Accepted: 5 June 2015

doi:10.1111/eva.12288

Abstract

Estimating the rate of exchange of individuals among populations is a central concern to evolutionary ecology and its applications to conservation and management. For instance, the efficiency of protected areas in sustaining locally endangered populations and ecosystems depends on reserve network connectivity. The population genetics theory offers a powerful framework for estimating dispersal distances and migration rates from molecular data. In the marine realm, however, decades of molecular studies have met limited success in inferring genetic connectivity, due to the frequent lack of spatial genetic structure in species exhibiting high fecundity and dispersal capabilities. This is especially true within biogeographic regions bounded by well-known hotspots of genetic differentiation. Here, we provide an overview of the current methods for estimating genetic connectivity using molecular markers and propose several directions for improving existing approaches using large population genomic datasets. We highlight several issues that limit the effectiveness of methods based on neutral markers when there is virtually no genetic differentiation among samples. We then focus on alternative methods based on markers influenced by selection. Although some of these methodologies are still underexplored, our aim was to stimulate new research to test how broadly they are applicable to nonmodel marine species. We argue that the increased ability to apply the concepts of cline analyses will improve dispersal inferences across physical and ecological barriers that reduce connectivity locally. We finally present how neutral markers hitchhiking with selected loci can also provide information about connectivity patterns within apparently well-mixed biogeographic regions. We contend that one of the most promising applications of population genomics is the use of outlier loci to delineate relevant conservation units and related eco-geographic features across which connectivity can be measured.

Introduction

Inferring population connectivity from molecular data within a population genetic framework can shed light on the evolutionary and ecological processes that shape patterns of genetic diversity (Clobert et al. 2012). Population genetic approaches offer convenient methods to evaluate the rate and scale of dispersal (or migration) when the movement of individuals cannot be assessed by other means such as mark–recapture field experiments. This problem is particularly acute in the marine environment, where the distribution and migratory pathways of organisms are hidden to human eyes underneath the surface of the oceans (Hellberg 2009; Selkoe and Toonen 2011). The

potential of genetic methods, illustrated by successful studies in reef species (Selkoe et al. 2010; Puebla et al. 2012), has led to increased expectations for inferring marine connectivity patterns from molecular markers, especially for conservation and management purposes.

The majority of marine species, however, display combinations of life history traits (e.g. high fecundity, large population sizes, high dispersal potential often combined to complex life cycles) that produce weak patterns of genetic differentiation or even no differentiation at all (Ward et al. 1994; Waples 1998; Palumbi 2003; Hedgecock et al. 2007). A lack of genetic differentiation may result from a range of situations spanning from nearly complete demographic independence among large-sized populations to the existence of a unique panmictic population (Palumbi 2003; Waples and Gaggiotti 2006; Waples et al. 2008) (Fig. 1). Spatial genetic homogeneity may thus hide a large diversity of scenarios with regard to the contemporary rates of demographic exchanges among groups of individuals inhabiting different parts of a species range. This is of particular concern because the per-generation number of migrants, which is sufficient to lead to apparent genetic panmixia, may not be high enough to ensure demographic connectivity and rescue effects (Waples 1998; Lowe and Allendorf 2010). This discrepancy between the objective of inferring demographic connectivity for conservation biology and management purposes and the limitations inherent to most population genetic approaches has motivated several reviews in the field (Waples and Gaggiotti 2006; Broquet and Petit 2009; Hellberg 2009; Lowe and Allendorf 2010). Our goal here is not to provide a new synthesis of existing methods to infer connectivity, which have been thoroughly addressed in those reviews. We rather aim at considering the new perspectives offered by the increasing number of markers in population genomic studies, with a special focus on the use of loci influenced by selection. The rapid spread of next-generation sequencing (NGS)-based genotyping methods in the molecular ecologists' toolbox has considerably enhanced our ability to identify and characterize genetic variation from population samples (Davey et al. 2011). Still, it remains unclear which approaches will benefit the most from this massive amount of sequence data. One direct benefit of analyzing thousands of markers is an increased precision in measuring genetic differentiation and a higher statistical power to detect small genetic differences among populations (Waples 1998). However, populations with large effective sizes, high migration rates or both may remain virtually undifferentiated, and thus, multiplying neutral markers in such cases may still fail to reveal the current level of demographic connectivity.

Another major achievement offered by NGS approaches is to facilitate the discovery of genetic markers that are influenced by selection (Allendorf et al. 2010; Stapley et al. 2010). These outlier loci can reveal genetic differentiation

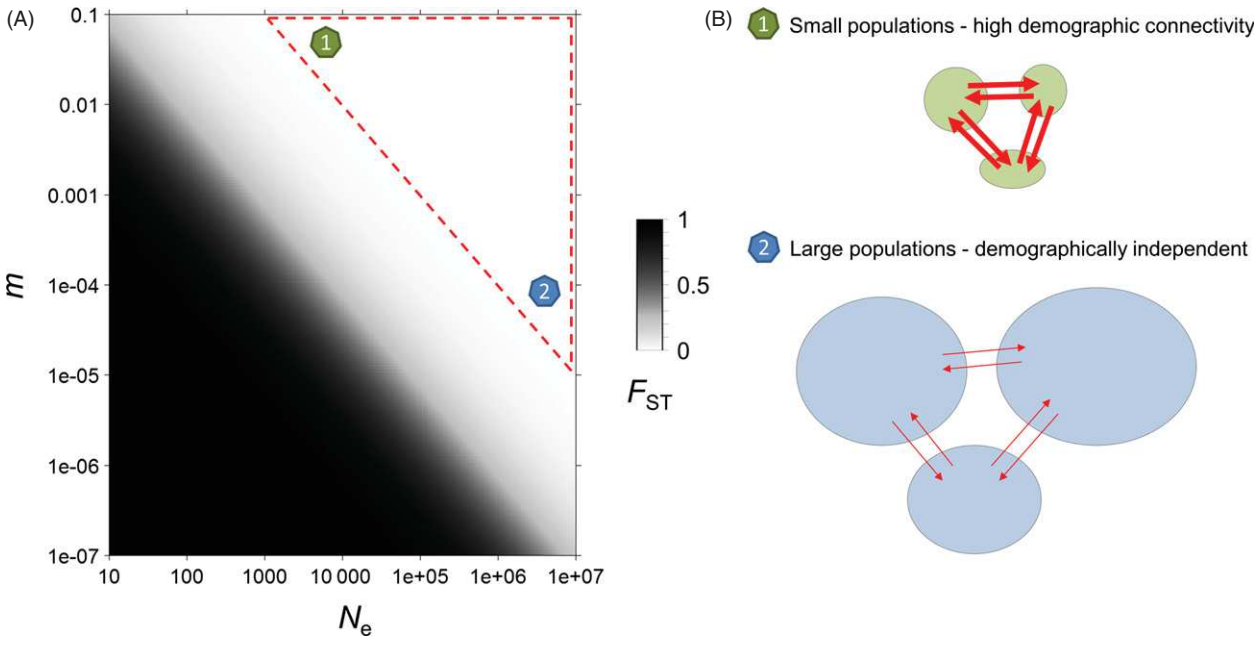


Figure 1 The demographic parameters values behind weak F_{ST} values. (A) Because of the nonlinear relationship between F_{ST} and N_em in the island model, genetic differentiation (F_{ST}, in color scale) rapidly shrinks as the per-generation number of migrants (N_em) increases. (B) At equilibrium, weak to null genetic differentiation is expected for small ($N_e = 10^3$) and highly connected ($m = 10^{-1}$) populations, but also for large ($N_e = 10^7$) and demographically independent ($m = 10^{-5}$) populations.

patterns at the place where neutral markers often remain uninformative, and therefore, it has been suggested that the signal held by outlier loci could be used to delineate locally adapted stocks and redefine conservation units (Nielsen et al. 2009, 2012; Funk et al. 2012). This approach is appealing because selection may be much more efficient than drift in opposing the homogenizing effect of migration, in particular when populations have large effective sizes. However, outlier loci may arise through a wide variety of evolutionary mechanisms apart from local adaptation, which is the primary target of most genome scan studies looking for adaptive variation (Bierne et al. 2013). These evolutionary mechanisms thus need to be identified before using outlier loci to evaluate connectivity.

Allele frequency shifts at outlier loci are expected to be concentrated in particular geographic regions where strong ecological gradients promote local adaptation (Schmidt et al. 2008). Hotspots of genetic differentiation may also arise through the trapping of tension zones by natural barriers to dispersal (Barton 1979a), or through the coupling between exogenous and endogenous reproductive barriers (Bierne et al. 2011). These predictions are corroborated by well-known hotspots of genetic differentiation in the sea (e.g. the Almeria-Oran front, the Siculo-Tunisian strait, Cape Agulhas, Cape Cod, Oresund, Point Conception, among others). However, marine conservation and management issues often require measures of connectivity in areas located outside these particular zones. In a last section, we explore alternative mechanisms that generate disequilibrium at neutral hitchhiker loci even outside the cline of the selected locus itself. These indirect effects of selection can reveal cryptic genetic structure within apparently well-mixed areas. These effects are of two sorts: (i) gradients of introgression (or introgression tails) originating from a geographically distant contact zone (Gagnaire et al. 2011) and (ii) hitchhiking clines that are transiently generated during the propagation of a selective sweep (Bierne 2010). Large population genomic datasets now provide molecular ecologists with the means to use these patterns to study marine connectivity. Therefore, there is a good hope that gathering theoretical background with these new data will further catalyze research in the field.

Genetic approaches to marine connectivity using neutral markers: successes and limits

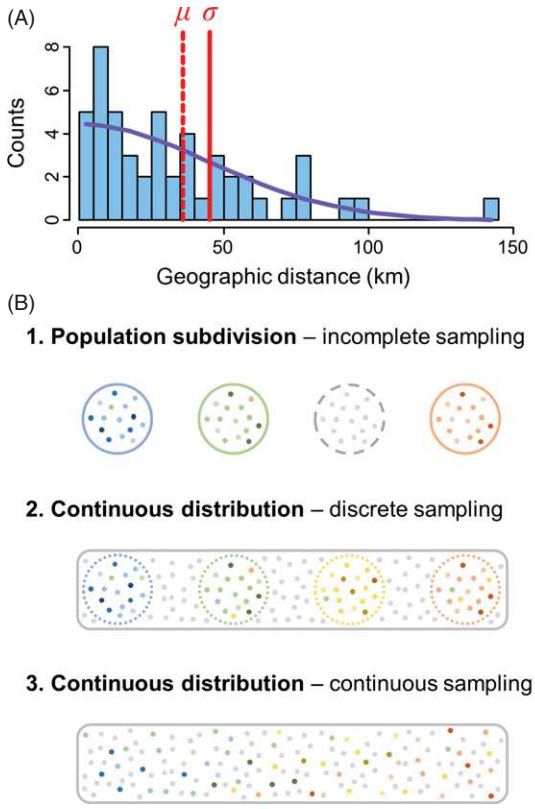
Quantitative methods for inferring dispersal with neutral genetic markers fall into two broad categories. One first class of methods looks for the effects of gene flow on the level of genetic differentiation between populations. These methods rely on population genetics models that integrate all relevant evolutionary forces, apart from the effect of mutation which can be neglected for a wide range of migration rates (Box 1). For instance, migration can be indirectly inferred using observations of genetic structure and a model formalizing how gene flow might have produced these observations. The second option is to detect individual dispersal events directly to reconstruct the distribution of dispersal distances, which can be done through genealogy inference (e.g. parentage assignment) or clustering analysis to ascertain population membership of individuals. Not all methods strictly take one of these two routes, but we mention this broad dichotomy here because it gives a good indication of the underlying assumptions, the amount of data required, the nature of the dispersal (or migration) parameter to be estimated, and the spatial and temporal scales over which each method is pertinent and

Box 1: What is the right F_{ST} estimator in high gene flow species?

Since the advent of multi-allelic loci in population genetics, it has been pointed out that the maximum value taken by Wright's F_{ST} at a given locus is bounded by its level of genetic diversity so that $F_{max} \leq 1 - H_S$, where H_S is the average within-sample diversity (reviewed in Meirmans and Hedrick 2011). Various methods (i.e. estimators aimed at scaling the maximum possible value to 1) have been proposed to correct what is perceived under certain circumstances as a bias, or even a drawback for measuring allelic differentiation between populations (Jost 2008). In the case of bi-allelic loci, all measures (e.g. D, F', G, θ; see definitions in Meirmans and Hedrick 2011) give the same equilibrium estimation as Wright's F_{ST}, which can be transformed into migration rate at migration–drift equilibrium. However, the problem becomes more complex when more than two allelic states occur, and one wishes to take mutation and homoplasy into account. In the island model, $F_{ST} \simeq \frac{1}{1+4Nm+4N\mu}$, so the relative role of mutation and migration becomes a key issue. In the case of high gene flow species, it is generally admitted that $\mu \ll m$. Hence, the main criticism against the use of F_{ST} (i.e. $m = 0$, $\mu > 0$, which produces a multi-allelic F_{ST} estimate tending toward 0 with time despite maximal differentiation and the absence of gene flow) is not justified in such species. On the contrary, $m \gg \mu$ would imply that the variation detected in one deme is mostly replenished by migration from the metapopulation rather than by locally arisen mutations. Under this assumption, the small differentiation generally observed in most marine species would not be an artifact of using multi-allelic markers, but the consequence of high migration. This has two consequences for our interpretation of genetic variation in high gene flow species: (i) homoplasy is likely to play a very limited role because homoplastic alleles will be almost equally distributed throughout the metapopulation by migration, (ii) high heterozygosity values reflect large metapopulation effective size rather than locally high population size, a pattern that remains true even with relatively low migration between populations.

Box 2: Neutral methods to infer genetic connectivity.

(A) The distribution of dispersal distances can be estimated in two ways: through the direct detection of discrete dispersal events (blue bars), or the indirect estimation of dispersal parameters like the standard deviation of parent–offspring dispersal distances (σ). The mean dispersal distance (μ) can be obtained from σ by assuming a normal distribution of dispersal distance (blue line). (B) The three sampling strategies commonly used in marine population genetic studies: 1. A geographically subdivided species range is discretely sampled, but some populations are not sampled (gray dotted circle); 2. A continuously distributed species is sampled discretely, and the geographic distance between samples is of the same order as σ; 3. Continuous sampling of individuals separated by distances of the same order as σ. Colored points are sampled individuals, gray points indicate nonsampled individuals. (C) The information about dispersal that can be obtained from indirect and direct methods for each sampling strategy.



(C)

	1. **Population subdivision** incomplete sampling	2. **Continuous distribution** discrete sampling	3. **Continuous distribution** continuous sampling
Indirect methods			
IBD	σ	σ	σ
F_{ST}	N_em	-	-
Direct methods			
Parentage assignment	Distribution of dispersal distances	Distribution of dispersal distances	Distribution of dispersal distances
Genetic assignment	Self-recruitment *vs.* dispersal	-	-
Clustering	Limits to dispersal	-	-

its estimates reliable. The applicability of these methods to three sampling strategies commonly deployed in marine population genetic studies is summarized in Box 2. Below, we briefly describe their broad properties rather than providing an exhaustive catalog, highlighting why both types of dispersal inference methods may be limited in their application for many marine species.

Inferring genetic connectivity using indirect methods

A representative example of indirect methods is the estimation of dispersal from IBD patterns. The IBD model can be used to estimate dispersal from the increase in genetic differentiation with increasing geographic distances between populations (Rousset 1997) or individuals (Rousset 2000) when dispersal is spatially limited (Box 2). Another example is the estimation of the absolute number of migrants per generation (N_em) from F_{ST} in the island model (Wright 1951). Other indirect methods include estimators of N_em or m under various extensions of the island model or other more refined population structures (Broquet and Petit 2009). All these methods are associated with a number of generally strong assumptions regarding the structure of populations (e.g. constant and equal size of demes, homogeneous migration, and population density), the life cycle of the species (e.g. nonoverlapping generations, identification of pre- and postdispersal stages and random mating

within demes), and the role of each evolutionary force (e.g. negligible effect of selection and negligible or known mutation rate). Model-based methods share at least two important properties. First, a measure of genetic structure never easily translates into an estimate of migration rate (Whitlock and McCauley 1999; Marko and Hart 2011). In particular, a low F_{ST} does not necessarily mean that migration is strong as genetic differentiation is influenced by both effective size (N_e) and migration rate (m) (Fig. 1). For instance, little dispersal is required to limit the global differentiation in an island model or a stepping-stone migration model (Rousset 2004; and for a recent empirical example: Puebla et al. 2012). Second, dispersal estimates often depend on other known parameters relevant to other evolutionary forces. For instance, the effect of genetic drift must be estimated independently (usually using density estimates) to infer dispersal under the isolation-by-distance model (Pinsky et al. 2010). The main advantage of model-based inference methods is that they require a small amount of data (for the less demanding methods, say about 10 sampling sites with 20 individuals per site). These methods produce estimates of migration rates (m) or moments of the distribution of dispersal distances (such as σ, the standard deviation of axial dispersal distances, Box 2) which can be difficult to interpret in an ecological context. Finally, such estimates, which have the merit of integrating the effects of evolutionary forces over longer time scales than direct approaches, rely on the questionable hypotheses that dispersal is stable over time and that the migration–drift equilibrium has been reached.

Indirect methods have been applied in a series of case studies in marine organisms. For instance, keeping IBD as a typical example of model-based inference, empirical estimates of σ have been reported for a variety of species (Rose et al. 2006; Puebla et al. 2009, 2012; Ledoux et al. 2010; Pinsky et al. 2010). However, such indirect approaches of dispersal can fail on two grounds. First, if genetic drift is too weak to generate population differentiation, then dispersal cannot be inferred using a model that relies on the migration/drift balance. This problem is often encountered in species with extremely large population sizes, such as many marine fishes and invertebrates (DeWoody and Avise 2000; McCusker and Bentzen 2010). For instance, there is no detectable genetic differentiation among populations of the California sea mussel *Mytilus californianus* across 4000 km of its distribution range (Addison et al. 2008). Second, species with large effective population sizes may show patterns of genetic structure that are not at mutation–migration–drift equilibrium. Indirect estimators of dispersal are based on different statistics that evolve at their own speed. Therefore, the rate of approach of equilibrium for a given estimator has to be evaluated to determine whether or not equilibrium is a strong assumption in particular case studies. For that reason, the uncertainty of indirect dispersal estimates due to a possible departure from equilibrium is generally unknown (Pogson et al. 2001).

Direct estimates of genetic connectivity

In contrast with indirect approaches, the direct detection of migrants through parentage analysis or individual assignment makes much fewer assumptions. For instance, population or parentage assignment methods allow estimating dispersal rates or distances without necessarily relying on demo-genetic models. On the downside, these approaches generally require a great deal of data, very good knowledge of the species distribution, and make the sampling design critical as it must be representative of the postdispersal distribution of individuals (evaluation of long-distance dispersal might be especially difficult due to constraints in the size of the study area). In the case of parentage analysis, an additional constraint stems from the necessity to sample a large fraction of the potential parents. Assignment methods specifically applied to detect immigrants without identifying their origin require less extensive sampling, but their efficiency reduces quickly with decreasing genetic differentiation (but see Gaggiotti et al. 2002). Parental assignments or first-generation migrant tracking methods provide measures of dispersal distances that are relevant to the dispersal episode preceding sampling. Moreover, these methods yield estimates of individual movement rather than gene flow, as immigrants may or may not reproduce locally following dispersal. Finally, although interpreting the results produced by these methods is generally more intuitive than those of indirect approaches, care must be taken regarding the effect of type I errors (i.e. incorrectly identifying a local individual as an immigrant) and unsampled putative parents or source populations (Paetkau et al. 2004; Waples and Gaggiotti 2006). For example, even with high statistical power (no type II error), accepting a 5% type I error for detecting migrants can spuriously increase the estimate of migration rate.

Direct methods have been successfully applied to some marine species. For instance, genetic assignment has yielded useful dispersal information in seals (Gaggiotti et al. 2002), reef fish (Saenz-Agudelo et al. 2011), and corals (Underwood et al. 2007). Similarly, parentage assignment has proven efficient in a number of case studies focusing especially on reef fishes (Jones et al. 2005; Planes et al. 2009; Christie et al. 2010; Almany et al. 2013). Besides a minute type I error, the success of such studies relies upon the fraction of potential source populations or parents that are sampled. These approaches thus require a high-density sampling at a relevant geographic scale, and their application in the marine environment is therefore limited to species with population sizes and distribution ranges that are well documented and small

enough for such sampling to be realistic. Although recent studies have shown that larval dispersal oftentimes occurs over smaller spatial scales than previously believed (Swearer et al. 2002; Almany et al. 2007; van der Meer et al. 2012; Puebla et al. 2012), many marine species typically have high fecundity rates, large distribution ranges, and population size (Palumbi 1994). These methods are thus inapplicable to the majority of marine animal species (from invertebrates to pelagic fishes) that have medium to large population sizes, elusive population contours and for which only a minute fraction of the individuals can be sampled for genetic studies.

Investigating genetic connectivity with clustering methods

When the study species is subdivided into discrete populations, there is a need to first determine the number of populations before evaluating gene flow (Waples 1998). Clustering methods which detect genetic discontinuities and limits to gene flow have been proposed as a way to identify both populations (stocks) and migrants (Pritchard et al. 2000; Broquet et al. 2009). The different clustering approaches (Pritchard et al. 2000; Corander et al. 2003) have their own limits, such as departures from the underlying models (François and Durand 2010). In particular, patterns of isolation by distance may lead to artificial clustering (Schwartz and McKelvey 2009; Blair et al. 2012; Aurelle and Ledoux 2013), and peculiar reproductive systems like partial selfing can induce spurious admixture patterns (Gao et al. 2007). The power of clustering methods generally increases with the amount of genetic differentiation among populations (Latch et al. 2006). For that reason, they are mostly suited to infer genetic connectivity in species with intrinsically or behaviorally limited dispersal abilities and relatively small local population sizes (Ledoux et al. 2010; Wilson and Eigenmann Veraguth 2010; Mokhtar-Jamaï et al. 2011; Perrier et al. 2011; Ansmann et al. 2012; Lukoschek and Shine 2012), which are not representative of the majority of marine species.

Population genomics using neutral markers for marine connectivity studies: what way forward?

Estimating connectivity from genetic data is a challenging task, which is made even more difficult by the particular life history traits and demographic characteristics of many marine species. More markers may enhance the statistical power of genetic studies and yield more precise estimates of small genetic differentiation values (Patterson et al. 2006), but the signature of dispersal contained in the data may remain intrinsically small or inexistent. In particular, it is not clear whether increasing the number of loci will help in situations where large effective population size keeps genetic structure down, even with restricted migration. As the number of markers rapidly increases, the nonindependence of loci in large population genomic datasets is also becoming another important issue which requires further investigation (Waples 2015).

Despite well-recognized limitations, there is still a good hope that population genomic datasets will improve the usefulness of indirect methods by increasing the power and precision of small genetic differentiation estimates. Although fairly robust estimates of dispersal were already obtained from IBD patterns among populations or discrete geographic samples using tens of markers, greater improvement is expected for methods based on genetic differentiation between individuals (Rousset 2000). This should be achieved through a more accurate estimation of pairwise genetic differentiation between individuals, just like population genomic datasets have improved the inference of relatedness between pairs of individuals for heritability estimation (Visscher et al. 2008). Because the power of isolation-by-distance regression scales with the number of observed pairwise geographic and genetic distances, a continuous sampling of individuals separated by distances in the order of σ (Rousset 2000) may be preferable to a discrete sampling of groups of individuals (Box 2).

Analyzing thousands of markers should also increase the power of direct methods, although the type I error issue underlined above is unlikely to be fully resolved even with high power, and sampling requirements cannot be alleviated by intensifying the genetic coverage of each individual. On the other hand, population genomic datasets may also contain useful information on migration events that trace back to several generations in the past. Therefore, extending direct estimates of dispersal beyond the identification of parent–offspring or sibling relationships seems appealing. This should encourage the development of methods that take the full spectrum of relatedness into account.

Whether large datasets will significantly improve the ability of clustering methods to detect existing structure when genetic differentiation is small remains to be tested with recent programs that have been specifically developed for rapidly processing population genomic data (Popescu et al. 2014; Raj et al. 2014). The use of principal component analysis (PCA) methods already proved useful for detecting fine-scale structure between human populations exhibiting low levels of genetic differentiation (Patterson et al. 2006; Novembre et al. 2008). This type of analysis may benefit from the informativeness of rare variants to detect fine-scale population structures (O'Connor et al. 2014), especially in the case of large populations that only exchange few migrants per generation.

Genome-wide polymorphism data that contain information about haplotype phase may open other interesting possibilities for studying connectivity. Immigration fol-

lowed by successive rounds of sexual reproduction with local residents produces individuals with mixed genetic ancestry. Across generations, the original immigrant chromosomes are progressively broken down by recombination, so that the genome of admixed individuals is composed by a mosaic of segments originating from different ancestral populations (Gompert and Buerkle 2013). The length distribution of such admixture tracts (also called migrant tracts) can be used to infer migration rates between populations (Pool and Nielsen 2009; Gravel 2012). In practice, this approach requires that the ancestry of admixture tracts can be accurately inferred, and this might be possible only when admixture stems from divergent populations. A related approach is based on the analysis of identical genomic segments that are inherited by pairs of individuals. The genomic proportions of long segments that are identical by descent between individuals from the same or different populations are directly related with migration rate (Palamara and Pe'er 2013). Referred to as 'haplotype sharing', this approach may be better suited to infer relatively recent migration between populations, although so far it has only been tested using high marker density datasets in species with a high-quality reference genome. These methods are currently under development (Gravel 2012; Liang and Nielsen 2014) and need to be evaluated for their potential to estimate migration in nonmodel species with weakly structured populations contemporarily exchanging migrants. Below, we develop another avenue of research that takes advantage of large population genomic datasets by focusing on genetic markers affected by selection.

Using selected and hitchhiker loci as an alternative approach to infer marine connectivity

As developed above, the approaches to infer demographic parameters from genetic data classically rely on neutral models that assume a balance between migration and genetic drift (Whitlock and McCauley 1999). As in large populations the effect of drift is very weak, even the most sophisticated methods based on this balance may lack power to infer migration, not to mention that disentangling the effects of N_e and m it is very difficult under these models (Waples 1998; Fig. 1). Alternatively, selection can act as a more efficient antagonistic evolutionary force than drift to counteract the homogenizing effect of migration (Lenormand et al. 1998). As the efficiency of selection scales up with population size, the counterbalancing effect of directional or divergent selection is expected to be greater in marine species with large population sizes (Allendorf et al. 2010). The detection of selected genes has long been a challenging prerequisite, but large marker datasets have considerably enhanced the power of genome scans to identify loci with extreme levels of differentiation (Stapley et al. 2010), the so-called 'F_{ST} outliers' supposed to be directly or, more probably, indirectly affected by selection (Luikart et al. 2003; Storz 2005). Recent conservation genetic studies have proposed to delineate locally adapted units based on the signal held by outlier loci (Funk et al. 2012; Nielsen et al. 2012), but without providing means to explicitly assess connectivity between such units. Before providing further guidance for using selected markers to infer the rate and scale of dispersal, we consider some of the problems that specifically arise with this category of markers.

Important concerns related to outlier detection

A common issue encountered in population genomic studies is that the different methods that can be used for identifying F_{ST} outliers usually detect only partially overlapping sets of loci. These inconsistencies across methods partly reflect the influence of unknown genetic structure and demographic history on outlier detection tests (for recent reviews, see Narum and Hess 2011; De Mita et al. 2013; De Villemereuil et al. 2014; Lotterhos and Whitlock 2014). In particular, the most commonly used methods for detecting F_{ST} outliers have a high rate of false-positive detection under nonequilibrium scenarios (Fraïsse et al. 2014; Lotterhos and Whitlock 2014), hierarchical population genetic structure (Excoffier et al. 2009), and IBD patterns (Meirmans 2012; Fourcade et al. 2013), while suffering at the same time from limited sensitivity (false-negative detection). To circumvent these problems, combining differentiation-based methods with genotype–environment association tests was suggested as a more reliable outlier identification approach (De Villemereuil et al. 2014). In addition, new methods have been developed that are expected to account for correlated ancestry among samples (Excoffier et al. 2009; Bonhomme et al. 2010; Duforet-Frebourg et al. 2014; Foll et al. 2014). However, even if F_{ST} outlier tests perform rather well when selection acts on few loci with large effects, they are more seriously challenged by selection acting on many small-effect loci or when the marker loci are loosely linked to the target loci. Because adaptation involving quantitative traits most often evolves through polygenic selection (Pritchard and Di Rienzo 2010), the small changes in allele frequencies resulting from polygenic adaptation may remain below the detection limit of most outlier detection methods (Le Corre and Kremer 2012). In light of recent simulation-based studies that have investigated the performance of outlier tests, it thus appears that outlier candidates should be submitted to validation by combining different statistical approaches, or more directly by comparing allele frequencies before (e.g. in the larval pool) and after (e.g. in juveniles or adults) selective mortality whenever possible (Gagnaire et al. 2012).

An important question that stems from acknowledging the limited power to detect polygenic selection is how to treat the signal held by the most differentiated selected loci, which may have a long, complex, and unanticipated history of divergence, when many others remain undetectable? Apart from the fact that undetected selected loci are expected to bias the neutral-based estimations of connectivity described above, these loci may be more informative than neutral loci for delineating genetic clusters, even if polygenic selection produces small allele frequency changes. Principal component-based analyses combining neutral and selected loci may thus be used as a naive approach to test whether genetic variation is continuously distributed in space, or partitioned into discrete genetic clusters to which individuals can be assigned to estimate the rate and scale of dispersal. Because many species probably match the polygenic selection model, this approach may be appropriate to improve the delineation of discrete genetic clusters in study systems where neutral marker datasets have been uninformative. However, the gain of power offered by large population genomic datasets is difficult to predict and requires further examination using simulated data under different dispersal scenarios.

Another concern when reliable outlier candidates can be identified relates to the nature of the selective effects behind their detection. Several selective mechanisms can increase genetic differentiation above the genome-wide average (Bierne et al. 2013), including underdominance, local and background selection (Charlesworth et al. 1997), but also convergent evolution in response to uniform selection (Ralph and Coop 2010). Importantly, each type of selection can affect neutral loci through linkage, which means that outlier loci could most often result from the indirect effect of selection at other loci. The pervasive effect of selection at linked sites has been well documented in *Drosophila* (e.g. Langley et al. 2012). In such species with large population effective sizes, background selection (Charlesworth et al. 1993) and genetic hitchhiking (Maynard Smith and Haigh 1974) can easily generate correlation between local recombination rates and genetic diversity. Such a correlation has been recently described in the stickleback (Roesti et al. 2012) and the European sea bass (Tine et al. 2014), which confirms that selection at linked sites can also have a dominant effect on genetic diversity in marine species.

In the next sections, we consider geographic patterns generated under various types of selection and provide a guide to infer genetic connectivity using existing and newly developed theoretical frameworks. Applied in local areas with environmental and hydrological singularities, some of these approaches will provide quantitative estimates of dispersal. In other cases, where the effects of selection are less well resolved, genomic data will be helpful to detect genetic discontinuities and provide qualitative assessment of connectivity.

Estimating dispersal distances using genetic clines

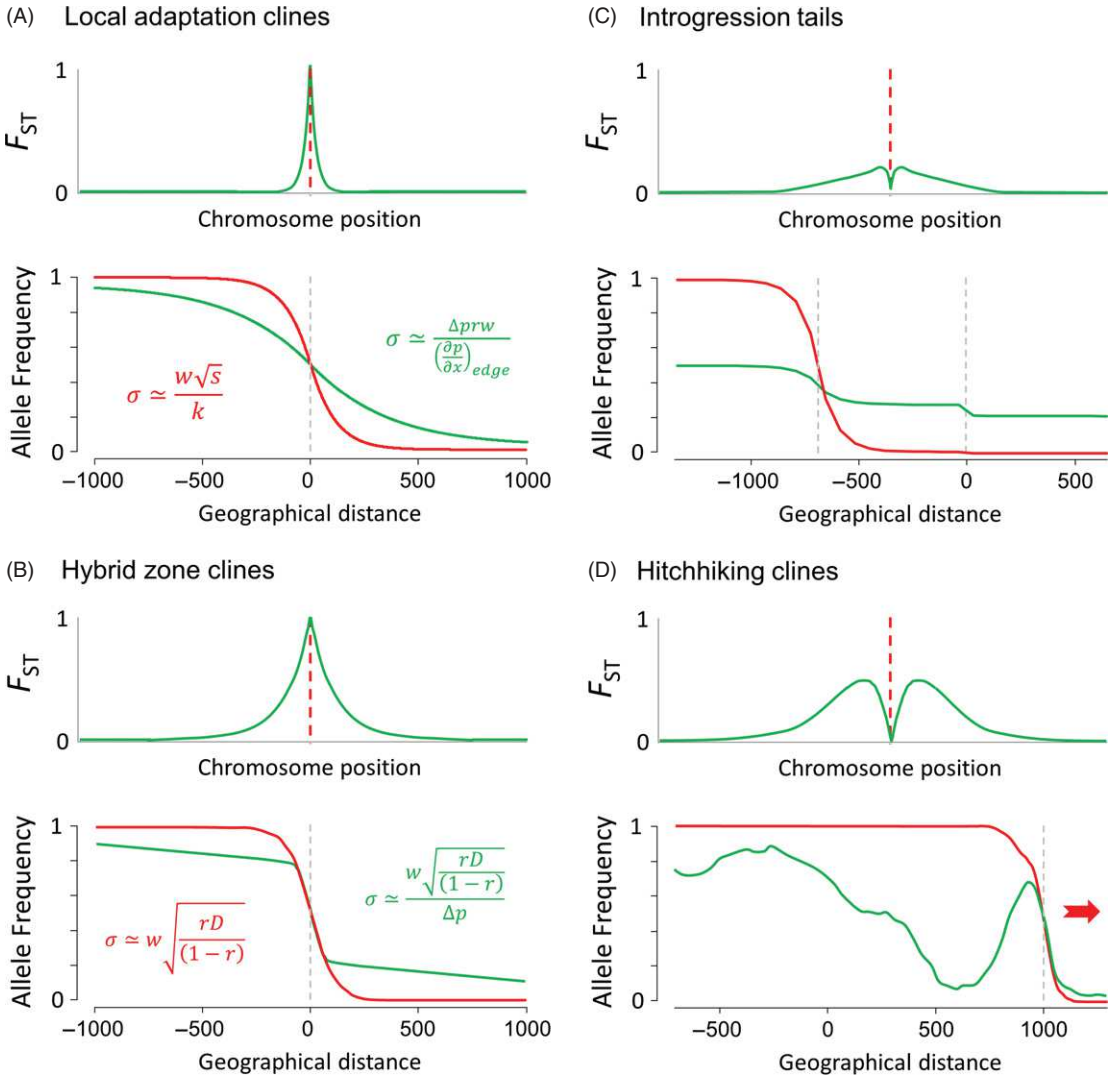
Genome scan studies in marine species have reported several empirical examples of outlier loci exhibiting clinal variation patterns, usually coinciding in space with environmental gradients, ecotones, or boundaries between biogeographic regions (Murray and Hare 2006; Bradbury et al. 2010; Colbeck et al. 2011; Gagnaire et al. 2011; Lamichhaney et al. 2012; Limborg et al. 2012). It is well established that selected markers analyzed in light of cline theory can provide robust estimates of dispersal distances (Barton and Gale 1993; Lenormand et al. 1998; Sotka and Palumbi 2006). Cline shape is basically determined by a balance between migration and selection, which allows under quasi-equilibrium conditions to derive the dispersal parameter in the geographic region of the cline. Empirically inferred dispersal distances may not be precise when only one locus is available and when linkage disequilibrium between selected loci is unknown, but even then they should be of the right order of magnitude (Sotka and Palumbi 2006). Population genomic studies have now the power to detect loci exhibiting clinal variation in species previously believed to be genetically homogeneous, so the potential for discovering new cases of local adaptation clines and cryptic hybrid zones is high (Bierne et al. 2011).

Using local adaptation clines to infer dispersal

We refer to local adaptation clines as monogenic clinal variation patterns maintained by a balance between the divergent effects of selection and the homogenizing effects of migration. Such clines occur along environmental gradients or at the frontier between habitats when alternative alleles have antagonistic fitness effects in different environmental conditions (Powers and Place 1978; Koehn et al. 1980). Allele frequencies vary as a sigmoid function of geographic distance (Box 3A) without necessarily reaching fixation if selection cannot purge the inflow of maladapted genotypes (Slatkin 1973). Local adaptation clines can be used to estimate dispersal distance (σ) if the selection coefficient (s) can be measured, which actually represents a serious challenge to most case studies. However, a measure of selection can sometimes be obtained using experimental populations or genotype frequency comparisons between larvae and adults sampled from the same cohort. By contrast, inferring dispersal from a neutral hitchhiker locus only requires the recombination rate with the selected locus (Box 3A). This can be more readily obtained by studying the signature left by selection in

Box 3: Using selected and hitchhiker loci to infer genetic connectivity.

Plots show the chromosomal and geographic signatures of selection under four different selective processes. Selected and neutral loci are colored in red and green, respectively. Genetic differentiation (F_{ST}) along the chromosome is measured between spatial coordinates −500 and 500. (A) A local adaptation cline lying at the frontier between two environments where selection acts in opposite directions (s = 0.1, σ = 30). The cline width parameter (w) is defined as the inverse of the maximum slope at the cline center, and k is a coefficient that depends on the selection regime (Slatkin 1973; Nagylaki 1975; Endler 1977; Barton and Gale 1993; Kruuk et al. 1999). A neutral hitchhiker locus with a recombination rate r with the selected locus makes a shift (Δp) in the central region of the cline, and an external gradient of allele frequency ($\partial p/\partial x$) directly outside the cline (Barton 1979b). (B) Hybrid zone cline between two partially reproductively isolated populations with selection acting against hybrid genotypes (s = 0.5). The amount of linkage disequilibrium (D) between selected loci is measured after dispersal at the center of the overlapping clines. (C) A tail of introgression produced by the inflow of foreign alleles entering a subdivided population (see Fig 3 for details). (D) Local connectivity patterns revealed by a global sweep. An unconditionally favorable mutation (s = 0.05) appears on the left side of a chain of demes (at an initial frequency of $1/2N_e$) and then propagates to the right side from deme to deme (m = 0.01), leaving behind a complex allele frequency pattern at a neutral hitchhiking locus (r = 0.001). Local connectivity between adjacent demes is transiently revealed by the structure of the neutral hitchhiking locus, as long as gene flow re-homogenizes allele frequencies. The chromosomal signatures of selection can take the form of narrow regions of differentiation (A), large genomic islands (B), or shoulders of differentiation (C and D) centered on the selected loci.



Glossary

Connectivity : *The exchange of individuals among populations or subpopulations. Lowe and Allendorf (2010) interestingly distinguished demographic connectivity and genetic connectivity*

Demographic connectivity : *The relative contribution of net immigration and local recruitment to the population growth rate of a population. Depends both on intrinsic characteristics (survival, reproduction, emigration) of the focal population and the extrinsic contribution of dispersal from other populations*

Genetic connectivity : *The absolute number of individuals coming into the focal population through immigration from other populations, as measured with genetic data; this may be a measure of individual movement or of gene flow according to the approach used*

Dispersal/migration : *The movement of individuals between populations*

Dispersal/migration rate : *The probability of a randomly sampled individual being an immigrant*

Genomic islands of differentiation : *A region of the genome where genetic differentiation increases above its genomewide average due to the presence of a genetic barrier to gene flow*

Hybrid zone : *A region where genetically distinct populations are in contact and interbreed*

Introgression : *The movement of genes between populations or species due to repeated backcrossing*

Local adaptation : *Higher performance of individuals in the habitat where they were born compared to immigrants*

Metapopulation : *A group of subpopulations exchanging migrants*

Next-generation Sequencing (NGS) : *High-throughput sequencing techniques (e.g. Illumina sequencing) in contrast to Sanger-based sequencing*

Outlier loci : *Loci with atypical patterns of genetic differentiation indicative of direct or indirect selection processes*

Partially reproductively isolated species : *Species that still exchange genes through introgressive hybridization*

Population : *A group of individuals living in the same habitat and reproducing with each other*

Species : *A group of individuals which are not interbreeding with other such groups (i.e. sensu* biological species definition)

Subpopulation, deme : *A group of individuals within a population that mate randomly and exchange migrants with other such groups*

the chromosomal neighborhood of individual outlier loci. For instance, resequencing the region around outliers may help to determine which polymorphism is actually under selection (i.e. the one showing the highest F_{ST} value, surrounded by decreasing differentiation on both sides; Box 3A) and provides data to estimate local recombination rates around the selected locus without needing a recombination map (Stumpf and McVean 2003). The chromosomal signature left by local selection in high gene flow species is usually limited to very narrow regions, even when selection acts on *de novo* mutations (Fig. 2). Therefore, high-density genome scans are usually required for efficiently detecting local adaptation loci.

As with parentage assignment methods, the dispersal parameter estimated from local adaptation clines is mostly relevant over short time scales in the geographic area where the shift in allele frequency is observed. However, discordant clines arising in distinct locations in response to spatially uncorrelated selective factors should provide independent local estimates of dispersal across a species range. Local adaptation clines might thus offer valuable alternatives to estimate migration in high gene flow marine species, keeping in mind that the underlying models assume that each cline evolves independently. Therefore, caution must be taken in distinguishing oligogenic from highly polygenic clines. For instance, a high-density genome scan in *Drosophila melanogaster* revealed the existence of several latitudinal clines (Fabian et al. 2012) that geographically overlap with classical clines attributed to local adaptation (e.g. the *Adh* locus, Berry and Kreitman 1993). As for *Drosophila* (Bergland et al. 2015), some classical clines found in marine organisms, such as the *Ldh* cline in

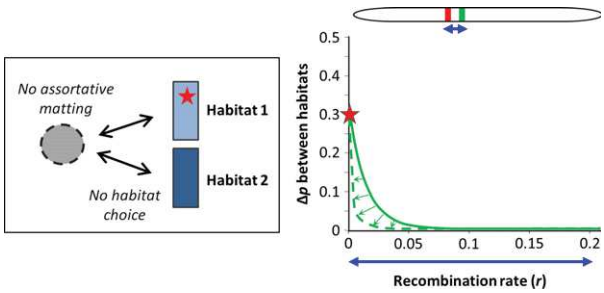


Figure 2 The chromosomal signature of local selection acting on a *de novo* mutation in panmixia. We consider a two habitats Levene's model (Levene 1953) represented in the left box, with random mating (in the dotted circle) and random dispersal (arrows) across two habitats of equal size (rectangles). A new selected mutation (allele *a*, red star) appears in habitat 1 on a haplotype bearing rare neutral variants (in green) at variable recombination distances (the initial frequency is $1/2N_e$). The selected mutation has symmetrical antagonistic effects on the fitness of genotypes with respect to habitat (Habitat 1: $\omega_{AA}/\omega_{Aa} = 0.5$, $\omega_{aa}/\omega_{Aa} = 2$; Habitat 2: $\omega_{AA}/\omega_{Aa} = 2$, $\omega_{aa}/\omega_{Aa} = 0.5$). At equilibrium, varying selection among genotypes and habitats results in differentiation between habitats at the selected locus (in this example $\Delta p \approx 0.3$). During the progress toward equilibrium, neutral variants hitchhike with the selected allele, transiently producing a narrow chromosomal region where genetic differentiation is increased around the selected locus (green line). As the selected allele progressively recombines away from its haplotypic background, differentiation at neutral alleles rapidly vanishes (green arrows). After a few thousands of generations, differentiation is almost limited to the selected locus (dashed green line).

the killifish *Fundulus heteroclitus* (Powers and Place 1978), turned out to occur in secondary contact zones (Durand et al. 2009). This suggests that some of the few clinal outliers that were detected trough candidate gene or low-density genome scans may only reflect the emerged part of the iceberg and that polygenic clines and cryptic hybrid zones coinciding with environmental boundaries may be more common than usually believed (Bierne et al. 2011). When significant linkage disequilibrium is detected among clines, the hybrid zone theory offers a more appropriate framework to infer dispersal.

Using hybrid zone clines to infer dispersal

Many clines evidenced in marine population genetics studies actually result from selection acting at multiple loci, as revealed by the finding of concordant clines in contact zones between hybridizing taxa, that is hybrid zones (Duggins et al. 1995; Bierne et al. 2003; Sotka et al. 2004; Murray and Hare 2006; Zbawicka et al. 2014). In such clines, each locus cumulates the indirect selective effects from other loci (transmitted through linkage disequilibrium) in addition to its own selection coefficient (Barton 1983; Kruuk et al. 1999). The magnitude of indirect effects depends on the amount of linkage disequilibrium and therefore on selection, recombination, and dispersal. The associations among selected alleles in hybrid zones can be used in combination with cline width to infer dispersal (Box 3B, Barton and Gale 1993). As for single locus clines, outlier loci showing concordant clines are not necessarily the actual targets of selection but more likely neutral loci presenting various degrees of linkage with the genes involved in the barrier. Therefore, the shift in allele frequency in the central region of the cline is often much less than 1 for neutral markers, and linkage disequilibrium needs to be corrected for the effect of introgression to estimate dispersal distance (Box 3B).

The cumulative effects of direct selection and indirect selection acting on other loci produce a typical cline shape characterized by a central sigmoid step with two exponential tails of introgression on either side (Barton 1983; Barton and Gale 1993). Allele frequency data collected across a hybrid zone transect can be used to fit a model of cline shape and estimate its parameters, including cline center and width within the narrow region of abrupt change (Szymura and Barton 1986). Hybrid zones analysis programs like *HZAR* provide useful functions for fitting clines along geographic transects (Derryberry et al. 2014).

Genetic tagging in hybrid zones

A conceptually different approach to estimating connectivity in contact zones is to perform individual genetic assignments to identify migrants. This approach which is similar to the one detailed in the above section (i.e. direct estimates of genetic connectivity) takes advantage of the substantial genetic differences existing between populations or species that are on both sides of the hybrid zone. Minimal dispersal distances can be obtained through the identification of parental genotypes that crossed a hybrid zone and successfully settled in a foreign parental population or species. An even more precise estimation of larval dispersal distance can be made when the source of dispersing larvae is known, as for first-generation hybrids dispersing outside a hybrid zone. Using this strategy, patterns of larval movement among neighboring patches of blue mussels have been examined by measuring realized larval dispersal based on the genetic identification of recently settled juveniles (Gilg and Hilbish 2003). This approach provides an interesting alternative to the measures of dispersal offered by the analysis of genetic clines. In the blue mussel example, both approaches provide comparable estimates: Gilg and Hilbish (2003) found a dispersal distance of 30 km which is in accordance with the 38 km width of the *LAP* cline in Long Island (Lassen and Turano 1978) and the 52 km width of the cline between *M. edulis* and *M. trossulus* in the Oresund (Väinölä and Hvilsom 1991). Because genetic tagging relies on a clear distinction between parental genotypes, introgressed individuals, and real hybrids, individual

assignments should be done using the most highly differentiated (and preferentially diagnostic) markers identified in genome scans.

Using introgression tails to reveal cryptic population structure

Previous methods based on cline width analysis are constrained in their application by the geographic localization of cline centers. However, estimates of population connectivity are often required outside these singular regions, for instance when it is necessary to determine whether there is limited dispersal between populations within areas delimited by ecological or biogeographic boundaries, a relatively common concern for conservation and stock management issues (Allendorf et al. 2010). A potential solution, when the migration–drift equilibrium is not informative, is to search for evidence of spatial structure revealed by introgression (Gagnaire et al. 2011). Introgression may generate gradients (or steps) in allele frequencies along a geographic axis originating at the edge of a contact zone. These tails of introgression may extend to large distances beyond cline centers and can arise for several reasons.

The first one is the free diffusion of neutral alleles following secondary contact between two genetically differentiated populations, or two partially reproductively isolated species. This process creates a transient gradient of

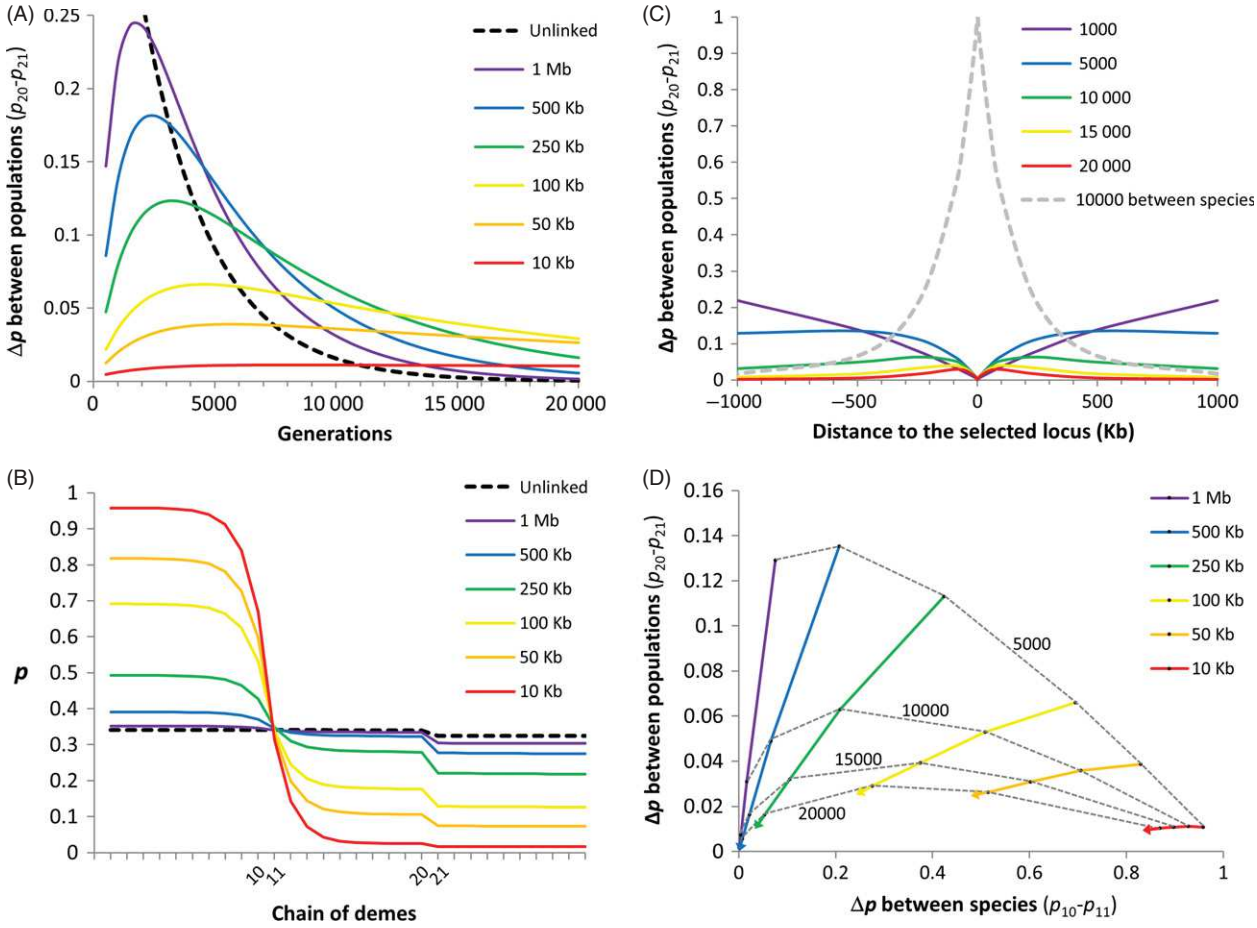


Figure 3 Using the inflow of foreign alleles to reveal within-species connectivity patterns. At generation zero, two partially reproductively isolated species meet on a linear stepping-stone model between demes 10 and 11 and start to exchange genes. The auto-recruitment rate is $1 - m$, and migration to adjacent demes is $m/2$ (with $m = 0.5$). A weak barrier to gene flow ($m = 0.01$) was set between demes 20 and 21, in the middle of the range of the species localized on the right side. Strong selection ($s = 0.5$) acts against heterozygote genotypes at a reproductive isolation locus, which is linked to neutral markers located at variable recombination distances (from closely linked to unlinked). A recombination rate of 1 cM per Mb was used to convert genetic into physical distances. (A) The step size, calculated as the difference in allele frequency between demes 20 and 21 (Δp), as a function of the number of generations postcontact. (B) Spatial allele frequency patterns after 10 000 generations of introgression showing the frequency step between demes 20 and 21. (C) The step size between demes 20 and 21 as a function of the physical distance to the reproductive isolation locus. (D) The step size between demes 20 and 21 as a function of the difference in allele frequency between species.

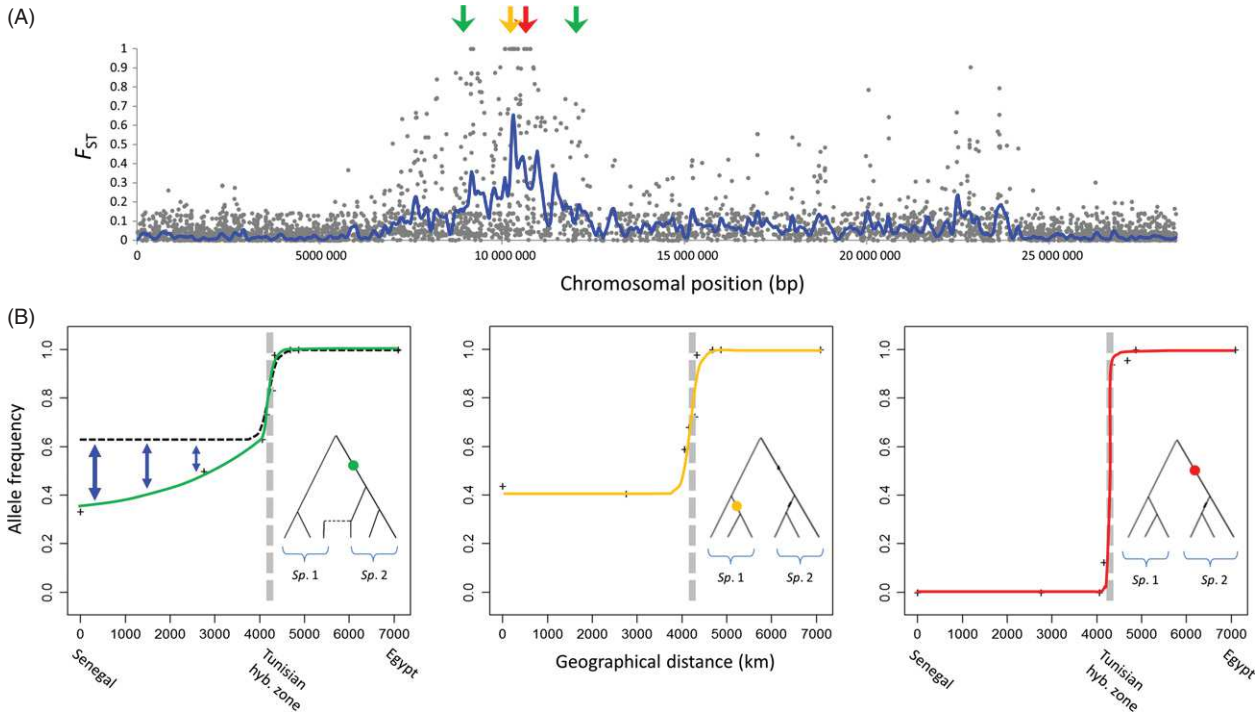


Figure 4 Genomic islands of differentiation and the information therein. (A) A genomic island of differentiation between Atlantic and Mediterranean sea bass lineages (*Dicentrarchus labrax*) on chromosome 7 (RAD-sequencing data from Tine et al. 2014). (B) Geographic clines between two partially reproductively isolated species of sole, *Solea senegalensis* (*Sp*. 1, left side) and *Solea aegyptiaca* (*Sp*. 2, right side) assessed by RAD-Sequencing (A. Souissi, P.-A. Gagnaire, L. Bahri-Sfar, F. Bonhomme, unpublished). Red and orange clines correspond to expectations near reproductive isolation loci (i.e. at the center of a genomic island, where there is no introgression), for a diagnostic locus (red) and a locus only polymorphic in *S. senegalensis* (orange) due to incomplete lineage sorting. The green cline shows a gradient (or a tail) of introgression due to the inflow of *S. aegyptiaca* alleles in the *S. senegalensis* background. At this locus, the shared allele is a consequence of secondary introgression instead of incomplete lineage sorting. Such gradients of introgression are expected to be found at loci showing intermediate degrees of linkage with reproductive isolation loci (i.e. located in the periphery of a genomic island, where introgression is reduced but not zero). Introgression tails may be used to reveal cryptic genetic structure where freely recombining neutral loci remain uninformative (black dashed line), as it is the case in *S. senegalensis*.

introgressing alleles if the rate of introgression is higher or equal to the rate of homogenization within the introgressed population (i.e. the introgression/homogenization rate ratio is ≥1). Importantly, the gradient only appears if dispersal is spatially limited, otherwise spatial homogenization occurs immediately as foreign alleles enter the introgressed population. In order to illustrate how this mechanism can be used to detect a local barrier to dispersal, we simulated a contact zone between two partially reproductively isolated species and the introgression of foreign alleles within one of the two species which is geographically subdivided (Fig. 3). The extent of genetic differentiation within the introgressed species was measured between two populations separated by a weak barrier to gene flow (m = 0.01), other adjacent demes being otherwise highly connected in the standard linear stepping-stone model. During a few thousands of generations postcontact, introgression generates a step in allele frequency between the two populations of the introgressed species, and the step then disappears as allele frequencies equilibrate between species (the black dashed line Fig. 3A). Two important properties emerge from these simulations. As the introgression/homogenization rate ratio approaches 1, the magnitude of the frequency step decreases, but the maximum step magnitude is reached later and the step lasts longer. A direct application of these properties is that variable introgression rates among loci provide with the means to detect a weak barrier to gene flow even when introgression has started thousands of generations in the past. For instance, a snapshot taken after 10 000 generations of introgression shows that while the step has completely vanished at freely recombining neutral loci, neutral loci in partial linkage with reproductive isolation loci have retained the signal of differentiation between populations due to their reduced effective migration rate (Fig. 3B). Therefore, differential introgression between parapatric species can be used as a powerful tool to detect cryptic population structure outside the contact zone.

Tails of introgression may be also influenced by selection acting outside the tension zone. In this case, the gradient of

allele frequency within the range of the introgressed species may be steepened by a gradient of selection (e.g. an environmental gradient). Because secondary contact zones commonly coincide with environmental gradients (Bierne et al. 2011), introgression tails may be commonly encountered within biogeographic regions separated by environmental boundaries (e.g. the Baltic Sea).

These mechanisms show how much it is important to sample not only the whole distribution range of a species but also other divergent populations, or closely related species that live in parapatry or in sympatry before interpreting spatial genetic variation patterns (Gagnaire et al. 2011; Cullingham et al. 2013; Gosset and Bierne 2013). Now that NGS tools begin to reveal genomic islands of differentiation between cryptic species that were previously considered as populations of the same species (Hemmer-Hansen et al. 2013; Karlsen et al. 2013; Tine et al. 2014), polymorphisms located in the periphery of these islands may become a powerful new type of markers to infer connectivity within species, as illustrated in Fig. 4. Importantly, the spatial range of application of genomic-island associated loci could be large if markers are taken at various recombination distances from the central region of a genomic island of differentiation.

Hitchhiking clines

Another scenario that generates outlier loci happens during the spread of an unconditionally favorable allele in a spatially subdivided population. This process leaves a transient footprint at neutral markers in the chromosomal vicinity of the sweeping allele. When the overall genomic differentiation is low, as it is typically the case in marine species, this process generates an elevated level of differentiation on both sides of the selected locus (Bierne 2010), which corresponds to the locations of sweep shoulders (Schrider et al. 2015). The reason is that recombination progressively breaks the association between the selected locus and the hitchhiker neutral locus, while the sweeping wave propagates. Therefore, the hitchhiking effect is strong at the birthplace of the favorable mutation, while it progressively softens as the wave travels. The effect is stronger for intermediate recombination rates between the selected and the hitchhiker neutral locus, because when linkage is tight, associations remain during the spread of the wave while when linkage is loose the hitchhiking effect is weak right from the beginning. For a similar reason as for the case of introgression clines (Fig. 3C), global hitchhiking therefore generates two shoulders of differentiation on each side of the selected locus on the chromosome. In the deterministic model, the spatial structure generated is a gradient in allele frequency called 'hitchhiking cline', but when stochasticity is introduced, for example random genetic drift, the spatial structure can be more complex and sometimes results in nonmonotonic variations in allele frequency (a patchy genetic structure as shown in Box 3D). Detecting the genomic signature of a global sweep requires a high-density screening of the genetic differentiation in the chromosomal neighborhood of the selected locus. Therefore, only few examples that fit the predictions have been studied, with only two cases in marine species (in the blue mussel, Bierne 2010; and the stickleback, Roesti et al. 2014), and some possible cases in highly polymorphic terrestrial species such as maize (Gore et al. 2009; Bierne 2010) and nematodes (Jovelin et al. 2014). By adjusting the global hitchhiking model to the mussel data, it has been possible to estimate the minimal migration rate needed to obtain the observed F_{ST} value between the two geographically distant populations of *M. edulis* (Faure et al. 2008) which proved to be surprisingly low ($m < 10^{-8}$), as well as the position of the selected locus (-3 kb 5′ of the start codon of the *EF1α* gene), the selection coefficient ($s = 0.01$), and incidentally the local recombination rate of the chromosomal region ($\rho = 1.7$ cM/Mb, Bierne 2010). This result nicely closes the loop of our argumentation by showing how two populations of mussels that are demographically largely independent for thousands of years do not depart from apparent genetic panmixia. Recent analysis based on NGS data (Fraïsse et al. 2015) revealed that deep sampling of the neutral fraction of the genome does not reveal a clear genetic structure between the two populations and that local adaptation is either extremely rare or extremely difficult to evidence (Gosset et al. 2014). Only the indirect effect of selection transiently generated at a linked neutral hitchhiker locus has revealed a sufficiently clear pattern to demonstrate demographic independence.

Conclusion

Substantial progresses in our understanding of connectivity in nonmodel organism can be achieved with large population genomic datasets. High-density genome scans have reached the power to detect outlying patterns of genetic differentiation at different spatial scales, enabling conservation geneticists to identify genetic differences reflecting restriction to gene flow where classical neutral markers were hitherto most often largely uninformative, as in high gene flow species. The scope of the applications of outlier loci for assessing connectivity patterns in marine species needs further investigations, in particular through gathering a larger set of empirical data. Some of the methodologies that were proposed in this review are still underexplored, and we hope that our work will stimulate new research to test how broadly they are applicable to nonmodel marine species. Although spatially explicit methods are directly applicable to continuously distributed ses-

sile species, selected and hitchhiker loci also have the potential to reveal cryptic genetic structure in migratory species with natal homing (Gagnaire et al. 2011) or feeding migrations. A growing question will be to determine whether all the genetic differences revealed by outlier loci are relevant for conservation and species management. Genome scans will probably confirm the picture of major biogeographic boundaries as hotspots of cryptic genetic structure between populations and partially reproductively isolated species pairs. They may also reveal new and unexpected barriers to gene flow. Such zones are likely to delineate stocks and populations that are important from a conservation point of view. Besides, genome scans may also reveal unusual outlier patterns that are difficult to relate to a clearly identified evolutionary mechanism. The shift to using selected and hitchhiker loci will probably open this can of worms, irrespective to their utility to assess connectivity in the marine realm.

Acknowledgements

This study was supported by the *Marine French Connection* research group GDR CNRS-Ifremer 3445 *MarCo* and the ANR grant LABRAD-SEQ 11-PDOC-009-01. The authors thank the Associate Editor Craig Primmer, Arnaud Estoup, Robin Waples, and two anonymous reviewers for their constructive comments. This is ISEM publication 2015-116.

Literature cited

Addison, J. A., B. S. Ort, K. A. Mesa, and G. H. Pogson 2008. Range-wide genetic homogeneity in the California sea mussel (*Mytilus californianus*): a comparison of allozymes, nuclear DNA markers, and mitochondrial DNA sequences. Molecular Ecology **17**:4222–4232.

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart 2010. Genomics and the future of conservation genetics. Nature Reviews Genetics **11**:697–709.

Almany, G. R., M. L. Berumen, S. R. Thorrold, S. Planes, and G. P. Jones 2007. Local replenishment of coral reef fish populations in a marine reserve. Science **316**:742–744.

Almany, G. R., R. J. Hamilton, M. Bode, M. Matawai, T. Potuku, P. Saenz-Agudelo, S. Planes et al. 2013. Dispersal of grouper larvae drives local resource sharing in a coral reef fishery. Current Biology **23**:626–630.

Ansmann, I. C., G. J. Parra, J. M. Lanyon, and J. M. Seddon 2012. Fine-scale genetic population structure in a mobile marine mammal: inshore bottlenose dolphins in Moreton Bay, Australia. Molecular Ecology **21**:4472–4485.

Aurelle, D., and J.-B. Ledoux 2013. Interplay between isolation by distance and genetic clusters in the red coral *Corallium rubrum*: insights from simulated and empirical data. Conservation Genetics **14**:705–716.

Barton, N. H. 1979a. The dynamics of hybrid zones. Heredity **43**:341–359.

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

Barton, N. H. 1983. Multilocus clines. Evolution **37**:454–471.

Barton, N. H., and K. S. Gale 1993. Genetic analysis of hybrid zones. In: R. G. Harrison, ed. Hybrid Zones and the Evolutionary Process, pp. 13–45. Oxford University Press, New York.

Bergland, A. O., R. Tobler, J. Gonzalez, P. Schmidt, and D. Petrov. 2015. Secondary contact and local adaptation contribute to genome-wide patterns of clinal variation in *Drosophila melanogaster*. bioRxiv, doi:10.1101/009084.

Berry, A., and M. Kreitman 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134**:869–893.

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

Bierne, N., P. Borsa, C. Daguin, D. Jollivet, F. Viard, F. Bonhomme, and P. David 2003. Introgression patterns in the mosaic hybrid zone between *Mytilus edulis* and *M. galloprovincialis*. Molecular Ecology **12**:447–461.

Bierne, N., J. Welch, E. Loire, F. Bonhomme, and P. David 2011. The coupling hypothesis: why genome scans may fail to map local adaptation genes. Molecular Ecology **20**:2044–2072.

Bierne, N., D. Roze, and J. J. Welch 2013. Pervasive selection or is it. . .? why are FST outliers sometimes so frequent? Molecular Ecology **22**:2061–2064.

Blair, C., D. E. Weigel, M. Balazik, A. T. H. Keeley, F. M. Walker, E. Landguth, S. A. M. Cushman et al. 2012. A simulation-based evaluation of methods for inferring linear barriers to gene flow. Molecular Ecology Resources **12**:822–833.

Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. SanCristobal 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics **186**:241–262.

Bradbury, I. R., S. Hubert, B. Higgins, T. Borza, S. Bowman, I. G. Paterson, P. V. R. Snelgrove et al. 2010. Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. Proceedings of the Royal Society B: Biological Sciences **277**:3725–3734.

Broquet, T., and E. J. Petit 2009. Molecular estimation of dispersal for ecology and population genetics. Annual Review of Ecology, Evolution, and Systematics **40**:193–216.

Broquet, T., J. Yearsley, A. H. Hirzel, J. Goudet, and N. Perrin 2009. Inferring recent migration rates from individual genotypes. Molecular Ecology **18**:1048–1060.

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

Charlesworth, B., M. Nordborg, and D. Charlesworth 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical research **70**:155–174.

Christie, M. R., B. N. Tissot, M. A. Albins, J. P. Beets, Y. Jia, D. M. Ortiz, S. E. Thompson et al. 2010. Larval connectivity in an effective network of marine protected areas. PLoS ONE **5**:e15715.

Clobert, J., M. Baguette, T. G. Benton, J. M. Bullock, and S. Ducatez 2012. Dispersal Ecology and Evolution. Oxford University Press, Oxford.

Colbeck, G. J., J. Turgeon, P. Sirois, and J. J. Dodson 2011. Historical introgression and the role of selective vs. neutral processes in structuring nuclear genetic variation (AFLP) in a circumpolar marine fish, the capelin (*Mallotus villosus*). Molecular Ecology **20**:1976–1987.

Corander, J., P. Waldmann, and M. J. Sillanpää 2003. Bayesian analysis of genetic differentiation between Populations. Genetics **163**:367–374.

Cullingham, C. I., J. E. Cooke, and D. W. Coltman 2013. Effects of introgression on the genetic population structure of two ecologically and economically important conifer species: lodgepole pine (*Pinus contorta* var. *latifolia*) and jack pine (*Pinus banksiana*). Genome **56**:577–585.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics **12**:499–510.

De Mita, S., A.-C. Thuillet, L. Gay, N. Ahmadi, S. Manel, J. Ronfort, and Y. Vigouroux 2013. Detecting selection along environmental gradients : analysis of eight methods and their effectiveness for outbreeding and selfing populations. Molecular Ecology **22**:1383–1399.

De Villemereuil, P., E. Frichot, E. Bazin, O. François, and O. Gaggiotti 2014. Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology **23**:2006–2019.

Derryberry, E. P., G. E. Derryberry, J. M. Maley, and R. T. Brumfield 2014. HZAR: hybrid zone analysis using an R software package. Molecular Ecology Resources **14**:652–663.

DeWoody, J. A., and J. C. Avise 2000. Microsatellite variation in marine, freshwater and anadromous fishes compared with other animals. Journal of Fish Biology **56**:461–473.

Duforet-Frebourg, N., E. Bazin, and M. G. Blum. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. Molecular Biology and Evolution, **31**:2483–2495.

Duggins, C. F., A. A. Karlin, T. A. Mousseau, and K. G. Relyea 1995. Analysis of a hybrid zone in *Fundulus majalis* in a northeastern Florida ecotone. Heredity **74**:117–128.

Durand, E., F. Jay, O. E. Gaggiotti, and O. François 2009. Spatial inference of admixture proportions and secondary contact zones. Molecular Biology and Evolution **26**:1963–1973.

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

Excoffier, L., T. Hofer, and M. Foll 2009. Detecting loci under selection in a hierarchically structured population. Heredity **103**:285–298.

Fabian, D. K., M. Kapun, V. Nolte, R. Kofler, P. S. Schmidt, C. Schlötterer, and T. Flatt 2012. Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. Molecular Ecology **21**:4748–4769.

Faure, M. F., P. David, F. Bonhomme, and N. Bierne 2008. Genetic hitchhiking in a subdivided population of *Mytilus edulis*. BMC Evolutionary Biology **8**:164.

Foll, M., O. E. Gaggiotti, J. T. Daub, A. Vatsiou, and L. Excoffier 2014. Widespread signals of convergent adaptation to high altitude in Asia and America. The American Journal of Human Genetics **95**:394–407.

Fourcade, Y., A. Chaput-Bardy, J. Secondi, C. Fleurant, and C. Lemaire 2013. Is local selection so widespread in river organisms? Fractal geometry of river networks leads to high bias in outlier detection. Molecular Ecology **22**:2065–2073.

Fraïsse, C., C. Roux, J. J. Welch, and N. Bierne 2014. Gene-flow in a mosaic hybrid zone: is local introgression adaptive? Genetics **197**:939–951.

Fraïsse, C., K. Belkhir, J. J. Welch, and N. Bierne. 2015. Local inter-species introgression is the main cause of outlying levels of intra-specific differentiation in mussels. Molecular Ecology [in press].

François, O., and E. Durand 2010. Spatially explicit Bayesian clustering models in population genetics. Molecular Ecology Resources **10**:773–784.

Funk, W. C., J. K. McKay, P. A. Hohenlohe, and F. W. Allendorf 2012. Harnessing genomics for delineating conservation units. Trends in Ecology & Evolution **27**:489–496.

Gaggiotti, O. E., F. Jones, W. M. Lee, W. Amos, J. Harwood, and R. A. Nichols 2002. Patterns of colonization in a metapopulation of grey seals. Nature **416**:424–427.

Gagnaire, P.-A., Y. Minegishi, S. Zenboudji, P. Valade, J. Aoyama, and P. Berrebi 2011. Within-population structure highlighted by differential introgression across semipermeable barriers to gene flow in *Anguilla marmorata*. Evolution **65**:3413–3427.

Gagnaire, P.-A., E. Normandeau, C. Côté, M. M. Hansen, and L. Bernatchez 2012. The genetic consequences of spatially varying selection in the panmictic American eel (*Anguilla rostrata*). Genetics **190**:725–736.

Gao, H., S. Williamson, and C. D. Bustamante 2007. A Marko Chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. Genetics **76**:1635–1651.

Gilg, M. R., and T. J. Hilbish 2003. The geography of marine larval dispersal: coupling genetics with fine-scale physical oceanography. Ecology **84**:2989–2998.

Gompert, Z., and C. A. Buerkle 2013. Analyses of genetic ancestry enable key insights for molecular ecology. Molecular Ecology **22**:5278–5294.

Gore, M. A., J. M. Chia, R. J. Elshire, Q. Sun, E. S. Ersoz, B. L. Hurwitz, J. A. Peiffer et al. 2009. A first-generation haplotype map of maize. Science **326**:1115–1117.

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

Gosset, C. C., J. Do Nascimento, M. T. Augé, and N. Bierne 2014. Evidence for adaptation from standing genetic variation on an antimicrobial peptide gene in the mussel *Mytilus edulis*. Molecular Ecology **23**:3000–3012.

Gravel, S. 2012. Population genetics models of local ancestry. Genetics **191**:607–619.

Hedgecock, D., P. H. Barber, and S. Edmands 2007. Genetic approaches to measuring connectivity. Oceanography **20**:70–79.

Hellberg, M. E. 2009. Gene flow and isolation among populations of marine animals. Annual Review of Ecology, Evolution, and Systematics **40**:291–310.

Hemmer-Hansen, J., E. E. Nielsen, N. O. Therkildsen, M. I. Taylor, R. Ogden, A. J. Geffen, D. Bekkevold et al. 2013. A genomic island linked to ecotype divergence in Atlantic cod. Molecular Ecology **22**:2653–2667.

Jones, G. P., S. Planes, and S. R. Thorrold 2005. Coral reef fish larvae settle close to home. Current Biology **15**:1314–1318.

Jost, L. O. U. 2008. GST and its relatives do not measure differentiation. Molecular Ecology **17**:4015–4026.

Jovelin, R., J. S. Comstock, A. D. Cutter, and P. C. Phillips 2014. A recent global selective sweep on the age-1 phosphatidylinositol 3-OH kinase regulator of the insulin-like signaling pathway within *Caenorhabditis remanei*. G3: Genes | Genomes | Genetics **4**:1123–1133.

Karlsen, B. O., K. Klingan, A. Emblem, T. E. Jørgensen, A. Jueterbock, T. Furmanek, G. Hoarau et al. 2013. Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. Molecular Ecology **22**:5098–5111.

Koehn, R. K., R. I. Newell, and F. Immermann 1980. Maintenance of an aminopeptidase allele frequency cline by natural selection. Proceedings of the National Academy of Sciences **77**:5385–5389.

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.

Lamichhaney, S., A. M. Barrio, N. Rafati, G. Sundström, C.-J. Rubin, E. R. Gilbert, J. Berglund et al. 2012. Population-scale sequencing reveals

genetic differentiation due to local adaptation in Atlantic herring. Proceedings of the National Academy of Sciences **109**:19345–19350.

Langley, C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, J. E. Pool, S. A. Langley et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. Genetics **192**:533–598.

Lassen, H. H., and F. J. Turano 1978. Clinal variation and heterozygote deficit at the Lap-locus in *Mytilus edulis*. Marine Biology **49**:245–254.

Latch, E. K., G. Dharmarajan, J. C. Glaubitz, and O. E. Rhodes Jr 2006. Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. Conservation Genetics **7**:295–302.

Le Corre, V., and A. Kremer 2012. The genetic differentiation at quantitative trait loci under local adaptation. Molecular Ecology **21**:1548–1566.

Ledoux, J. B., J. Garrabou, O. Bianchimani, P. Drap, J. P. Féral, and D. Aurelle 2010. Fine-scale genetic structure and inferences on population biology in the threatened Mediterranean red coral, *Corallium rubrum*. Molecular Ecology **19**:4204–4216.

Lenormand, T., T. Guillemaud, D. Bourguet, and M. Raymond 1998. Evaluating gene flow using selected markers: a case study. Genetics **149**:1383–1392.

Levene, H. 1953. Genetic equilibrium when more than one ecological niche is available. The American Naturalist **87**:331–333.

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

Limborg, M. T., S. J. Helyar, M. De Bruyn, M. I. Taylor, E. E. Nielsen, R. O. B. Ogden, G. R. Carvalho et al. 2012. Environmental selection on transcriptome-derived SNPs in a high gene flow marine fish, the Atlantic herring (*Clupea harengus*). Molecular Ecology **21**:3686–3703.

Lotterhos, K., and M. Whitlock 2014. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. Molecular Ecology **23**:2178–2192.

Lowe, W. H., and F. W. Allendorf 2010. What can genetics tell us about population connectivity? Molecular Ecology **19**:3038–3051.

Luikart, G., P. R. England, D. Tallmon, S. Jordan, and P. Taberlet 2003. The power and promise of population genomics: from genotyping to genome typing. Nature Reviews Genetics **4**:981–994.

Lukoschek, V., and R. Shine 2012. Sea snakes rarely venture far from home. Ecology and Evolution **2**:1113–1121.

Marko, P. B., and M. W. Hart 2011. The complex analytical landscape of gene flow inference. Trends in Ecology & Evolution **26**:448–456.

Maynard Smith, J., and J. Haigh 1974. The hitch-hiking effect of a favourable gene. Genetical Research **23**:23–35.

McCusker, M. R., and P. Bentzen 2010. Positive relationships between genetic diversity and abundance in fishes. Molecular Ecology **19**:4852–4862.

van der Meer, M. H., J.-P. A. Hobbs, G. P. Jones, and L. van Herwerden 2012. Genetic connectivity among and self-replenishment within island populations of a restricted range subtropical reef fish. PLoS ONE **7**:e49660.

Meirmans, P. G. 2012. The trouble with isolation by distance. Molecular Ecology **21**:2839–2846.

Meirmans, P. G., and P. W. Hedrick 2011. Assessing population structure: FST and related measures. Molecular Ecology Resources **11**:5–18.

Mokhtar-Jamaï, K., M. Pascual, J. B. Ledoux, R. Coma, J. P. Féral, J. Garrabou, and D. Aurelle 2011. From global to local genetic structuring in the red gorgonian *Paramuricea clavata*: the interplay between oceanographic conditions and limited larval dispersal. Molecular Ecology **20**:3291–3305.

Murray, M. C., and M. P. Hare 2006. A genomic scan for divergent selection in a secondary contact zone between Atlantic and Gulf of Mexico oysters, *Crassostrea virginica*. Molecular Ecology **15**:4229–4242.

Nagylaki, T. 1975. Conditions for the existence of clines. Genetics **80**:595–615.

Narum, S. R., and J. E. Hess 2011. Comparison of F_{ST} outlier tests for SNP loci under selection. Molecular Ecology Resources **11**(Suppl. 1):184–194.

Nielsen, E. E., J. Hemmer-Hansenn, P. F. Larsen, and D. Bekkevold 2009. Population genomics of marine fishes: identifying adaptive variation in space and time. Molecular Ecology **18**:3128–3150.

Nielsen, E. E., A. Cariani, E. Mac Aoidh, G. E. Maes, I. Milano, R. Ogden, M. Taylor et al. 2012. Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. Nature Communications **3**:851.

Novembre, J., T. Johnson, K. Bryc, Z. Kutalik, A. R. Boyko, A. Auton, A. Indap et al. 2008. Genes mirror geography within Europe. Nature **456**:98–101.

O'Connor, T. D., W. Fu, NHLBI GO Exome Sequencing Project, ESP Population Genetics and Statistical Analysis Working Group, E. Turner, J. C. Mychaleckyj, B. Logsdon et al. 2014. Rare variation facilitates inferences of fine-scale population structure in humans. Molecular Biology and Evolution **32**:653–660.

Paetkau, D., R. Slade, M. Burden, and A. Estoup 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Molecular Ecology **13**:55–65.

Palamara, P. F., and I. Pe'er. 2013. Inference of historical migration rates via haplotype sharing. Bioinformatics **29**:i180–i188.

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

Palumbi, S. R. 2003. Population genetics, demographic connectivity, and the design of marine reserves. Ecological Applications **13**:S146–S158.

Patterson, N., A. L. Price, and D. Reich 2006. Population structure and eigenanalysis. PLoS Genetics **2**:e190.

Perrier, C., R. Guyomard, J.-L. Bagliniere, and G. Evanno 2011. Determinants of hierarchical genetic structure in Atlantic salmon populations: environmental factors vs. anthropogenic influences. Molecular Ecology **20**:4231–4245.

Pinsky, M. L., H. R. Montes Jr, and S. R. Palumbi 2010. Using isolation by distance and effective density to estimate dispersal scales in anemonefish. Evolution **64**:2688–2700.

Planes, S., G. P. Jones, and S. R. Thorrold 2009. Larval dispersal connects fish populations in a network of marine protected areas. Proceedings of the National Academy of Sciences **106**:5693–5697.

Pogson, G. H., C. T. Taggart, K. A. Mesa, and R. G. Boutilier 2001. Isolation by distance in the Atlantic cod, *Gadus morhua*, at large and small geographic scales. Evolution **55**:131–146.

Pool, J. E., and R. Nielsen 2009. Inference of historical changes in migration rate from the lengths of migrant tracts. Genetics **181**:711–719.

Popescu, A.-A., A. L. Harper, M. Trick, I. Bancroft, and K. T. Huber 2014. A novel and fast approach for population structure inference using kernel-PCA and optimization (PSIKO). Genetics **114**:171314.

Powers, D. A., and A. R. Place 1978. Biochemical genetics of *Fundulus heteroclitus* (L.). I. Temporal and spatial variation in gene frequencies of Ldh-B, Mdh-A, Gpi-B, and Pgm-A. Biochemical Genetics **16**:593–607.

Pritchard, J. K., and A. Di Rienzo 2010. Adaptation – not by sweeps alone. Nature Reviews Genetics **11**:665–667.

Pritchard, J. K., M. Stephens, and P. Donnelly 2000. Inference of population structure using multilocus genotype data. Genetics **155**:945–959.

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

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

Raj, A., M. Stephens, and J. K. Pritchard 2014. Variational inference of population structure in large SNP datasets. Genetics **114**:164350.

Ralph, P., and G. Coop 2010. Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics **186**:647–668.

Roesti, M., A. P. Hendry, W. Salzburger, and D. Berner 2012. Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. Molecular Ecology **21**:2852–2862.

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

Rose, C. G., K. T. Paynter, and M. P. Hare 2006. Isolation by distance in the eastern oyster, *Crassostrea virginica*, in Chesapeake Bay. Journal of Heredity **97**:158–170.

Rousset, F. 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics **145**:1219–1228.

Rousset, F. 2000. Genetic differentiation between individuals. Journal of Evolutionary Biology **13**:58–62.

Rousset, F. 2004. Genetic Structure and Selection in Subdivided Populations, vol **40**. Princeton University Press, Princeton.

Saenz-Agudelo, P., G. P. Jones, S. R. Thorrold, and S. Planes 2011. Connectivity dominates larval replenishment in a coastal reef fish metapopulation. Proceedings of the Royal Society B: Biological Sciences **278**:2954–2961.

Schmidt, P. S., E. A. Serrão, G. A. Pearson, C. Riginos, P. D. Rawson, T. J. Hilbish, S. H. Brawley et al. 2008. Ecological genetics in the North Atlantic: environmental gradients and adaptation at specific loci. Ecology **89**(sp11):S91–S107.

Schrider, D. R., F. K. Mendes, M. W. Hahn, and A. D. Kern 2015. Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. Genetics **200**:267–284.

Schwartz, M. K., and K. S. McKelvey 2009. Why sampling scheme matters: the effect of sampling scheme on landscape genetic results. Conservation Genetics **10**:441–452.

Selkoe, K. A., and R. J. Toonen 2011. Marine connectivity: a new look at pelagic larval duration and genetic metrics of dispersal. Marine Ecology Progress Series **436**:291–305.

Selkoe, K. A., J. R. Watson, C. White, T. B. Horin, M. Iacchei, S. Mitarai, D. A. Siegel et al. 2010. Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. Molecular Ecology **19**:3708–3726.

Slatkin, M. 1973. Gene flow and selection in a cline. Genetics **75**:733–756.

Sotka, E. E., and S. R. Palumbi 2006. The use of genetic clines to estimate dispersal distances of marine larvae. Ecology **87**:1094–1103.

Sotka, E. E., J. P. Wares, J. A. Barth, R. K. Grosberg, and S. R. Palumbi 2004. Strong genetic clines and geographical variation in gene flow in the rocky intertidal barnacle *Balanus glandula*. Molecular Ecology **13**:2143–2156.

Stapley, J., J. Reger, P. G. D. Feulner, C. Smadja, J. Galindo, R. Ekblom, C. Bennison et al. 2010. Adaptation genomics: the next generation. Trends in Ecology and Evolution **25**:705–712.

Storz, J. F. 2005. Using genome scans of DNA polymorphism to infer adaptive population divergence. Molecular Ecology **14**:671–688.

Stumpf, M. P., and G. A. McVean 2003. Estimating recombination rates from population-genetic data. Nature reviews. Genetics **4**:959–968.

Swearer, S. E., J. S. Shima, M. E. Hellberg, S. R. Thorrold, G. P. Jones, D. R. Robertson, S. G. Morgan et al. 2002. Evidence of self-recruitment in demersal marine populations. Bulletin of Marine Science **70** (Suppl. 1):251–271.

Szymura, J. M., and N. H. Barton 1986. Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, Near Cracow in Southern Poland. Evolution **40**:1141–1159.

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

Underwood, J. N., L. D. Smith, M. J. H. Van Oppen, and J. P. Gilmour 2007. Multiple scales of genetic connectivity in a brooding coral on isolated reefs following catastrophic bleaching. Molecular Ecology **16**:771–784.

Väinölä, R., and M. M. Hvilsom 1991. Genetic divergence and a hybrid zone between Baltic and North Sea *Mytilus* populations (Mytilidae: Mollusca). Biological Journal of the Linnean Society **43**:127–148.

Visscher, P. M., W. G. Hill, and N. R. Wray 2008. Heritability in the genomics era – concepts and misconceptions. Nature Reviews Genetics **9**:255–266.

Waples, R. S. 1998. Separating the wheat from the chaff: patterns of genetic differentiation in high gene flow species. Journal of Heredity **89**:438–450.

Waples, R. S. 2015. Testing for Hardy-Weinberg proportions: have we lost the plot? Journal of Heredity **106**:1–19.

Waples, R. S., and O. Gaggiotti 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. Molecular Ecology **15**:1419–1439.

Waples, R. S., A. E. Punt, and J. M. Cope 2008. Integrating genetic data into management of marine resources: how can we do it better? Fish and Fisheries **9**:423–449.

Ward, R. D., M. Woodwark, and D. O. F. Skibinski 1994. A comparison of genetic diversity levels in marine, freshwater, and anadromous fishes. Journal of Fish Biology **44**:213–232.

Whitlock, M. C., and D. E. McCauley 1999. Indirect measures of gene flow and migration: FST≠1/(4Nm+1). Heredity **82**:117–125.

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

Wright, S. 1951. The genetical structure of populations. Annals of Eugenics **15**:323–354.

Zbawicka, M., T. Sanko, J. Strand, and R. Wenne 2014. New SNP markers reveal largely concordant clinal variation across the hybrid zone between *Mytilus* spp. in the Baltic Sea. Aquatic Biology **21**:25–36.