

Histoire évolutive et influence de la sélection sur la diversité génétique des annélides polychètes d'environnements extrêmes

Alexis Bioy

► To cite this version:

Alexis Bioy. Histoire évolutive et influence de la sélection sur la diversité génétique des annélides polychètes d'environnements extrêmes. Biologie moléculaire. Sorbonne Université, 2018. Français. NNT: 2018SORUS156 . tel-02467893

HAL Id: tel-02467893 https://theses.hal.science/tel-02467893

Submitted on 5 Feb 2020 $\,$

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.





Sorbonne Université

Ecole doctorale 227 – Science de la Nature et de l'Homme Station biologique de Roscoff – UMR 7144 Adaptation et diversité en milieu marin / ABICE (Adaptation et Biologie des Invertébrés en Conditions Extrêmes

Histoire évolutive et influence de la sélection sur la diversité génétique des annélides polychètes d'environnements extrêmes

Par Alexis BIOY

Thèse de doctorat de Génétique des population / Biologie Marine

Dirigée par Stéphane HOURDEZ & Didier JOLLIVET

Présentée et soutenue publiquement le 26 juin 2018

Devant un jury composé de :

Mme. ARNAUD-HAOND Sophie, chercheuse à l'IFREMER, Rapporteur

M. BACKELJAU Thierry, professeur à l'Université d'Antwerp (Belgique), Rapporteur

M. BONHOMME François, directeur de recherche DRCE CNRS, Examinateur

M. CHEVALDONNE Pierre, Directeur de recherche DR2 CNRS, Examinateur

Mme. SAMADI Sarah, professeur MNHN, Examinatrice



A ma mère qui m'a toujours laissé faire mes propres choix mais surtout les a respectés



Remerciements

Je voudrais tout d'abord remercier les membres de mon jury, Thierry BACKELJAU et Sophie ARNAUD-HAOND d'avoir accepté d'être rapporteurs de ma thèse, ainsi que François BONHOMME, Pierre CHEVALDONNE et Sarah SAMADI d'avoir accepté de participer. Merci à Christophe, Fred, Eric et encore François d'avoir accepté de suivre mes péripéties et de m'avoir prodigué de bons conseils au cours de mes comités de thèse.

Merci à l'ensemble de l'équipe ABICE de m'avoir accueilli pendant ces quatre années et quelques. Un merci particulier à Didier et Jean grâce auxquels j'ai pu rentrer par la petite porte du M2 dans le monde des sources hydrothermales, aussi passionnant qu'obscure (pas que littéralement...). Un autre merci tout aussi particulier à Stéphane et Didier d'avoir eu confiance en moi pour ce sujet de thèse qui a élargie mon horizon au froid polaire de l'Antarctique et au milieu tempéré Breton (pour lequel il suffisait d'ailleurs d'un demi-tour de chaise d'ordinateur pour profiter d'une fenêtre avec vue imprenable sur lui). Merci à Anne-So et Dominique pour leur aide et leurs conseils à mon arrivée dans ces labos inconnus qui deviendraient sous peu le lieu de mes longs rendez-vous en tête à tête avec l'HPLC (pour HyPer Long et Ch...ool). Merci à ceux qui sont arrivé après moi (et oui quatre ans de carrière quand même !), Stéphanie et Victor M. qui m'ont sauvé quand le temps a commencé à presser et que des banques RAD et des analyses informatiques me manquaient cruellement. Ah et comment parler d'ABICE sans parler du bureau 116 qui a vu et verra un sacré paquet d'étudiants défiler et grâce auquel j'ai pu rencontrer ma meilleure mais surtout unique stagiaire qui est devenu une super collègue de bureau, Camille TB. Avant ça, Kevin, Camille D, Gonzalo, Lola, Perrine, puis Bérénice et le dernier né Victor L. Un grand merci à toute l'équipe.

Mais parce que la station, c'est aussi tout un tas de monde, heureusement qu'il y avait la machine à café pour rencontrer tous ceux qui travaillent sur des sujets plus compliqués les uns que les autres. Je commencerais donc par Céline et « le monde secret de l'administration », qui en plus de m'aider toujours gentiment sans me montrer que mes questions sont particulièrement débiles, est une super collègue de pose cigarette/journée poster/barbecue de la station/.... Bon en vrai vous êtes beaucoup à la pause-café alors merci à tous ;)

Et le Ty Pierre de mercredi (et pas que), le rendez-vous incontournable pour se plaindre sur l'épaule chaleureuse d'une petite bière en compagnie d'autre malheureux de la pipette ! Heureusement que ça n'était pas que ça ! Tous ces bons moments passés avec les stagiaires de passage, les thésards et post-doc et puis quelques permanents quand même. Donc merci à

Tortillo, Solenn, Momo, Valérie, Jojo, Loïc, Claire, Basil, Théo, Justine, Laure, Hugo, Doriane, Tristan, Haaaa je sens que j'oublie plein de monde je suis désolééé d'avance. Ah mais merci au Ty Pierre bien sûr, particulièrement Goul, Marie et Dorian qui sont restés toujours de bonne humeur même quand en arrivant à 5 pour finir à 35, on leur bouleversait le plan de salle occasionnant des comptes incompréhensibles.

Tant qu'on en est dans ce qui se mange et se boit, le Gulf Stream, enfin surtout le Gulf pour les intimes. L'endroit où en trois repas tu prends 5kg, pas juste parce que t'es en Bretagne et que le beurre c'est la vie, mais parce que c'est HYPER bon et que tu oublis que ton estomac n'est pas le sac de Marie Poppins. Mais bon, ça fait pas tout, on est surtout content d'y aller pour les sourires, la bonne humeur générale et le super accueille qui nous y est toujours fait.

En pensant au Gulf on pense aussi au café de l'AJC, et donc à l'AJC et ses fêtes ! Je n'oublierais pas cette année de mandat présidentiel héhé ! Un super groupe avec qui je suis vraiment heureux d'avoir partagé cette année associative. Et puis merci quand même aux autres promos ! D'ailleurs Victor L, après une rapide reflexion, je pense que je suis ton arrière-grand-père en terme d'AJC, mais ton père si on voit l'héritage de bureau... ça ressemble vachement aux histoires d'Eloïse ça...

Justement Eloïse, et Laure aussi. Qui ont bien voulu d'un thésard sur la fin en garde alternée. Un grand merci !!! Mais vous n'étiez pas là que pour ça, merci à vous deux, mais aussi à Peche, Damien, Ulysse, Gus, Jérémy, reVictor L, reCamille TB, reVictor M, re un paquet de nom d'avant, Mathilde, Solène, ..., et Marine !!!

Marine, je te décerne le prix du cumul des mandats de super rencontre improbable le jour des oraux de l'ED, super coloc, super collègue (qui m'a apporté beaucoup pendant nos discussions « travail »), super amie (qui m'a apporté beaucoup pendant nos discussions « surtout pas travail »). Je suis vraiment heureux d'avoir fait ma thèse en même temps que toi !

Merci aussi à Margot avec qui l'on se suit depuis le master EBE, aussi originale que passionnée. Ta délicatesse face au stress des autres aura été d'une efficacité prodigieuse sur la diminution du stress à la fin de ma rédaction. Et un grand, gros, énorme merci pour cette aquarelle qui trônera en bonne place chez moi.

Un grand merci aussi quand même à SciHub, qui m'a permis d'économiser approximativement le pris de trois villas sur la côte d'azur en lectures d'articles, qui en plus n'étaient pas toujours utiles...

Bon et parce que ça fait quand même énormément du bien de déconnecter, et que le mieux pour ça c'est la famille et les amis. Alors merci à Benjamin et Ludivine qui ont eu le temps de se construire une jolie petite famille le temps que je sois à Roscoff (et oui toujours 4ans), merci à

Mickael, Baptiste. Merci aux BIOY d'Hasparren, de Limeil et de Créteil (c'est comme ça qu'on dit chez moi), et donc merci aussi aux Gautherot de Dijon.

Bon et puis le meilleur pour la fin ! Parce qu'on parle du calvaire des conjoints de femmes enceintes, mais essayez de donner une barquette de fraises à un thésard en pleine crise existentielle ... Et en plus le mien a dû supporter une petite prolongation de fin de parcours. Merci de m'avoir supporté et aidé et tout simplement d'avoir été là pour moi, Merci Thibaut.

Sommaire

I - Introduction	1
1 - Les mécanismes évolutifs façonnant le génome : sélection, démographie, traits	
d'histoire de vie.	5
A - Le mode de reproduction et les traits d'histoire de vie	6
B - L'histoire démographique des populations	7
C - La sélection naturelle	8
2 - Les environnements extrêmes	12
A - Les sources hydrothermales profondes	13
B - Le milieu polaire Antarctique	16
C - La zone intertidale tempérée	19
3 - Les Annélides polychètes	20
Aphroditiformia	22
4 - Les objectifs de la Thèse	24
II - Obtention et traitement des données génomiques	25
1 - Traitement des échantillons et construction des banques individuelles d'ADN	28
A - Préparation des ADN	28
B - Construction des banques ddRAD	29
2 - Les mesures de la diversité génétique et de différenciation	33
A - Mesures intra-population	33
B - Mesures inter populations	36
3 - Nettoyage des lectures et identification des populations	37
A - Filtrage, nettoyage et assignation (mapping) des lectures (reads) à un génome de	;
référence	37
B - Constructions des catalogues de locus	41
C - Structure génétique des populations	45
4 - Analyse de la diversité génétique des espèces	49
A - Estimation des diversités synonyme et non-synonyme : Reads2snp	49
B - Calcul d'indices de diversité et tests de neutralité : Genepop/DNAsp	51
5 - Identification des locus « outliers » (ou aberrants) sous sélection positive avec diffé	rents
modèles de populations : Lositan & DetSel	52
6 - Bilan	54
III - Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processu	ıs de
spéciation chez les espèces du genre Alvinella	55
1 - Introduction	55
2 - Matériel et Méthodes	65
3 - Résultats	66
A - Evaluation de la qualité des banques	66
B - Cartographie des lectures individuelles sur génome de référence de A. pompejan	a.69
C - Création des catalogues de locus : Stacks	72
D - Identification des structures de populations avec <i>a priori</i> sur l'appartenance des	
individus à des localités géographiques ou à un habitat thermique	74
E - Identification des locus 'outlier' en Fst entre populations et effet de la sélection	
diversifiante	87
F - Identification des gènes associés aux locus sous sélection positive dans l'analyse	de
la différenciation entre les populations de l'EPR nord et sud.	106
G - Diversités des polychètes Alvinella de part et d'autre de la barrière équatoriale	108
4 - Discussion	111

	111
A - Limites methodologiques	111
B - Role de la barriere equatoriale dans l'evolution des populations d'Alvinella :	110
speciation en cours ou remise en contact secondaire des populations nord et sud ?	113
C - Identification des génes et fonctions sous sélection diversifiante	120
5 - Conclusion	127
IV - Structure, connectivite et diversite des populations d' <i>Harmothoe fuligineum</i> en	100
Antarctique	129
I - Introduction	129
A - Température et adaptabilité des espèces	129
B - Histoire géologique et évolution de la température en Antarctique	131
C - Variabilité contemporaine de la température en Antarctique	133
D - Choix du modèle Polynoidae Harmothoe fuligineum	137
2 - Matériel et Méthodes	138
A - Echantillonnage	138
B - Assignation populationnelle	139
3 - Résultats	139
A - Contrôle qualité des banques	139
B - Construction des catalogues de locus	140
C - Définition des populations analysées	144
D - Assignation de l'individu de Kerguelen à une population continentale	148
E - Diversité génétique des trois échantillons de <i>H. fuligineum</i>	149
F - Etude de la différenciation génétique des populations locus par locus et rôle de	
l'étagement/éloignement géographique dans cette différenciation	150
4 - Discussion	159
A - Connectivité autour de l'Antarctique et avec les îles sub-antarctiques	159
B - Impact des glaciations passées sur l'histoire évolutive de l'espèce	164
C - Structure des populations en réponse à l'adaptation aux contraintes liées à la	
profondeur	167
5 - Conclusion	171
V - Effet de la variabilité environnementale sur la diversité génétique des invertébrés m	arins
et sur la thermotolérance de leurs enzymes	173
1 - Introduction	173
A - Les traits d'histoire de vie	173
B - Variabilité environnementale et sélection purifiante	174
C - Les phénomènes démographiques : extinction/recolonisation, goulots	
d'étranglement et expansion démographique	175
D - Les contacts secondaires entre populations génétiquement différenciées	175
E - Implications fonctionnelles	176
2 - Matériel et Méthodes	178
A - Echantillonnage	178
B - Estimation des diversités nucléotidiques globales, synonymes et non-synonyme	es.180
3 - Distribution des mutations rares ou en fréquences intermédiaires dans le polymor	ohisme
des espèces (spectres de fréquences alléliques)	181
A - Distribution des mutations polymorphes non-synonymes selon la nature des ac	ides
aminés	181
B - Mesures d'activité de la cMDH entre espèces et détermination de la températur	e de
dénaturation thermique (Tm) des différentes isoformes	182
C - Mesure d'activité et détermination du Tm	182
D - Test de Tukey	183
4 - Résultats	183

A - Contrôle qualité des banques	
B - Alignement des séquences sur les transcriptomes de référence	
C - Distribution des valeurs de diversité nucléotidique et des valeurs du test de	e Tajima
monolocus entre les différentes espèces	
D - Catégorisation du polymorphisme protéique et identification des remplace	ements
préférentiels selon les espèces	196
E - Etude de la variabilité du Tm de la cMDH entre les différentes espèces sel	on le type
d'environnement	
5 - Discussion	
A - Milieu hydrothermal, hyper variabilité thermique et température élevées	
B - L'Antarctique, une stabilité thermique dans l'extrême	
C - Le milieu côtier tempéré, un milieu hypervariable mais cyclique	
D - Relations entre traits d'histoire de vie et diversité synonyme	
6 - Conclusion	
VI – Conclusion, Discussion, Perspectives	
1 - Conclusion / Discussion	
A - Sélection et Evolution	
B - Sélection et Différenciation	
2 - Perspectives	
Bibliographie	

I - Introduction

Les populations naturelles sont définies par un ensemble de similitudes permettant d'identifier les membres lui appartenant ou non, mais ces similitudes n'impliquent pas à une identité stricte des individus la composant. Comme le dit Buffon, « *Le premier obstacle qui se présente dans l'Histoire Naturelle, vient de cette grande multitude d'objets ; mais la variabilité de ces mêmes objets, et la difficulté de rassembler les productions diverses des différents climats, forment un autre obstacle à l'avancement de nos connaissances* » (Buffon, Histoire naturelle 1749).

Dans L'Origine des espèces publié le 24 novembre 1859, Charles Darwin présente un ensemble d'observations réalisées sur des espèces domestiques et sauvages et en conclut des règles ou des vérités générales sur les mécanismes évolutifs régissant le monde du vivant. Il commence par présenter des cas de sélection humaine sur des espèces animales et végétales domestiquées dont l'intérêt majeur est qu'elle est systématiquement finaliste puisqu'orientée par l'homme. Puis en s'appuyant sur ces premières observations, il tente de décoder les mécanismes en cours lorsque cette sélection se fait de façon naturelle. Il explique ainsi pour le cas des pinsons de Galápagos que la diversité de formes et de tailles de becs que l'on trouve en comparant les différentes îles est une réponse adaptative aux différentes ressources disponibles sur chacune de ces îles. De plus il hiérarchise ses idées en allant de celles qui seront le plus facilement tolérables par la communauté scientifique à celles qui seront les plus sujettes à controverse (filiation de l'homme avec les primates). Darwin part d'une première observation : les caractères morphologiques d'un individu varient entre variétés d'une plante cultivée ou d'un animal domestiqué, ainsi qu'au sein d'une même espèce à l'état naturel. Le deuxième constat est que les caractéristiques individuelles sont héritées par les descendants. Dans le cas de la domestication, l'homme a sélectionné certaines caractéristiques propres à ses besoins (alimentation, caractéristiques physiques) et les a accumulées sur plusieurs générations pour former de nouvelles variétés pouvant être très différentes les unes des autres. Dans le cas de la sélection naturelle, il part du principe qu'une population sans limite croit de façon géométrique dans un milieu nécessairement limité, ce qui implique de la compétition entre les individus pour l'espace et la nourriture. Les individus présentant les caractéristiques les plus adaptées à leur environnement survivent mieux et laissent plus de descendants. Ces caractéristiques sont donc héritées et accumulées dans les générations suivantes jusqu'à obtenir de nouvelles espèces. Il aborde également l'effet évolutif d'une compétition pour la reproduction entre individus du même sexe et, par conséquent, sur la sélection sexuelle qui s'opère chez les mâles en général et qui complète la sélection naturelle. Il apparaît alors que, quel que soit le type de sélection en jeu, celui-ci ne peut s'opérer que sur des différences entre individus, et donc sur le polymorphisme inhérent d'une espèce. Il propose donc que les espèces apparaissent via un tri des caractères présents dans les populations en réponse à des conditions de sélection (humaine ou naturelle) différentes.

Très rapidement cette variabilité existante au sein des espèces et les règles régissant sa transmission ont intrigué. Les mécanismes génétiques à l'origine du maintien de ces polymorphismes vont être alors établis par Gregor Mendel au XIX siècle sur le petit pois (Bateson & Mendel, 2013). Mendel réalise une étude sur l'hérédité dans laquelle il s'intéresse à différentes caractéristiques morphologiques de ses plants à travers la graine, sa forme ou la couleur des cotylédons, la couleur des fleurs, la forme des cosses ou encore la taille et la structure générale du plant. Pour des raisons d'avancement technique, ces caractères ont constitué les premières preuves et bases de questionnement sur la variabilité intra-spécifique et l'héritabilité des caractères. Ces études se sont aussi bien concentrées sur du polymorphisme observé dans le temps, entre sexes ou entre environnements (Bessey, 1880 ; Harnly, 1936 ; Lacassagne, 1940 ; Lutz, 1913).

C'est alors que deux problématiques vont se mêler : comment accéder au polymorphisme de populations en apparence homogène d'une part, et, comment identifier l'effet de ces variations sur les individus s'il y en a un d'autre part.

Dans un premier temps, l'étude de la variation génétique pouvait se faire de deux façons (Lewontin, 1991) : i - en étudiant les modifications phénotypiques induites lorsqu'il y en avait afin de déterminer les relations de dominance opérant en utilisant le même type de protocole que celui utilisé par Mendel. Mais les caractères identifiés étaient difficilement corrélables avec une quelconque différence de valeur sélective ii – en caractérisant le polymorphisme d'arrangement chromosomique que l'on observait dans les populations. Ce polymorphisme d'arrangement chromosomique a pu quant à lui être corrélé avec des différences de valeurs sélectives par suivi des fréquences chromosomiques dans des populations expérimentales de drosophiles à différentes températures (Dobzhansky, 1937). Cette seconde méthode présentait cependant l'inconvénient de ne pas permettre d'interprétation à l'échelle du gène. En effet il était impossible de savoir si un ou plusieurs gènes étaient impliqués et les relations de dominance n'étaient donc pas non plus extrapolables (Lewontin, 1991).

Par la suite, avec l'invention de la technique de l'électrophorèse des protéines (Picton & Linder, 1897) et son utilisation pour caractériser le polymorphisme de charge des allozymes et ainsi accéder directement à une partie du polymorphisme existant entre les protéines résultant de différences dans leurs séquences d'acides aminés (Lewontin & Hubby, 1966 ; Harris 1966). Cette technique permet alors de mettre en évidence l'existence d'un polymorphisme affectant apparemment la majorité des gènes sans effet apparent (Harris 1966), probablement maintenu même entre populations isolées (Lewontin & Hubby, 1966) et ne semblant pas être corrélé à des paramètres environnementaux (Selander *et al.*, 1970).

En 1964, Kimura a proposé différents scénarios permettant d'expliquer le maintien d'un polymorphisme dans les populations. Il part du postulat que pour une séquence nucléotidique donnée, en raison du fait que chaque base puisse subir une substitution, que des réarrangements sont possibles et qu'on peut assister à des insertions ou des délétions le nombre de séquences alternatives possible est astronomique. Lorsqu'une nouvelle séquence apparait dans une population, il est possible qu'en raison d'un effet délétère celle-ci soit immédiatement purgée. Mais le simple fait de la dérive en raison de sa présence en faible fréquence peut déjà impliquer une perte de celle-ci quand bien même elle aurait un effet positif sur la valeur sélective de l'individu. Il montre ainsi que dans une population dans laquelle les hétérozygotes sont positivement sélectionnés, la population va voir augmenter le nombre d'allèles pour ce gène. Mais plus le nombre d'allèles sera élevé, plus l'effet de la dérive génétique sera marqué. Ce nombre d'allèles sera grandement dépendant de la taille efficace (nombre d'individus participant effectivement à la reproduction) et du taux de mutation du gène en question. Mais il montre aussi que dans le cas neutre où la sélection n'opère pas, le nombre d'allèles qu'une population peut maintenir n'est alors plus dépendant que de la taille efficace de la population (N_e) et du taux de mutation (v) et est égal à $4.N_e.v+1$.

Par la suite cette théorie de la neutralité sera en partie révisée, en donnant encore moins de poids à la sélection. En effet d'un point de vue mathématique, la probabilité de fixation d'une mutation dans une population (u) sera directement égale à la fréquence de cette mutation (p) dans le cas du modèle neutre, et elle sera déterminée par la formule suivante dans le cas d'une mutation soumise à un coefficient de sélection (s) (Ohta, 2002) : $u = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}}$, avec N la taille efficace de la population. Ceci implique que dans le cas de populations qui présentent de petites tailles efficaces ou dans celui où le coefficient de sélection est bas, la probabilité de fixation d'une mutation même sous sélection est très proche de celle d'une mutation neutre.

D'un autre côté, la comparaison des séquences d'acides aminés de l'hémoglobine chez plusieurs vertébrés a permis de mettre en évidence d'une part des similitudes de polymorphisme identifié à certaines positions entre différentes espèces et d'autre part une certaine régularité dans l'apparition de ces changements au cours de l'évolution (Zukerkandl & Pauling 1965). Les auteurs ne pouvant pas vérifier l'effet de toutes ces mutations, ils évitent soigneusement de se risquer à parler de convergence évolutive. Mais ils suggèrent que ces changements bien qu'ils soient neutralement contraints par le code génétique qui ne permet pas en une seule mutation de passer d'un acide aminé à n'importe quel autre, ils peuvent être en partie sélectionnés sur la base des propriétés de ceux-ci pour assurer le fonctionnement de la protéine.

Malgré cela de nombreuses études vont réussir à identifier des augmentations ou diminutions de valeur sélective en lien avec la présence d'un variant potentiellement « adaptatif » avec, par exemple, une forme allélique majoritaire en présence d'un polluant (Nevo et al., 1984), l'existence d'un allélisme et une corrélation entre la présence de l'un d'eux et une sensibilité au cuivre cumulée à une tolérance au zinc expliquant possiblement un maintien d'une forme alternative (Lavie & Nevo, 1982) ou encore des cas plus complexes avec une corrélation entre présence d'un allèle, succès reproducteur et faible taux de survie, permettant le maintien d'un polymorphisme variable dans les populations (Mitton & Koehn, 1975).

Par la suite de nouvelles techniques permettant d'accéder au polymorphisme génétique sont apparues dont entre autres les techniques de RFLP (« Restriction Fragment Length Polymorphism » ou polymorphisme de taille des fragments après digestion de l'ADN génomique par des enzymes de restriction) (Saiki *et al.*, 1985), de dénaturation différentielle des brins d'ADN ou le séquençage de l'ADN. Mais dans tous les cas une grande majorité des polymorphismes identifiés sont effectivement maintenus de façon neutre dans les populations d'où l'intérêt d'être vigilant quant aux interprétations à faire au vu d'une situation donnée.

Mais les avancées de la génétique permettent aujourd'hui d'accéder au polymorphisme à l'échelle du génome entier et de distinguer ce qui tient de la sélection de ce qui se produit de façon neutre. Ainsi des comparaisons de taux de recombinaison le long du génome permettent de mettre en évidence des zones particulièrement maintenues. Celles-ci trahissent un balayage sélectif ayant induit une augmentation en fréquence dans la population d'un haplotype de

grande taille incluant potentiellement d'autres gènes, l'identification pour certaines parties du génome d'une corrélation négative entre polymorphisme populationnel et divergence nonsynonyme entre espèces permet l'identification d'une zone soumise à la sélection purifiante et différenciant des espèces de mouches (Shapiro et al., 2007). La comparaison du polymorphisme des espèces de grands singes permet de mettre en évidence une diminution significative de la diversité nucléotidique autour de certains gènes traduisant là encore du balayage sélectif et révèle une importante perte fonctionnelle fixée de gènes dont la comparaison inter-espèces permet d'identifier les relations de parenté entre elles et permet ainsi d'en valider le rôle adaptatif (Prado-Martinez et al., 2013). Une étude comparative entre l'homme et le chimpanzé a notamment permis, en comparant leurs génomes ne différant que de 1,24%, de montrer qu'il y avait une corrélation positive entre les taux de réarrangement chromosomique et de substitution à l'échelle des chromosomes (Navarro & Barton, 2003). Ceci montre d'une part que la divergence entre espèces est dépendante de la structure du génome, et d'autre part que celle-ci ne se distribue pas de façon homogène le long de celui-ci. Bien que ce genre d'analyses présente un intérêt évident dans l'étude de la structuration des populations, voire des espèces, il est nécessaire de disposer de génomes de référence complets ; ce qui est rarement le cas lorsque l'on étudie des métazoaires non-modèles dont le génome est de grande taille, à intérêt économique quasi nul.

Mais il reste tout de même possible, en réalisant ces comparaisons entre espèces, entre individus d'une population ou entre parties d'un même génome de faire la part des choses entre phénomènes démographiques, sélection, arrangement chromosomique, etc...

1 - Les mécanismes évolutifs façonnant le génome : sélection, démographie, traits d'histoire de vie.

Le polymorphisme d'une espèce peut donc s'observer à différentes échelles allant du phénotype (diversité de formes, comportements, propriétés) à la séquence nucléotidique d'un gène ou d'une région non-codante (accumulation de mutations au cours du temps). Elle varie d'une espèce à une autre et entre des régions différentes d'un génome (Ellegren & Galtier, 2016). Les causes de cette variabilité sont nombreuses et prennent en compte la position d'un gène le long du chromosome, l'hétérogénéité du taux de mutation entre gènes, la taille efficace de la

population, les forces de sélection, les traits d'histoire de vie de l'espèce et son mode de reproduction ainsi que l'histoire démographique de celle-ci (Singhal et al., 2017).

A - Le mode de reproduction et les traits d'histoire de vie

Chez les espèces qui pratiquent la reproduction sexuée, les réarrangements chromosomiques lors de la méiose pour la production de gamètes haploïdes (et donc la recombinaison intragénique) sont un des premiers mécanismes influençant la diversité génétique d'une espèce. On peut placer le mode de reproduction d'une espèce sur un axe imaginaire allant de l'appariement purement aléatoire des gamètes au sein d'une population (panmixie) au cas extrême où les gamètes ne s'apparient qu'au sein d'un même individu (autofécondation). L'autofécondation (aucun mélange de patrimoine génétique entre individus) conduit à une évolution particulière de la diversité génétique vers le monomorphisme. En augmentant l'homozygotie, elle va diminuer les possibilités de recombinaison entre allèles. Elle va aussi limiter l'action de la dérive génétique qui est corrélée à la diversité génétique en restreignant peu à peu le nombre d'allèles disponibles d'une génération à l'autre. Ce mode de reproduction va aussi intensifier la purge des allèles délétères puisque l'homozygotie permet de mettre en évidence les allèles récessifs et les éliminer par sélection purifiante (Crow, 1970 ; Hartfield et al. 2017). Ces effets peuvent être aussi mis en relation avec des modes de multiplication sans reproduction comme la parthénogénèse ou la clonalité (Arnaud-Haond et al., 2005). En ce qui concerne la reproduction sexuée entre individus et la proximité géographique entre ceux-ci, deux facteurs sont à prendre en compte quand on s'intéresse à l'évolution de la diversité génétique ; la dispersion des gamètes ainsi que celle des produits de la reproduction (géniteurs, œufs, larves, etc...) influence la diversité génétique des populations (Ellingson & Krug, 2016 ; Gohli et al., 2016 ; Dalongeville et al., 2016). Ainsi plus les individus d'une espèce se reproduisent avec leurs plus proches voisins, plus on observe une différenciation génétique entre les populations et une diminution de la diversité au sein de chaque population en raison de la diminution de la taille efficace accompagnée d'une dérive génétique plus marquée. Les traits d'histoire de vie sont aussi fortement corrélables avec la diversité que l'on observe dans une population. On observe en effet une forte corrélation négative entre la longévité, la taille adulte, la faible fécondité ou les soins parentaux avec la diversité génétique (Romiguier et al., 2014).

B - L'histoire démographique des populations

D'un point de vue démographique, une population est inféodée à un lieu et une niche donné (Jokinen et al., 2016; Callaghan et al., 2016) qui elle-même est définie par des caractéristiques environnementales et des relations biotiques particulières (compétition, mutualisme, etc). Et sans prendre en compte une quelconque connectivité avec d'autres populations, de même que cet habitat est défini dans l'espace, cette population est limitée dans l'espace et donc aussi en nombre d'individus. Or la taille d'une population et plus précisément sa taille efficace (nombre d'individus reproducteurs qui transmettent leurs gènes à la génération suivante) est le premier facteur à influencer la diversité génétique d'une population (Hague & Routman, 2016 ; Ellegren & Galtier, 2016). Cette taille efficace est susceptible de varier beaucoup au cours du temps, notamment lorsque les conditions locales deviennent défavorables, en réduisant alors la taille, ou au contraire permettent son expansion. Ces variations passées ont alors une influence sur la diversité génétique actuelle d'une espèce. Ainsi lorsqu'une grande partie de la population disparaît (phénomène appelé goulot d'étranglement ou « bottleneck »), on assiste à une perte d'une partie de la diversité génétique que cette population avait accumulée au cours de son histoire. La suppression aléatoire de nombreuses lignées alléliques aura alors tendance à renforcer la proportion des lignées les plus divergentes. L'effet de la dérive génétique sera d'autant plus prononcé que la nombre d'individus survivants sera réduit, pouvant aller jusqu'au maintien d'une seule lignée allélique et donc à un monomorphisme de nombreux gènes lorsque l'effet de goulot d'étranglement est très prononcé. De même dans le cas d'une colonisation d'un nouveau territoire/habitat à partir de quelques individus de la population mère, la population nouvellement formée fonde son patrimoine génétique sur les individus ayant participé à l'évènement de colonisation, on parle alors d'effet fondateur. Dans un cas comme dans l'autre, la diminution de la diversité sera d'autant plus forte que le nombre de colonisateurs est petit, c'est-à-dire d'autant plus forte que le nombre d'individus ayant survécu à l'évènement de mortalité et s'étant reproduit est petit (Jangjoo et al., 2016 ; Olsen et al., 2016 ; Hillman et al., 2017).

Lors de la colonisation d'un nouvel environnement, avec ou sans contact avec la population mère, on assiste ensuite à ce que l'on appelle une expansion démographique si les conditions sont favorables. Au fil du temps, l'aire de répartition de la population va s'étendre et la population va accumuler des mutations au niveau de la zone d'expansion. Dans un modèle en nombre infini de mutations, les individus porteurs de nouvelles mutations sont à une fréquence de 1/2N, ce qui génère un excès de mutations rares au niveau de la population, le temps de

fixation d'une nouvelle mutation par dérive étant de 4N générations. Ces nouvelles mutations appelées singletons ne sont retrouvées qu'une seule fois dans la population et s'accumulent dans la population jusqu'à ce qu'un équilibre s'établisse entre la mutation et la dérive. Ce genre de phénomène entraîne une ponctuelle mais extrême diversification des séquences haplotypiques en relation directe avec le taux de mutation s'exerçant sur le gène et la taille de la population qui augmente de façon exponentielle avec le temps.

C - La sélection naturelle

La théorie de la neutralité de Kimura (1983) prédit que la diversité d'une population est fonction de sa taille efficace et de son taux de mutation, or l'observation de cas concrets permet de mettre en évidence des situations où la diversité génétique et sa distribution ne suit pas l'attendu neutre. Des études se sont notamment intéressées au décalage que l'on observe souvent en milieu marin lorsque l'on estime la taille globale d'une population et sa taille efficace en utilisant des données de polymorphisme. La première explication est que la taille efficace d'une espèce (i.e. Ne, nombre d'individus participant effectivement à la production de la génération suivante) est souvent très nettement inférieure à la taille réelle de la population en raison de la grande variabilité du succès reproducteur des individus au sein de la population que l'on peut assimiler à une véritable loterie de la reproduction ou "sweepstake reproductive success" (Hedgecock, 1994). La reproduction peut aussi se faire préférentiellement dans une certaine gamme d'âges ou de stades de développement, tout comme la reproduction entre sexes différents avec des sexe ratio biaisés, limite encore les possibilités de reproduction pour certaines catégories (Charlesworth, 2009). Une autre possibilité est l'occurrence récente de «bottlenecks» réduisant la diversité haplotypique de la population ainsi que des gamètes et mimant ainsi une faible taille efficace (Frankham, 1995; Charlesworth, 2009). La deuxième explication la plus répandue est que la diversité neutre est systématiquement réduite par la sélection positive à certains gènes par un effet d'entraînement sur l'environnement génomique qui leur est lié (Maynard Smith & Haigh, 1974). Lorsqu'à un locus donné, un allèle avantageux se fixe dans la population, il entraîne avec lui une baisse de la diversité neutre aux sites qui lui sont génétiquement liés, on parle alors d'effet auto stop ou de « hitchhiking » (Maynard Smith & Haigh, 1974). On assiste au même genre de phénomène lorsque des allèles sont contre-sélectionnés et purgés dans une population, on parle dans ce cas de sélection d'arrière-plan ou « background selection » (Charlesworth et al., 1995). Ainsi la sélection opère un rôle de filtre sur le génome, en sculptant la diversité le long de celui-ci, entraînant par la même occasion des purges ou des augmentations en fréquence de certaines portions du génome simplement en raison de leur proximité avec une séquence sous sélection jouant ainsi sur cette diversité n'impactant pourtant pas la valeur sélective « fitness » des individus.

En ce qui concerne la sélection directe sur des individus porteurs d'une mutation en particulier, l'effet sur la diversité du ou des gène(s) qui lui sont associé(s) peut être très différente selon la nature de cette sélection. On peut en effet distinguer trois grands types de sélection sur le phénotype : la sélection directionnelle, la sélection balancée et la sélection diversifiante. Ce que l'on appelle la sélection correspond en fait à la capacité des individus à survivre et à se reproduire pour perpétuer leur patrimoine génétique. Dans le cas de la sélection naturelle, les individus ayant le patrimoine génétique leur conférant une meilleure « fitness » vont se reproduire plus que les autres. Dans le cas d'une sélection artificielle ou domestication, c'est l'homme qui décidera de favoriser la reproduction des individus ayant un phénotype particulier pouvant avoir un intérêt alimentaire ou esthétique. Ainsi, quelle que soit sa nature, la sélection opère en conférant une valeur sélective différente aux individus en fonction de leur génotype.

-La sélection directionnelle

Dans le cas d'un locus bi-allélique pour lequel les homozygotes ont une « fitness » différente (1 et 1-s ; s correspondant au coefficient de sélection), en théorie et selon la relation de dominance entre allèles, la valeur sélective des individus hétérozygotes peut prendre différentes valeurs 1, 1-s/2 ou 1-s selon que l'allèle alternatif soit récessif, additif ou dominant.

Dans le cas d'un effet additif avec s = 0,01 (Li & Graur, 1991) on peut voir dans la figure I-1 que l'allèle conférant la meilleure fitness va systématiquement augmenter en fréquence dans la population au détriment de l'autre allèle. Lorsqu'une nouvelle mutation favorable apparaît, elle augmente rapidement en fréquence dans la population et on parle alors de balayage sélectif et/ou de sélection directionnelle. Ce balayage dans la population entraîne avec lui une partie du génome environnant et induit ainsi une baisse locale de la diversité proportionnée à la distance au site sous sélection (Prado-Martinez *et al.*, 2013).



Figure I-1 : Evolution de la fréquence d'un allèle codominant avantageux pour un coefficient de sélection s=0,01 au cours des générations pour une apparition de cet allèle à la génération 0 (Li & Graur, 1991).

-La sélection balancée

Deux types de sélection balancée ont été décrits : la sélection balancée avec avantage à l'hétérozygote et la sélection balancée fréquence-dépendante.

La sélection balancée avec avantage à l'hétérozygote est un cas qui a déjà été documenté pour de nombreuses situations de résistance à des pathogènes (Gemmell & Slate, 2006). En effet dans ces cas, le fait de présenter une hétérozygotie augmente les types de réponses face à l'arrivée d'un pathogène et augmente ainsi les chances de survie. En théorie, bien que les génotypes homozygotes aient une « fitness » moindre que les hétérozygotes, à l'exception des cas de létalité provoquée par l'homozygotie, on retrouve tout de même des homozygotes en fréquence rare et constante dans les populations simplement par appariement aléatoire des chromosomes lors de la reproduction.

Un autre cas de sélection balancée est la sélection fréquence dépendante négative ou l'avantage est donné aux allèles les plus rares. Dans une population polymorphe pour un trait donné, le coefficient de sélection est variable et inversement corrélé à la fréquence de chaque allèle. Ainsi un allèle est sélectionné sur la base de sa rareté dans une logique clef/serrure. Ce mécanisme est particulièrement marqué dans les populations naturelles de poissons guppy (Olendorf *et al.*,

2006). On y observe une grande diversité dans la coloration des mâles et l'intensité de celle-ci ; et cette variabilité est maintenue dans l'espèce par sélection balancée. En effet, quel que soit le trait considéré, le phénotype le plus rare est systématiquement celui qui a le meilleur taux de reproduction impliquant ainsi un maintien de polymorphisme dynamique dans la population. Bien que ce type de sélection semble être rare (Hedrick, 2012), le maintien de plusieurs formes alléliques va dans ce cas induire une augmentation locale de la diversité, puisque chaque lignée allélique va accumuler du polymorphisme propre.

-La sélection diversifiante

Dans le cas d'une sélection diversifiante par l'habitat qui porte aussi le nom de sélection disruptive ou divergente, on observe, contrairement à la sélection balancée, un avantage à l'homozygote dans une niche particulière, on parle aussi de sous-dominance. Dans ce cas particulier, deux lignées alléliques ou plus sont sélectionnées dans une même population en raison d'une hétérogénéité des conditions (Marchinko et al., 2014) ou en raison d'une perturbation induisant une scission de la population cis à vis d'un caractère (Landi et al., 2015). Mais ce genre de situation n'est pas stable dans le temps dans un environnement stable du fait que la probabilité pour que le coefficient de sélection des différents allèles soit exactement identique est quasi nulle (Hedrick, 1999).

Ce cas de figure peut effectivement se retrouver dans des cas où l'environnement présente une hétérogénéité spatial ou temporelle (Caldwell, 1982; Brakefield, 1985; Hedrick 1986) comparables à une subdivision de l'habitat en plusieurs niches écologiques dans le temps où l'espace proposée pour la première fois par Levene en 1953.

Tout comme dans le cas précédent, ce type de sélection induisant le maintien de plusieurs lignées alléliques, il va par le biais de l'évolution différente de chacune permettre une accumulation plus importante de diversité génétique.

Cette liste de facteurs influençant la diversité n'est pas exhaustive, d'autre mécanismes sont en effet connus pour impacter le polymorphisme comme par exemple les recombinaisons ou les réarrangements chromosomiques qui entrainent tous les deux une réorganisation du patrimoine entre ou au sein même d'un chromosome ou les éléments transposables dont l'un des effets majeurs est de générer la diversité au sein des espèces (Morgante *et al.*, 2005).

2 - Les environnements extrêmes

En termes probabilistes, un milieu est défini comme extrême lorsque l'amplitude des variables environnementales qui le caractérisent correspond à la queue de la distribution normale de ces mêmes valeurs sur l'ensemble des habitats du globe. Cette appréciation peut s'étendre au niveau de variation que peut prendre cette valeur environnementale au cours du temps, en effet un organisme peut tolérer des variations temporelles d'un paramètre du moment que l'ampleur ou la fréquence celles-ci n'excèdent pas un certain point (limite physiologique de l'organisme). D'un point de vue biologique, un environnement peut être considéré comme extrême pour une espèce donnée si un stress peut être identifié à partir de données montrant une baisse significative de valeur sélective ou « fitness » (Chevin & Hoffmann, 2017). De même que le stress subit par une espèce dépend de l'histoire adaptative de celle-ci, la distribution des valeurs environnementales à considérer doit être faite en ne prenant en compte que les valeurs auxquelles les individus de l'espèce considérée sont exposés. Cette façon d'envisager la tolérance d'une espèce à différentes valeurs et variations d'une variable donnée peut être représentée comme dans la figure I-2. Ainsi pour une variable environnementale X, l'individu considéré aura une valeur sélective d'autant plus élevée qu'il se trouvera dans le préférendum de l'espèce en bleu correspondant à une gamme donnée avec une variabilité dont l'ampleur ou la fréquence n'excèdent pas un maximum lui étant propre.



Figure I-2 : Illustration de la valeur de fitness en fonction de celle d'une variable quelconque X et de sa variabilité ΔX .

Au vu de cette définition de ce qu'est un milieu extrême, on pourra considérer que l'Antarctique par exemple peut être considéré comme stressant du point de vue des valeurs environnementales qui le caractérisent sans pour autant que ce soit le cas du point de vue de la variabilité de ces mêmes valeurs. Nous verrons aussi que des environnements tels que le milieu hydrothermal peuvent être considérés comme extrêmes à la fois en raison de l'hyper-variabilité thermique qui les caractérise mais aussi en raison des hautes températures qu'expériencent certains organismes s'y développant.

A - Les sources hydrothermales profondes

La première observation d'un champ hydrothermal actif a été faite sur la dorsale des Galápagos lors d'une plongée du submersible américain Alvin en 1976 (Lonsdale, 1977). Alors que cette plongée avait pour but de réaliser des observations directes d'une supposée circulation hydrothermale à l'axe des dorsales prédite par la théorie des modèles thermodynamiques, les premières observations ont permis, outre de vérifier l'existence de cette circulation hydrothermale profonde, de découvrir un peuplement faunistique original à forte biomasse associé dont l'existence n'avait pas été imaginée. En effet, la matière organique produite par photosynthèse près de la surface des océans est rapidement dégradée lors de sa sédimentation dans la colonne d'eau. Il en résulte qu'une très faible quantité atteint le fond des océans et ne permet pas le maintien de communautés à forte biomasse. Ces communautés sont en fait basées sur une production primaire locale bactérienne, alimentée par l'oxydation de composés réduits (chimiosynthèse) émis du plancher océanique dans le fluide hydrothermal (Fisher & Childress, 1992). Ce fluide est produit par l'interaction de l'eau de mer avec les roches profondes dans des zones géologiquement actives (dorsales océaniques, bassins arrière-arc) et est enrichi en H₂S, CO₂, CH₄ et divers éléments métalliques.

L'eau de mer pénètre dans le système hydrothermal par des crevasses et des fissures creusées sur les flancs de la dorsale. Au contact de la roche chaude (~150°C à 200°C), l'eau de mer perd sa charge en calcium (Ca²⁺), magnésium (Mg²⁺) et sulfates (SO4²⁻) par précipitation de sulfate de calcium et de minéraux riches en magnésium (extrait du site web de la Woods Hole Oceanography Institution, <u>http://www.divediscover.whoi-edu/vents</u>). En montant en température et en pression, la circulation de l'eau va induire un lessivage des basaltes et ainsi enrichir l'eau en H⁺, en éléments réduits (fer ferreux, sulfures, méthane) et en divers métaux (e.g. manganèse, silicium, cadmium, cuivre, zinc...). Le fluide chaud est moins dense et remonte, avant de rentrer en contact avec l'eau de mer froide. Les éléments minéraux dont il s'est chargé précipitent et forment des édifices au niveau des points de sortie (Lalou 1991). Ces édifices sont de formes variées et dépendent de l'intensité du mélange du fluide et de l'eau de mer en sub-surface et de la composition des roches sous-jacentes. Pour les émissions les plus chaudes (350°C), anoxiques et acides (pH de 2-3), la précipitation en sulfures polymétaliques et en sulfate de calcium conduit à la formation de fumeurs noirs (Alt 1995 ; Hannington et al.

1995). En revanche lorsque la précipitation s'initie partiellement au sein même des réseaux de crevasses, les édifices formés à la sortie du fluide hydrothermal sont des fumeurs blancs puisque que les émissions se sont apauvries sulfures polymétalliques et sulfates de calcium au cours de leur remontée, et lors de la sortie de la croûte océanique, ce sont les silicates, les anhydrites et le sulfate de baryum qui vont participer à l'édification de la cheminée (Hannington et al., 1995). Il existe d'autres formations plus diffuses, lorsque l'eau de mer pénètre moins profondément dans la croûte, ou lorsque de l'eau de mer se mélange en sub-surface avec les fluides hydrothermaux dans les cellules qui sont issues de la vidange des lacs de lave qui s'épandent dans le *graben* axial des dorsales, on obtient alors un fluide intermédiaire moins concentré que les fluides hydrothermaux qui peut diffuser lentement et de façon étendue à travers les surfaces poreuses des structures minérales ou directement par des crevasses ou fissures des laves tectonisées à des températures inférieures à 50°C (Alt 1995 ; Hannington et al. 1995) (Figure I-3).



Figure I-3 : Ensemble des formations associées à la circulation hydrothermale (extrait du site web de la U.S. Geological Survey, <u>https://walrus.wr.usgs.gov/global-ocean-mineral-resources/news.html</u> publication de Hein & Mizell, 2013). (cf texte pour plus de détails).

Lors de son émission dans l'eau de mer du fond, froide et bien oxygénée, le fluide hydrothermal se mélange de façon chaotique avec celle-ci, ce qui induit la formation d'un gradient thermochimique avec des échelles variables allant de quelques centimètres à quelques mètres le long duquel vont se distribuer les organismes en fonction de leur tolérance physiologique et de leurs besoins métaboliques (Shank et al., 1998 ; Luther et al., 2001 ; Bates et al., 2005).

La distribution et la durée de vie des sites hydrothermaux sont liées à la vitesse d'expansion du plancher océanique, ou taux d'accrétion des dorsales médio-océaniques. La dynamique spatiale et temporelle de ce type d'environnement est régie par le volcanisme et la tectonique de la dorsale (Fouquet et al., 1994). En l'absence d'un nouvel épisode de volcanisme effusif, cette activité diminue progressivement par colmatage des conduits en sub-surface jusqu'à l'arrêt de toute circulation hydrothermale entraînant par la même une disparition de la faune associée (Shank et al., 1998). Cette instabilité se manifeste aussi par l'apparition de nouveaux sites plus ou moins proches lors de changements de la circulation profonde des fluides hydrothermaux. Des modèles ont d'ailleurs été proposés après la mise en évidence du déplacement de l'activité hydrothermale le long de la dorsale (Watremez & Kervevan, 1990) expliquant que les phénomènes d'extinction et d'apparition ont lieu au sein de champs hydrothermaux qui se maintiennent et se déplacent dans le temps (se rapprochant et s'éloignant) et pourrait ainsi expliquer pour certaines espèces aux capacité de dispersion limitées que l'on observe une homogénéité entre sites éloignés (Chevaldonne *et al.*, 1997 ; Jollivet *et al.*, 1999).

Le long des dorsales rapides telle que la dorsale Est Pacifique (East Pacific Rise – EPR), les émissions de fluides de haute température sont confinées sur une zone étroite au-dessus de la chambre magmatique et l'hydrothermalisme est souvent associé à des activités volcaniques récentes (Alt, 1995 ; Fornari & Embley, 1995). Les émissions sont principalement localisées à l'intersection des failles perpendiculaires et parallèles à l'axe de la dorsale et forment champs aux structures complexes et stables dans le temps (Alt, 1995 ; Fornari & Embley 1995). Les dorsales à taux d'expansion intermédiaire telles que celles de Juan de Fuca ou les Galápagos présentent un large spectre de types d'édifices hydrothermaux, allant d'un hydrothermalisme de type volcanique, à faible durée de vie, jusqu'à des structures plus importantes en taille et complexes à longue durée de vie liée à un hydrothermalisme de type tectonique (Fornari & Embley, 1995). Les sites hydrothermaux se distribuent donc de façon agrégative le long des dorsales et ont une durée de vie limitée. De manière générale, la durée de vie des sites hydrothermaux n'excède pas 100 ans, et l'activité hydrothermale se déplace progressivement le long de la dorsale (Lalou et al., 1993a). La stochasticité de la dynamique de ce type d'environnement conduit à la création et la disparition d'habitats en relation avec les évènements géologiques. A cette fluctuation de l'activité hydrothermale à l'échelle de la centaine d'années s'ajoute des fluctuations fortes des caractéristiques physico-chimiques de l'habitat allant de l'échelle de quelques secondes à plusieurs heures (Johnson et al., 1988, Chevaldonné et al., 1991, Le Bris et al., 2003).

De façon intuitive, une telle variabilité des conditions de milieu et leurs amplitudes semble extrême, de plus cette stochasticité des variations thermiques et ce à des intervalles de temps très courts impliquent que les organismes au plus près des émissions ne bénéficient en aucun cas de conditions environnementales normées. Ainsi cet environnement constitue un modèle d'étude hors du commun dans lequel l'adaptation à la température ou à tout autre paramètre lié aux émissions hydrothermales est beaucoup moins intuitive que dans d'autres milieux. Il a déjà été montré dans cet environnement l'existence de polymorphisme adaptatif à la variabilité thermique par exemple chez *Alvinella pompejana* avec la phosphoglucomutase (Jollivet et al., 1995; Piccino et al., 2004; Plouviez et al., 2010; Bioy *et al.*, soumis), ainsi qu'aux hautes températures avec notamment des adaptations du collagène cuticulaire (Gaill *et al.*, 1995), de l'hémoglobine (Hourdez & Weber, 2005), de la respiration mitochondriale (Dahlhoff *et al.*, 1991). Cette forte spécialisation de la faune hydrothermale et les distances qui séparent certains champs expliquent ainsi le fort degré d'endémisme observé chez les espèces des sources hydrothermales (Tunnicliffe *et al.*, 1998).

B - Le milieu polaire Antarctique

Si l'on cherche un environnement extrême aux antipodes des sources hydrothermales du point de vue de la stabilité thermique, l'Antarctique est sûrement en première position.

Avec une superficie de 14 millions de kilomètres carrés et recouvert à 98% de glace, l'Antarctique est le continent le plus froid, le plus sec et le plus venteux, de la planète. Il est entouré d'un vaste océan, l'Océan Austral (ou Antarctique) dont les conditions thermiques sont stables et froides. Ce continent est ceinturé par le courant circumpolaire Antarctique, qui isole l'océan Austral des autres océans (Strass *et al.*, 2017) (Figure I-4). Ce courant a deux conséquences importantes dans le contexte de notre étude : il limite les échanges thermiques avec les autres océans (au Nord) et limite aussi les flux de larves (et donc de matériel génétique). Il en résulte un très fort taux d'endémisme des espèces marines qui lui sont associées en partie lié à cet isolement mais également aux caractéristiques physiologiques particulières de ces organismes.



Figure I-4 : Représentation des différents courants présents autour de l'Antarctique et des températures moyennes annuelles sous forme de thermoclines (<u>http://www.cnrs.fr</u>).

En effet, la température de l'océan Austral au niveau du plateau continental se maintient dans une gamme allant de -1.8°C à 1,5°C (Jenkins et al., 2016 ; figure I-5), soit une variation annuelle saisonnière d'environ 3°C, bien plus limitée que celle rencontrée en milieu tempéré et comparable à celle trouvée sous les tropiques dans une gamme beaucoup plus chaude (figure I-6). Les variations de température sont donc très faibles dans cet environnement, en particulier si on exclut la Péninsule Antarctique. De plus cet isolement de la faune est très ancien puisqu'il résulte de la séparation avec la Tasmanie il y a 34 millions d'années (Scher & Martin, 2006) et de l'ouverture du passage de Drake il y a 25 millions d'années qui a conduit à un renforcement de courants préexistants pour former ce courant circumpolaire (Lefebvre & Donnadieu, 2012).



Figure I-5 : Enregistrement de la température à 15 m de profondeur à la station de Rothera (Péninsule Antarctique). D'après Clark et al., 2008.



Figure I-6: Variations saisonnières de température moyenne pour différentes latitudes en abscisse pour l'Océan Indien, l'Océan Atlantique, et l'Océan Pacifique. Modifié d'après Sverdrup, Johnson, and Fleming, 1942. Les axes correspondant aux valeurs d'amplitude thermique en ordonnées pour l'Océan Atlantique et Pacifique sont placés l'un au-dessus de l'autre sur la gauche du graphique et celui de l'océan Indien et quant à lui placé au centre.

D'un point de vue thermique, si on en revient à la définition d'environnement extrême donnée précédemment, l'Antarctique constitue un environnement relativement stable du point de vue de l'amplitude de température, même s'il présente une variation saisonnière (Jenkins et al., 2016). Mais il faut tout de même prendre en compte le fait qu'on se trouve à des températures proches du point de congélation de l'eau ce qui constitue en soi une caractéristique extrême. En revanche, même si d'un point de vue thermique cet environnement est stable, son positionnement au pôle sud implique d'autres particularités. Effectivement la marque des saisons y est particulière puisqu'à l'été austral, l'Antarctique bénéficie d'un ensoleillement permanant et au contraire durant l'hiver austral, ce continent se trouve dans l'obscurité durant tout autant de temps. Ceci entraine une forte variation de la production primaire entre ces deux périodes. Ainsi avec la forte production primaire estivale, on observe chaque année à l'apparition de « blooms » ou floraisons planctoniques à la même période (Nicholson et al. 2016). Ainsi cette grande stabilité Antarctique est toute relative à la variable prise en compte. Mais quoi qu'il en soit le rythme de ces variations est bien plus lent que dans les sources hydrothermales.

Les faibles températures auxquelles sont exposés les animaux dans cet environnement imposent des adaptations moléculaires puisque dans de telles conditions la mobilité moléculaire est extrêmement diminuée, ainsi des mécanismes tels qu'une plus grande flexibilité moléculaire des sites actifs enzymatiques (Fields & Somero, 1998) ou des changements structuraux permettant indirectement d'assurer la mobilité de celui-ci (Papaleo *et al.*, 2006) constituent des preuves de l'adaptation des espèces aux contraintes de cet environnement.

C - La zone intertidale tempérée

Les zones tempérées connaissent des variations de température de l'eau de mer de surface atteignant 8-10°C au cours de l'année (figure I-6) et constituent un environnement extrêmement variable. La zone intertidale, ou zone de balancement des marées, se caractérise par une grande variabilité régie par trois grands facteurs à effet cyclique. Le premier, et le plus évident, est le balancement des marées avec une cyclicité d'environ 12h30, à quoi s'ajoute l'alternance jour/nuit et la succession des saisons. Cette particularité produit une variation cyclique de la température entre le jour et la nuit ainsi qu'au cours de l'année (Figure I-7). Dans ce type d'environnement les espèces ont une disposition et des déplacements qui dépendent de nombreux facteurs tels que la fréquence d'émersion, l'étagement, la présence de sédiment ou de roche, et la cyclicité des conditions environnementales (Cosham et al., 2016). Mais en ce qui concerne les espèces sédentaires ou peu mobiles telles que les annélides, cette variabilité prédictible puisque cyclique doit impliquer des adaptations particulières. Il a déjà été montré que des variations environnementales au cours du développement embryonnaire peuvent entraîner une adaptation chez l'adulte (Loyau et al., 2015). Des espèces comme la drosophile, qui ont un temps de génération plus court que la durée de la saison favorable, montrent des adaptations différentielles selon les saisons, notamment des modifications de fréquences alléliques cycliques, ou l'abondance relative d'espèces selon leur fitness (Behrman et al., 2015 ; Dobzhansky & Pavan, 1950). On peut donc supposer que dès lors que dans un environnement tel que les sources hydrothermales, on observe de l'adaptation à des variations « hautes fréquences » de l'environnement, une variabilité cyclique de l'environnement puisse avoir un rôle important dans la mise en place de mécanismes adaptatifs particuliers, que ce soit dans la régulation de certains gènes comme dans la mise en place de polymorphismes « adaptatifs ».



Figure I-7 : Mesures de températures moyennées sur 10 jours réalisées dans l'air et dans l'eau dans la zone du Laber à Roscoff (Jacob, 1979)

L'alternance des marées implique que les espèces placées le plus haut sur l'estran sont exposées durant des lapses de temps relativement longs à la dessiccation, ainsi des études comparant des espèces colonisant différentes parties de l'estran ont pu montrer une résistance supérieure à la dessiccation des espèces colonisant la partie haute de l'estran chez des annélides (Ferronière, 1901) ou des gastéropodes (Brown, 1960). La hauteur sur l'estran est aussi en rapport avec les capacités d'adaptation aux variations de température qu'entraine l'émersion répétée et plus ou moins longue, et permet donc une zonation des espèces et des individus au sein de celles-ci en fonction des mécanismes de réponse au stress thermique (HSP), d'adaptation du métabolisme, de thermostabilité protéique et autres qu'elles possèdent (Somero, 2002).

De plus, la plupart des espèces ont une aire de répartition couvrant une gamme de latitudes et donc des gradients environnementaux, ainsi la connectivité existante entre les populations (Piñeda et al., 2007) laisse supposer à l'existence de brassages alléliques dont certains sont issus de l'adaptation à des conditions locales particulières.

Ainsi ces trois environnements sont extrêmement contrastés, aussi bien du point de vue des températures minimales et maximales qu'ils peuvent présenter que sur le plan de la variabilité de la température dans le temps ou dans l'espace. Nous avons pu voir que la vie dans ces environnements impliquait, ou au moins favorisait certaines solutions adaptatives chez les espèces les colonisant. Dans ce contexte, les Annélides polychètes se posent comme de bons modèles d'étude car ils sont présents dans ces trois environnements, sont diversifiés, et abondants.

3 - Les Annélides polychètes

Les annélides ou vers annelés constituent un phylum que l'on retrouve dans une grande variété d'environnements sur l'ensemble du globe à l'exception des milieux arides et aériens (Rouse

& Pleijel, 2006). De plus ce phylum, très diversifié en formes et modes de vie (Figure I-8), comprend plus de 15000 espèces décrites dont environ 9000 de polychètes. En milieu marin, il constitue la macrofaune benthique dominante depuis la zone intertidale jusqu'aux environnements profonds (Rouse & Pleijel, 2001 ; Struck et al., 2011).



Figure I-8: Illustration de Matthias Jacob Schleiden sur la diversité morphologique des annélides.

Les principales synapomorphies de ce phylum sont la métamérie (organisation antéropostérieure du corps en segments), la présence d'un système circulatoire clos et la présence de soies (perdues secondairement chez les achètes et fortement réduit en nombre chez les oligochètes).

Au sein de ce phylum deux sous-ordres des polychètes présentent eux aussi une distribution cosmopolite : les Terebellomorpha et les Aphroditiformia.

Terebellomorpha (Hatschek, Berthold 1893) La famille des Alvinellidae (Desbruyères & Laubier, 1986) est inféodée aux sources hydrothermales, et compte entre autre parmis ses rangs les deux espèces sœurs, syntopiques et emblématiques des sources hydrothermales, *Alvinella pompejana* et *Alvinella caudata* (figure I-9) que l'on retrouve exclusivement le



Figure I-9: Alvinella caudata à gauche et Alvinella pompejana à droite

long de la dorsale du Pacifique Oriental. Comme nous l'avons vu précédemment, ces deux espèces ont permis de mener de nombreuses études sur l'adaptation dans cet environnement et sur l'impact de l'organisation spatiale et temporelle de celui-ci sur la structure des populations (Jollivet et al., 1995 ; Piccino et al., 2004 ; Plouviez et al., 2010 ; Bioy *et al.*, soumis).

La famille des Terebellidae (Johnston, 1846) est très proche des Alvinellidae (groupe frère), tout comme cette dernière, elle se caractérise du point de vue morphologique, par une petite lèvre supérieure au-dessus de laquelle sont implantées les tentacules péribuccaux et la présence de tissus glandulaires ventraux positionnés par paires sur les métamères (Jirkov & Leontovich, 2013). Cette famille est entre autres présente en milieu côtier avec *Terebella lapidaria* (milieu intertidal) et *Amphitrite edwardsi* (milieu subtidal) et pour l'Antarctique avec une espèce d'*Amphitritides* et une espèce de *Thelepus (Figure I-10)*.



Figure I-10: De gauche à droite, Amphitrite edwardsi / Amphitritides gracilis / Thelepus sp. (de Taïwan)

Aphroditiformia (Levinsen, 1883)

Au sein des Aphroditiformia, la famille avec le plus grand nombre d'espèces (environ 900) est celle les Polynoidae. On les retrouve à la fois dans le milieu profond (hydrothermal et non hydrothermal), polaires (Arctique et Antarctique), et dans le milieu tempéré, y compris dans la zone de balancement des marées. Les Polynoidae sont caractérisés entre autres par la présence d'élytres sur l'ensemble du corps d'où leur nom anglais de « scaleworm » (littéralement vers à écailles). Ils présentent aussi souvent deux paires d'yeux sur le prostomium, et leur proboscis est pourvu de quatre mâchoires. Tout comme pour les Terebellomorphes, nous avons choisi deux espèces pour chacun de nos trois environnements d'étude (voir figure I-11 pour une illustration de trois des espèces d'etude).



Figure I-11: De gauche à droite, Harmothoe fuligineum / Alentia gelatinosa / Lepidonotopodium fimbriatum

On retrouve donc au sein de 2 sous ordres du phylum des annélides des espèces ayant colonisé des environnements extrêmement contrastés du point de vue de la température mais partageant des traits d'histoires de vie très semblables (grande taille de populations, fécondité élevée, phase larvaire dispersive avec développement planctonotrophique ou lécithotrophique). De plus, malgré une grande diversité de traits d'histoire de vie chez les annélides polychètes dans leur ensemble, McHugh et Fong (2002) se sont basés sur des données bibliographiques pour montrer d'une part que la variabilité de ces traits est relativement limitée au sein de chaque famille et, d'autre part, ils ont pu réaliser un bilan présentant des valeurs moyennes pour chaque trait d'histoire de vie au sein de 16 familles à partir des données récoltées chez plus de 200 espèces. Les valeurs qu'ils ont obtenues pour les Polynoidae et les Terebellidae (tableau I-1) nous montrent que ces deux familles sont assez proches pour ces caractéristiques. Ce résultat ne prend tout de même pas en compte le mode de développement différent des espèces hydrothermales (larve lécithotrophique) moins commun dans d'autre environnements (larve planctonotrophique).

Famille	Age de la	Durée de	Taille	Fécondité	Taille des
	première	vie	corporelle	(nombre d'œufs	œufs
	reproduction	(années)	(mm)	ou d'embryons)	(diamètre en
	(années)				μm)
Polynoidae	1,2	3,4	64	526000	120
Terebellidae	1,2	3,0	106	146021	242

Tableau I-1 : Traits d'histoire de vie moyen chez les Polynoidae et les Terebellidae (d'après McHugh & Fong, 2002).

De plus comparer les espèces d'environnements différents au sein de chaque groupe (Aphroditiformia et Terebellomorpha) nous fournit un réplicat écologique d'espèces pour valoriser les observations qui seront effectuées.

4 - Les objectifs de la Thèse

La question de la variabilité des environnements disponibles pour une population donnée s'est déjà posée par le passé. Levene (1953) a cherché à voir dans quelle mesure il est possible d'avoir un maintien du polymorphisme dans une situation où plus d'une niche écologique est disponible. Par la suite d'autres ont suivi en cherchant les conditions (forces de sélection, taille de la niche, etc) dans lesquelles un environnement variable dans l'espace, c'est-à-dire un environnement regroupant plusieurs niches pouvait induire ou maintenir du polymorphisme (Prout, 1968; Smith, 1970; Strobeck, 1974; Smith & Hoekstra, 1980). En revanche la variabilité intrinsèque de l'environnement est peu étudiée et donc par là même l'effet que celleci a sur les espèces qui y sont exposées.

En se basant successivement sur des analyses populationnelles et spécifique, le but de cette thèse est donc de déterminer la signature que laisse la variabilité environnementale sur le génome.

- La première partie de cette thèse a pour but de comprendre les rôles respectifs des mécanismes sélectifs et démographique dans l'histoire évolutive des espèces dans l'environnement hydrothermal en prenant pour exemple les Alvinellidae du Pacifique Oriental Alvinella pompejana et Alvinella caudata.
- (ii) Dans une deuxième partie, en prenant pour modèle *H. fuligineum*, nous nous intéresserons plus spécifiquement à l'histoire démographique et à la connectivité de populations exposées à différents niveaux de variation de la température en Antarctique. Nous chercherons d'éventuels liens de parenté entre un individu des îles Kerguelen avec l'une des localités continentales. Et nous essayerons à partir de ces résultats d'en déduire le potentiel adaptatif des espèces d'Antarctique au réchauffement climatique prévu pour les années à venir.
- (iii) Enfin dans une troisième et dernière partie, il s'agira d'investiguer l'effet des variations environnementales sur les patrons de sélection en cours sur le génome en termes de diversité génétique et phénotypique en comparant des espèces appartenant aux sous ordres des Aphroditiformia et des Terebellomorpha colonisant trois environnements contrastés (l'Antarctique, le milieu côtier Breton et les sources hydrothermales
II - Obtention et traitement des données génomiques

La génétique des populations analyse le polymorphisme des espèces pour inférer leur histoire évolutive et décrire les événements migratoires et démographiques au sein de leurs populations. Ainsi de même que le nombre de pixels d'une photo est important pour avoir une bonne résolution de l'objet photographié, plus nombreuses seront les données indépendantes acquises au sein d'un génome, et meilleur sera le diagnostic au sein de l'espèce en évitant ainsi de considérer l'histoire évolutive de quelques gènes comme une vérité générale pour l'espèce ou la population considérée.

La génétique des populations a longtemps reposé sur l'analyse en fréquence de variants morphologiques et/ou enzymatiques (allozymes). Elle a cependant connu deux avancées majeures d'un point de vue technique, au niveau de la détection des mutations, tout d'abord avec l'apparition du séquençage Sanger (Sanger *et al.*, 1997) dans lequel on génère une collection de fragments d'ADN d'un gène donné par terminaison de chaîne, puis avec la mise au point du séquençage de gènes en parallèle, dit de nouvelle génération (NGS).

Dans la technique Sanger, les désoxyribonucléotides triphosphates (dNTP) normaux incorporés par la polymérase pour construire le brin complémentaire du gène ciblé sont mélangés avec des didésoxyribonucléotides triphosphates (ddNTP) liés à des marqueurs fluorescents différents. Les ddNTP empêchent la suite de l'extension du brin complémentaire lors de la réaction et génèrent une collection de fragments de différentes tailles. Ces fragments possèdent tous en dernière position un ddNTP marqué par une couleur spécifique selon que ce soit une adénine, une thymine, une guanine ou une cytosine. Les fragments obtenus sont ensuite soumis à une électrophorèse et migrent en fonction de leur taille sur un gel d'acrylamide/résine d'un capillaire. La lecture du signal fluorescent au cours de cette migration permet ainsi d'identifier successivement la nature de chaque base du fragment séquencé.

Dans le séquençage de nouvelle génération ou NGS (« Next Generation Sequencing » en anglais), un grand nombre de gènes différents sont séquencés en parallèle. Cette méthode consiste en quatre grandes étapes, quelle que soit la technique utilisée. Dans un premier temps (Figure II-1 -A), les banques NGS sont préparées par une fragmentation de l'ADN génomique et les fragments obtenus sont ligués avec des adaptateurs spécifiques contenant un site d'amorçage pour le séquençage. Dans le cas du séquençage Illumina, les banques d'ADN sont ensuite chargées sur une surface appelée « flow cell » (Figure II-1 -B). Chaque fragment ligué

à la surface est amplifié localement grâce à un système de double hybridation des brins permettant le maintien des deux fragments au sein du groupe ou 'cluster' après un cycle d'amplification. Le séquençage (Figure II-1 -C) suit ensuite le principe de la méthode Sanger. A chaque cycle de séquencage, seule une base est ajoutée à chaque fragment. Les dNTP utilisés sont marqués par fluorescence avec une couleur différente par base, ce qui permet en identifiant cette couleur au sein d'un cluster (initié par un unique brin d'ADN) de déterminer à chaque cycle la base spécifiquement ajoutée à chaque 'cluster' (gène). Après la lecture, le marquage fluorescent est éliminé par photolyse, libérant ainsi l'extrémité 3'OH, qui permet de fixer une base au cycle suivant. Les séquences propres à chaque 'cluster' sont ensuite exportées dans un fichier texte avec un code « qualité » (phred score). Ces séquences courtes (50-150 bp ou lectures : 'reads' en anglais) sont ensuite traitées par différents outils bioinformatiques (Figure II-1 -D). Ces outils permettent l'alignement des lectures obtenues sur un génome de référence préexistant ou à partir de locus construits de novo à partir desdites séquences (par association homologue) afin de pouvoir identifier et caractériser le polymorphisme des échantillons étudiés. A ces bases de données individuelles, des traitements d'assignation, de nettoyage des données sont effectués pour obtenir le plus de locus communs entre individus séquencés et appréhender le polymorphisme de l'espèce à l'échelle du génome.



Figure II-12 : Etapes principales d'un séquençage NGS Illumina (source : Illumina.com) Ce travail de thèse étant basé en grande partie sur des données Illumina, le parti a été pris de présenter l'ensemble des protocoles et techniques d'analyses dédiés au séquençage haut débit dans cette première partie afin d'éviter toute redite entre les chapitres et faciliter par la même occasion la compréhension des différents types de traitements des données et leur enchaînement dans une suite d'analyses bioinformatiques et génétiques.

La méthode utilisée pour l'obtention des séquences au cours de cette thèse est le ddRADseq (Baird *et al.*, 2008). Le principe de cette technique est relativement simple. L'ADN génomique est digéré par 2 enzymes de restriction à haute fidélité de reconnaissance d'un motif d'ADN, l'un (MseI) coupant plus souvent que l'autre (PstI) dans le génome. Ceci permet d'accéder à un sous-échantillonnage du nombre de locus dans un génome en n'utilisant que les fragments d'ADN présentant les deux sites de restriction et séparés par une distance maximale de 300 bases. Cette technique permet donc d'échantillonner dans le génome des mutations polymorphes sans *a priori* mais d'une façon ciblée afin d'obtenir le plus grand nombre de locus

orthologues entre individus avec une couverture suffisante si la taille du génome le permet. Pour obtenir des données de polymorphisme sur nos espèces, nous avons donc :

- (1) construit nos banques de fragments d'ADN
- (2) envoyer ces banques à une plateforme de séquençage haut débit (McGill university, Canada)
- (3) traiter les données de séquences obtenues (nettoyage, filtrage, obtention des locus), et
- (4) analyser les patrons de diversité génétique au sein et entre espèces

Les étapes de construction des banques (méthodes de biologie moléculaire) avant le séquençage et le traitement des données de séquence (outils bioinformatiques) sont expliqués en détail dans les parties qui suivent.

1 - Traitement des échantillons et construction des banques individuelles d'ADN

A - Préparation des ADN

-Extraction de l'ADN génomique avec la méthode du Phénol/Chloroforme

Les échantillons de tissus (congelés (-80°C) ou alcoolisés (éthanol 80%) sont incubés à 55°C dans un mélange réactionnel contenant 500µL de tampon PK (Tris-HCl 50mM, NaCl 100mM, EDTA 25mM, SDS 1%) et 10µL de protéinase K (10mg/mL) pendant une nuit jusqu'à la lyse complète de l'échantillon. Les échantillons sont ensuite incubés 30 min à 37°C avec 10µL de RNAse A (10ng/mL) pour éliminer les ARNs. Ils sont enfin incubés 10min à 72°C pour dénaturer d'éventuelles traces de DNAses.

Du phénol tamponné (pH 8,0) est ensuite ajouté volume pour volume au mélange pour précipiter les protéines par émulsion pendant une minute puis centrifugé à 12000 rpm durant 5 min pour ne récupérer que la phase aqueuse. Cette étape est répétée 2 fois en remplaçant le phénol d'abord par un mélange phénol/chloroforme (50:50), puis par un mélange chloroforme/alcool isoamylique (24:1) pour éliminer lipides et carbohydrates. Enfin l'ADN est précipité en ajoutant à la phase aqueuse récupérée deux fois son volume en éthanol absolu glacé avant de placer le tout à -20°C sur la nuit.

-Nettoyage de l'ADN

Les ADN précipités sont centrifugés à 12500 rpm pendant 10 min pour les culoter et se débarrasser du surnageant. Trois nettoyages successifs sont réalisés, les deux premiers avec de l'éthanol 75% froid, après ajout de PVPP (1%) pour éliminer les polysaccharides et les polyphénols. Le dernier lavage est fait avec de l'éthanol absolu froid pour déshydrater l'ADN et faciliter son séchage. Pour chaque nettoyage, 500µL d'éthanol est ajouté au culot et vortexé pendant 10s avant d'être centrifugé à 12 500 rpm pendant 10 min afin d'éliminer l'éthanol par pipetage. Pour sécher complètement le culot d'ADN, l'échantillon est passé trois minutes au SpeedVAC puis re-suspendu dans 150µL d'eau milliQ autoclavée. L'ADN re-suspendu est ensuite purifié une deuxième fois à l'aide du kit Nucleospin gDNA Clean-up XS (Macherey-Nagel) en suivant les instructions du manuel puis conservé à -20°C jusqu'à son utilisation. Cette dernière purification est très importante chez de nombreuses espèces d'annélides polychètes pour éliminer toute trace de polyphenols/polysaccharides qui interagissent négativement avec la digestion de l'ADN par les enzymes de restriction.

Les ADN génomiques sont ensuite visualisés sous UV après une migration sur gel d'agarose (0,8%) contenant du GelRed et, quantifiés au Nanodrop pour obtenir une concentration d'environ 200 ng/µL.

B - Construction des banques ddRAD

-Double digestion de l'ADN génomique

150 à 200ng d'ADN dans un volume de 20 μ L sont utilisés pour la digestion. Il est additionné aux 30 μ L de mélange réactionnel contenant 2,5 μ L de Cutsmart buffer 10X (New England Biolaps), 0,4 μ L d'enzyme PstI HF (20u/ μ L; site de reconnaissance : CTGCAG), 0,2 μ L d'enzyme MseI (10u/ μ L; site de reconnaissance : TTAA), 1,9 μ L d'eau milliQ autoclavée. Le tout est incubé pendant 10h à 37°C puis 20min à 65°C avant de repasser à 4°C pour conservation.

-Production des adaptateurs P1 barcodé et P2 non barcodé

Les adaptateurs sont fournis sous forme d'amorces simple brin (100µM) qu'il est nécessaire de liguer entre eux au préalable (Encadré II-1).

La préparation des adaptateurs double brins contenant les barcodes Illumina est effectuée selon le protocole décrit par Petterson et al. (2012). Deux types d'adaptateurs P1 et P2 sont produits

selon qu'ils contiennent le site PstI ou MseI à partir d'amorces simple brin (produits par Eurofins). Pour les adaptateurs barcodés contenant le site de restriction PstI, les amorces forward et reverse P1-1 et P1-2 (10 μ M) sont hybridées entre elles sous l'effet d'une température décroissante (0,1°C/s) après une étape de dénaturation de 5 min à 95°C (effectuée avec machine PCR Biorad) dans un tampon d'hybridation (Tris-HCl 10mM, NaCl 50mM, EDTA 1mM, pH 8.0). Les adaptateurs P2 avec site de restriction MseI sont obtenus de la même façon avec des amorces complémentaires P2-F et P2-R (100 mM) pour obtenir 10 fois plus d'adaptateurs P2 que de P1 (l'enzyme MseI coupant beaucoup plus souvent que l'enzyme PstI).

Pour la ligation des adaptateurs aux fragments d'ADN issus de la digestion enzymatique, 20μ L d'ADN digéré sont utilisés dans un mélange contenant 3,4 μ L d'adaptateur P1 (double brin, 1 μ M), 1 μ L d'adaptateur P2 (double brin, 10 μ M), 0,4 μ L de T4 DNA ligase (1-3U/ μ L, Promega), 2 μ l de Cutsmart buffer (10X) et 3 μ L d'ATP (10mM) pour atteindre un volume réactionnel de 36,6 μ L, qui est ensuite incubé à 16°C pendant 6h, puis passé à 4°C.

Encadré II-1 : Liste des adaptateurs Illumina P1 et P2 utilisés. Les barcodes (combinaison de 6 bases à façon) sont soulignés en gris, les séquences identifiées en jaune et orange sont celles permettant l'hybridation des adaptateurs avec les amorces forward et reverse universelles utilisées pour l'amplification des banques.

P2.R_Msel	TAAGATCGGAAGAGCGAGAACAA
P2.F_Msel	<mark>G T G A C T G G A G T T C A G A C G T G T G T G C</mark> T C T T C C G A T C T
P1_PstI_01.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G G T C T T G C A
P1_PstI_02.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G C A
P1_PstI_03.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T A A G A T A T G C A
P1_PstI_04.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T A C T T C C T G C A
P1_PstI_05.F	A C A C T C T T T C C C T A C A C G A C G C T C T T T C C G A T C T T A C G G T G C A
P1_PstI_06.F	A C A C T C T T T C C C T A C G A C G C T C T T C C G A T C T A A C G A A T G C A
P1_PstI_07.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T T C A T T C A T T G C A
P1_PstI_08.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G C A
P1_PstI_09.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T T C C T G C A
P1_PstI_10.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T C A T G C A
P1_PstI_11.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G C C T G G T G C A
P1_PstI_12.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T T G C T T G C T G C A
P1_PstI_13.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T T C G C A T T G C A
P1_PstI_14.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G G T A G A T G C A
P1_PstI_15.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G G A G C G T G C A
P1_PstI_16.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T T G A A C T G C A
P1_PstI_17.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G C A
P1_PstI_18.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T C G A G G C T G C A
P1_PstI_19.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T C C A A C C G T G C A
P1_PstI_20.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G T A T G A T G C A
P1_PstI_21.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T G G A T T T G C A
P1_PstI_22.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T C C A G C T T G C A
P1_PstI_23.F	A C A C T C T T T C C C T A C A C G A C G C T C T T C C G A T C T A A C T C G T G C A
P1_PstI_24.F	<mark>A C A C T C T T T C C C T A C A C G A C G</mark> C T C T T C C G A T C T A C C A G A T G C A

-Sélection d'une gamme de taille de fragments supérieure à 200 bp à l'aide de billes AMPure XP (Beckman)

Cette étape permet de réaliser une sélection de taille des fragments digérés après ligation des adaptateurs. La taille minimale des fragments retenus sur les billes AMPure dépend du ratio entre le volume de billes (après homogénéisation) et le volume d'ADN fragmenté quelle qu'en soit la concentration.

La solution d'ADN + adaptateurs est complétée à 40μ L avec de l'eau milliQ autoclavée puis mélangée, à température ambiante pendant 5 min, avec 60μ L de billes AMPure XP (conservé à température ambiante pendant 30 min avant mélange) pour obtenir un ratio de 1,5X et permettre aux fragments de se lier aux billes. Le liquide contenant les fragments non désirés (<200bp) est alors éliminé après 2 lavages successifs à l'éthanol 80%. Pour le nettoyage, les billes sont maintenues par aimantation et rincées pendant 30s avec 200 μ L d'éthanol. Après cette étape, les billes sont séchées pendant 5 min à l'air libre. L'ADN est ensuite élué dans 40μ L de Tris 10mM jusqu'à re-suspension totale du culot de billes. Le mélange est laissé 5min à incuber à température ambiante puis replacé sur le support aimanté pour agréger les billes sur la paroi, et permettre le pipetage d'un peu plus de 35 μ L d'ADN en solution en prenant garde de ne pas pipeter de billes.

-Amplification des fragments par PCR avec les amorces contenant les index Illumina

Les fragments (>200 bp) issus de la double restriction enzymatique sont amplifiés spécifiquement à l'aide des amorces ILLPCR1 (forward) et ILLPCR2 (reverse) qui correspondent à une région complémentaire des 2 adaptateurs. Lors de cette phase d'amplification, l'amorce forward est indexée (12 index Illumina possibles), ce qui permet de marquer une deuxième fois nos fragments pour obtenir un double indexage (barcode-index illumina) de l'échantillon (Encadré II-2).

Encadré II-1: Liste des amorces Illumina ILLPCR1 et ILLPCR2 utilisées dans l'amplification spécifique des fragments contenant les deux sites de restriction MseI et PstI (les index Illumina sont surlignés en gris, les séquences identifiées en jaune et orange sont les régions s'hybridant avec les adaptateurs P1 et P2 lors de l'amplification du fragment).

ILLPCR1	a a T G A T A C G G C G A C C A C C G A G A T C T <mark>A C A C T C T T T C C C T A C A C G A C G</mark>
ILLPCR2_IND01	c a A G C A G A A G A C G G C A T A C G A G A T C G T G A T G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND02	c a A G C A G A A G A C G G C A T A C G A G A T A C A T C G G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND03	c a A G C A G A A G A C G G C A T A C G A G A T G C C T A A G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND04	c a A G C A G A A G A C G G C A T A C G A G A T T G G T C A G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND05	c a A G C A G A A G A C G G C A T A C G A G A T C A C T G T G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND06	c a A G C A G A A G A C G G C A T A C G A G A T A T T G G C G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND07	c a A G C A G A A G A C G G C A T A C G A G A T G A T C T G G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND08	c a A G C A G A A G A C G G C A T A C G A G A T T C A A G T G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND09	c a A G C A G A A G A C G G C A T A C G A G A T C T G A T C G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND10	c a A G C A G A A G A C G G C A T A C G A G A T A A G C T A G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND11	c a A G C A G A A G A C G G C A T A C G A G A T G T A G C C G T G A C T G G A G T T C A G A C G T G T G C
ILLPCR2_IND12	c a A G C A G A A G A C G G C A T A C G A G A T A C A A G G T G A C T G G A G T T C A G A C G T G T G C

Le mélange réactionnel de 40 μ L est composé de 8 μ L d'ADN fragmenté après ligation des adaptateurs, 8 μ L du tampon de la Q5 polymérase (5X, New England Biolabs), 3,2 μ L de dNTP (2,5mM chaque), 0,4 μ L de Q5 polymérase High-Fidelity (2u/ μ L), 11,04 μ L d'eau milliQ autoclavée et de 1,36 μ L d'un mélange contenant les amorces ILLPCR1 et ILLPCR2 (5 μ M chaque, en prenant une combinaison spécifique de barcode/index par échantillon). La première phase de normalisation de la quantité d'ADN consiste dans une première phase de dénaturation à 98°C pendant 30s puis de 15 à 17 cycles d'amplification (dénaturation 98°C/20s, hybridation 60°C/30s, élongation 72°C/40s) et une phase finale d'élongation à 72°C pendant 10 min.

Après la normalisation, le milieu réactionnel est subdivisé en 4 lots de 10 μ L auxquels sont ajoutés 0,2 μ L de tampon Q5, 0,8 μ L de dNTP (2,5mM chaque) et 0,4 μ L de mélange contenant les amorces pour un cycle final d'amplification (dénaturation 98°C/3min, hybridation 60°C/2min, élongation 72°C/12min). Les lots sont rassemblés dans un même puits et le profil d'amplification est contrôlé sous UV après une migration sur gel d'agarose (1,5%) dans un tampon TBE (Tris (89mM), Borate (89mM), EDTA (2mM)). Ces échantillons d'ADN fragmentés avec site de séquençage indexé Illumina constituent alors nos banques individuelles ddRADseq prêtes à être séquencées.

-Pool des individus et préparation pour séquençage

La concentration en ADN de chaque banque individuelle est mesurée avec un fluoromètre QuBit sur la plateforme Genomer de la SBR afin de faire un mélange iso-stoechiométrique des individus (quantité égale de chaque individu/banque dans le mélange à séquencer). Une fois le « pool » réalisé, une dernière sélection de taille est réalisée à l'aide du kit Pippin Prep (1,5% Agarose Gel Cassette, type de cassette 1,5% DF Marker K) pour ne garder que des fragments compris entre 200-600 bp. La distribution en taille des fragments de la banque multiindividuelle est alors visualisée avec un BioAnalyzer Agilent en utilisant une puce high sensitivity DNATM avant d'être envoyée à séquencer à la plateforme dédiée de l'université McGill au Québec. Les trois séquençages ayant permis l'acquisition des données de ce manuscrit ont été réalisés de façon pairée (séquençage des 250 premières bases de chaque extrémité) sur une ligne de séquençage avec un séquenceur HiSeq 2500/Illumina générant environ 400 millions de lectures par ligne en mode pairé. Les séquences ont ensuite été démultiplexées sur la base des index et un filtre des phred score a été appliqué pour supprimer les données les plus mauvaises.

2 - Les mesures de la diversité génétique et de différenciation

A - Mesures intra-population

-Diversité génétique à un locus donné

Lorsqu'un gène en particulier est étudié, son séquençage apporte des informations basiques apportant une première idée de la diversité génétique de la population échantillonnée. On peut retrouver entre autres le nombre de sites polymorphes ou ségrégeants (S), le nombre d'haplotypes observés (H) ou le nombre moyen de différences observées entre deux séquences aléatoires de la population pour ce locus. La comparaison entre ce dernier et l'estimateur de diversité S permet via le D de Tajima de mettre en évidence l'effet d'une expansion démographique faisant suite à un goulot d'étranglement puisque dans une telle situation on s'attend à une perte du nombre de sites ségrégeants et une élévation du nombre moyen de différences entre séquence en raison du maintien de lignées aléatoirement conservées (Tajima, 1989). On peut aussi calculer le thêta de Watterson (Θ w) qui correspond au nombre de sites ségrégeant rapporté à la taille de la séquence considérée (Watterson, 1975). Nombre moyen d'allèles par locus (A) ou allélisme, comme son nom l'indique, traduit la tendance moyenne des locus d'une population en termes de nombre d'allèles. A l'inverse du polymorphisme, cet indice ne prend pas en compte les fréquences respectives des différents allèles. Cet indice est couramment utilisé en écologie de la conservation (Asins & Carbonell 1987 ; Bataillon et al., 1996). Le polymorphisme nucléotidique (π) correspond au nombre moyen de différences nucléotidiques par site et par paire de séquences (Nei 1987, Tajima 1983) est un autre moyen de quantifier la variabilité. Un autre avantage de cette méthode est que l'on peut distinguer le polymorphisme synonyme π s du non-synonyme π n lorsque l'on s'intéresse aux régions codantes du génome.

-Tajima

Le test de Tajima est un test statistique utilisé en génétique des populations. Le D de Tajima est calculé comme la différence entre deux mesures de la diversité nucléotidique : le nombre moyen de différences par paire de séquences à un locus dans une population et le nombre total de sites variables trouvé sur l'ensemble des séquences à ce même locus dans la même population. Ces deux estimateurs de la diversité génétique θ telle que $\theta=4N.\mu$) sont égaux dans une population de taille constante et évoluant de façon neutre (D = 0). Ainsi pour un locus soumis à la sélection balancée, et donc où au moins deux allèles sont maintenus dans le polymorphisme, le nombre moyen de différences entre deux séquences va être plus élevé que l'attendu neutre puisque les allèles sous sélection ne recombinent pas et accumulent des mutations séparément. Dans ce cas, le D de Tajima prend des valeurs positives en raison d'une plus forte augmentation du nombre moyen de différences par paire de séquences.

Un effet démographique engendrant une réduction drastique de la taille de la population (ou goulot d'étranglement) conduit à un même effet sur le test de Tajima avec des valeurs significativement positives du test, mais à l'inverse de la sélection balancée qui n'affecte que le locus visé, un goulot d'étranglement aura une action généralisée sur le génome, et donc fournira des valeurs positives pour l'ensemble des locus.

A l'inverse, lorsqu'un locus est sous sélection purifiante ou subit un balayage sélectif, la sélection va conduire à une diminution du nombre moyen de différences entre deux séquences et a des valeurs négatives du D de Tajima. En particulier, avec un balayage sélectif, l'allèle associé à l'arrivée d'une mutation favorable va se répandre en quelques générations dans la population et éliminer le polymorphisme associé aux autres allèles. Ce processus évolutif est assez rare et dépend de la vitesse à laquelle les mutations apparaissent dans la population, et est

donc fortement dépendant de la taille efficace de la population. De ce fait, ces effets affecteront un nombre limité de locus. A l'inverse, une population subissant une expansion démographique après un goulot d'étranglement va se caractériser par un excès de nouvelles mutations en faible fréquence (singletons = 1/2N avec N= taille efficace). Cet excès de mutations rares est comparable à ce que l'on peut attendre d'un balayage sélectif mais cette fois affectera l'ensemble du génome. Dans notre analyse multilocus sur le génome, le nombre de locus présentant des valeurs de D significativement différentes de zéro nous permettra donc de trancher entre un effet démographique ou sélectif et le signe positif ou négatif de la différence entre estimateurs sur le mode de sélection ou l'effet démographique envisagé.

-Polymorphisme multi locus

La proportion de locus polymorphes (P), aussi appelée taux de polymorphisme ou polymorphisme correspond au nombre de locus variables par rapport au nombre total de locus étudiés. On considère généralement qu'une population est polymorphe à un locus donné si la fréquence de l'allèle le plus commun est inférieure à 0,95. En revanche cet indice ne permet pas de distinguer les locus polymorphes selon le nombre d'allèles ou d'allèles majoritaires. Taux d'hétérozygotie (h_e et h_o)

Le taux d'hétérozygotie ou diversité génétique de Nei (1973) peut se définir en h_o (hétérozygotie observée) qui correspond à la proportion observée d'individus hétérozygotes dans la population ou h_e (hétérozygotie attendue) correspondant à la proportion d'individus hétérozygotes sous l'équilibre d'Hardy-Weinberg. Pour n allèles d'un locus donné à la fréquence f_n dans la population, l'hétérozygotie attendue est calculée avec la formule suivante :

$$h_e = 1 - (f_1^2 + f_2^2 + f_3^2 + \dots + f_n^2) = 1 - \sum_{i=1}^n f_n^2$$

ou f_n^2 représente la fréquence des génotypes homozygotes. Si plusieurs locus sont considérés, on peut calculer l'hétérozygotie moyenne (He), représentant la moyenne arithmétique du taux d'individus hétérozygotes attendus pour chaque locus.

Indice de fixation (Fis)

Le F_{is} de Wright, initialement appelé coefficient de consanguinité (Wright, 1969), est obtenu avec la formule suivante : $F_{is} = (h_e - h_o)/h_e = 1 - (h_o/h_e)$

Avec h_o , l'hétérozygotie observée et h_e l'hétérozygotie attendue, calculée à partir des fréquences alléliques sous l'hypothèse de Hardy-Weinberg.

Cet indice traduit le degré de différenciation génétique des individus au sein d'une population. Dans le cas d'une population à l'équilibre de Hardy-Weinberg on a $h_e = h_o$ et donc $F_{is} = 0$. Lorsque le taux d'hétérozygotie est supérieur à l'attendue, ce qui peut être le cas avec de la sélection balancée favorisant les hétérozygotes par exemple, $h_e > h_o$ et F_{is} est négatif. Enfin dans le cas d'un avantage aux homozygotes par exemple ou d'un effet Wahlund (mélange d'individus provenant de populations différenciées) on va observer un déficit d'hétérozygotes dans la population par rapport à l'attendue ($h_e > h_o$) et le F_{is} sera d'autant plus proche de 1 que le phénomène sera marqué.

B - Mesures inter populations

De même que le Fis traduit la différenciation des individus au sein d'une population, Wright (1969) a aussi défini deux autres paramètres populationnels. Considérons un ensemble T formé de populations (ou sous-populations) S contenant des individus I. Wright a défini le Fit comme correspondant à la différenciation des individus au sein de l'ensemble considéré et le Fst comme indice de différenciation entre les populations. Le Fit est calculé de la même manière que le Fis en utilisant les hétérozygoties attendue et observée pour l'ensemble des individus (toutes populations confondues). Enfin la différenciation entre les populations est estimée à l'aide de la relation suivante (avec Fis étant une moyenne des Fis de chaque population considérée) :

$$Fst = 1 - \frac{1 - Fit}{1 - Fis}$$

Cette liste non exhaustive d'outils permettant d'estimer la diversité génétique nous donne déjà un aperçu des possibilités qui s'offrent à nous. Certains indices peuvent être plus ou moins sensibles à des phénomènes démographiques ou au contraire à de la sélection, ce qui nous permet en les alliant et en résonnant par élimination de se faire une première idée des causes expliquant les patrons de diversité observés.

3 - Nettoyage des lectures et identification des populations

A - Filtrage, nettoyage et assignation (mapping) des lectures (reads) à un génome de référence

-Tri et nettoyage des lectures - Process-Radtag

Dans un premier temps il faut assigner les lectures obtenues par séquençage Illumina à chaque individu présent à partir du double indexage des séquences (index illumina – barcode). En utilisant l'index illumina (6 bases) séquencé en amont de la lecture et le barcode (de 6 bases, situé sur le read1 après les 4 bases du site de restriction PstI), le programme Process_Radtags appartenant au pipeline Stacks (Catchen *et al.*, 2013) nous permet d'assigner chaque lecture à un individu donné, tout en prenant en compte les erreurs de séquençage ayant pu se produire en début et fin de chaque séquence en utilisant un nombre de paramètres optionnels :

- Suppression des lectures présentant plus d'un mésappariement dans le motif de reconnaissance du barcode ou de l'index Illumina
- Suppression des lectures dans lesquelles on retrouve une partie ou la totalité de la séquence de l'adaptateur (i.e. lecture d'une taille inférieure à 150 bp qui aurait échappé à la sélection de taille des fragments, la longueur du fragment séquencé étant ici de 125 bp)
- Suppression des lectures de basse qualité (un premier filtre ayant déjà été effectué lors du démultiplexage des index après séquençage par la plateforme de séquençage McGill, peu de lectures sont éliminées avec ce filtre lors de notre démultiplexage)
- Suppression des lectures dans lesquelles le motif de restriction attendu en début de séquence n'est pas présent.

Le démultiplexage à proprement parler est réalisé sous le Cluster de la plateforme de bioinformatique ABIMS avec la commande Unix suivante à partir de fichiers de séquences fastq zippés (.gz) :

process_radtags -i gzfastq -P -p A -b B -y fastq --inline_index -q -D -r --renz_1 pstI -renz_2 mseI --adapter_2 TGCA --adapter_1 TAAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

ou process_radtags est le nom du logiciel d'analyse avec : -i : format des fichiers fournis au programme -P : Indique au programme s'il s'agit de données pairées

-p A – nom des fichiers R1 et R2 séparés contenant les 'reads' pairés (si 'single reads', mettre un seul fichier, et supprimer l'option –P).

-b B – nom du fichier texte contenant la liste des individus et leur barcode/index

-y : format demandé pour les fichiers de sortie

--inline_index : pour signifier que le barcode est intégré dans la séquence et l'index dans le nom de la séquence

-q : filtre les lectures en fonction de leur qualité sur la base du Phred Score

-D : conserve dans un fichier supplémentaire les lectures filtrées

-r : conserve les barcodes et les index dans les fichiers de sortie

--renz_X : supprime le motif de l'enzyme de restriction utilisé en forward (1) ou reverse

(2)

--adapter_X : supprime la séquence de l'adaptateur en forward (1) ou reverse (2)

n.b. : Les fragments d'ADN contenus dans les banques subissant un double séquençage (par chaque extrémité), les données pairées correspondent aux séquences pour lesquelles on connait la correspondance de lecture de l'autre extrémité. L'ensemble de ces séquences est contenu dans deux fichiers distincts (R1 et R2) où celles-ci sont nommées de façon à pouvoir reconnaitre ces paires.

-« Mapping » des lectures sur un génome ou un transcriptome de référence – Bowtie2

Une fois les séquences attribuées à chaque individu, l'étape d'assignation des lectures à un locus est réalisée à l'aide du script Bowtie2 (Langmead & Salzberg, 2012). De nombreuses options sont disponibles dont, entre autres, le type d'alignement utilisé pour assigner les séquences à telle ou telle position sur le génome de référence. Ces deux types d'alignement sont nommés « Local » ou « End to End ».

L'alignement « End-to-end » essaye d'aligner les séquences sur la totalité de leur longueur et est plus souvent utilisée quand les séquences sont très similaires (e.g. lectures d'*Alvinella pompejana* sur son propre génome) et de même longueur (e.g. 120 bp après 'trimming'). Dans le cas de l'alignement « Local », une méthode heuristique rapide permet d'effectuer des calculs de scores adaptés qui évitent de pénaliser les régions non-homologues et ne calculent le score que sur les régions conservées. Cette deuxième option est privilégiée lorsque que l'on suppose l'existence d'une divergence entre les lectures et la référence utilisée à un locus homologue ou

entre les allèles d'un même locus. Une première étape de conditionnement du fichier de référence est nécessaire. Ce conditionnement est réalisé par le biais de la commande suivante :

bowtie2-build A B

avec : A – le nom du fichier contenant les séquences (génome, transcriptome ou autre) de référence

B - le nom attribué au fichier de sortie conditionné

Une fois le fichier contenant la séquence de référence conditionné, l'assignation des lectures est réalisée à l'aide de la commande suivante avec les options que j'ai plus spécifiquement choisies de mettre en œuvre pour mes analyses :

bowtie2 { -1 A -2 B --no-mixed --no-discordant | -U C } -x D -S { --end-to-end | --local -N 1 -L 18 } -I 0 -X 300

(pour traiter plusieurs individus à la fois, cette commande peut être dupliquée selon le nombre d'individus analysés en changeant le nom des fichiers d'entrée à chaque ligne, et insérée dans un shell pour lancer une commande qsub)

Avec : A et B – noms des fichiers .fastq des lectures (R1 & R2) pour chaque individu avec A contenant les R1 et B contenant les R2

--no-mixed : option indiquant que seuls les lectures pairées sont prises en compte

--no-discordant : option indiquant que seules les lectures assignées une seule fois sur le génome sont prises en compte.

Les conditions d'assignation sont définies en suivant les paramètres suivants :

C – nom du dossier contenant les fichiers fastq des lectures pairées et non pairées

D - nom du fichier contenant les séquences de référence conditionnées

S – nom du dossier de sortie pour les alignements au format .bam ou .sam

--end-to-end ou --local : type d'alignement désiré

-N: nombre de mésappariements tolérés dans la graine servant d'initiation à l'alignement

-L : longueur de la graine servant d'initiation à l'alignement

-I : taille minimale des fragments entre les sites de restriction MseI et PstI (en pairées)

-X : taille maximale des fragments entre les sites de restriction MseI et PstI (en pairées)

Le nombre de mésappariements sur la graine peut aller de 0 à 2 et la taille de la graine de quelques bases à la longueur totale de la lecture. Le choix de ces paramètres dépend de la divergence moyenne estimée entre les lectures et la séquence de référence. Dans le cas d'une assignation de lectures d'une espèce sur un transcriptome de référence d'une espèce proche, notre choix s'est porté sur 1 mésappariement sur une graine de 20 bases pour être proche d'une divergence de 5%.

Selon les analyses envisagées, les lectures d'une espèce ont été assignées aux génomes ou aux transcriptomes de référence des mêmes espèces obtenus par assemblage avec Velvet/Oases ou Trinity par l'équipe. Seul le génome d'*A. pompejana* a été fourni par le Consortium *Alvinella* (coord. A. Claridge –Chang, A-Star* Institute, Singapore) à partir d'un assemblage sous Opera. Les transcriptomes générés par Velvet/Oases ou Trinity contiennent en général plusieurs isoformes pour chaque transcrit (locus), une assignation directe des lectures sur ces transcriptomes aurait pour effet une dilution de la couverture des lectures à un locus donné. De plus les analyses envisagées utilisant les transcriptomes d'annélides visent à caractériser le polymorphisme synonyme et non-synonyme des gènes, et nécessitent donc une réduction des transcrits à leur séquence codante. L'assignation des lectures à un transcriptome nécessite donc de filtrer et réduire les transcrits afin d'obtenir un cadre de lecture ouvert (ORF : « open reading fragment » en anglais) pour chaque transcrit et réduire la redondance du transcriptome (réduction de 150 000 transcrits initialement à environ 30 000).

-Réduction des transcriptomes à un allèle par transcrit – Filter assemblies et Filtration des ORF – transcriptsToOrfs

Lors de la réduction en taille des transcriptomes, le script écrit par Eric Fontanillas et intégré à la plateforme Galaxy par Julie Baffard ne conserve que les transcrits dépassant 100 paires de bases et l'allèle le plus long pour le Phred score le plus élevé pour un transcrit donné. Le script conserve le premier des allèles rencontrés dans le cas où les allèles d'un même transcrit font la même taille. Le script effectue ensuite un Cap3 sur les transcrits filtrés (après suppression des extrémités 5' et 3') pour assembler des portions d'un même transcrit qui aurait été séparées par l'assembleur à cause d'erreurs de séquençage sur le début et la fin de la séquence (couverture moins importante en 3' et 5').

Après ce premier traitement, une recherche de l'ORF le plus probable est effectuée avec transcriptsToOrfs. Cet autre script intégré à la plateforme Galaxy recherche, dans les 6 cadres de lecture possibles, la séquence codante la plus longue entre 2 codons stop pour chaque transcrit.

B - Constructions des catalogues de locus

A la suite de cette assignation des lectures à un génome/transcriptome de référence (fichiers .bam), les données de séquençage sont traitées à l'aide des outils du logiciel Stacks (Catchen et al., 2013) pour la construction des catalogues de locus. Stacks ne prenant pas en compte le pairage des données, les lectures R1 et R2 sont concaténées l'une à la suite de l'autre avec la commande Unix 'cat –n fichier1 fichier2'.

-Catalogues individuels - U / P stacks

La première étape de cette construction passe par l'analyse indépendante d'homologie de séquences de chaque individu, ou toutes les lectures sont comparés deux à deux sur la base de paramètres définissant leur degré d'identité. Selon que l'on dispose ou non d'un génome de référence, deux programmes semblables sont utilisables pour réaliser cette tâche. Dans le cas de l'utilisation de données assignées à un génome de référence pour lesquelles chaque lecture à une position connue, P-stacks assemble au sein d'un même locus (pile) toutes les lectures partageant la même position sur le génome pour un individu donné et le nombre de lectures pour un locus correspond à sa couverture. Dans le cas où les lectures ne sont pas positionnées sur une référence, c'est le programme U-stacks qui effectue 'de novo' le regroupement des lectures en piles selon leur degré d'apparentement (%identité) et la contrainte d'une couverture minimale en nombre de lectures pour définir un même locus. Dans ce cas, 2 lectures pourront être assemblées au sein d'une même pile si elles ont un nombre de différences en paires de bases inférieur au paramètre M et un locus ne pourra être gardé que si le nombre de lectures pour un allèle donné est supérieur au paramètre de couverture m.

Pour lancer cette procédure, la commande utilisée pour l'utilisation de P-stacks et U-stacks est la suivante :

pstacks -t sam -f A -o B -i X -m 1 -p N ustacks -t fastq -f A -o B -i X -m 2 -M n -R -p N (pour traiter plusieurs individus à la fois, cette commande peut être dupliquée selon le nombre d'individus analysés en changeant le nom des fichiers d'entrée à chaque ligne et insérée dans un shell pour lancer une commande qsub)

avec : A – le nom du fichier d'entrée qui correspond aux lectures soit assignées à une référence (.bam ou .sam), soit non assignées (.fastq)

B – le nom du fichier de sortie

-t : format des fichiers fournis au programme (bam, sam fastq)

-i : indique au programme que les lectures 1 et 2 sont dans des fichiers séparés avec X
 le répertoire

-m : couverture minimale des piles (m=1 signifie qu'il faut au moins 2 lectures pour définir une pile, m=2, au moins 4 lectures, etc)

-M : nombre (n) de différences entre 2 lectures pour qu'elles appartiennent à un même locus

-R : conserve les lectures non utilisées dans un fichier à part

-p : permet une exécution du calcul en parallèle sur N cœurs du Cluster de la plateforme ABIMS

Remarque : La couverture jugée suffisante pour créer une pile n'a pas été la même selon les espèces étudiées et selon que les lectures aient été assignées ou non à un génome de référence. En effet dans le cas de données préalablement assignées, la couverture m=1 a été jugée suffisante puisque que l'assignation sur génome nous permet d'être confiants vis-à-vis des locus obtenus. Dans le cas de la construction des piles '*de novo*' (pour *H. fuligineum*), une divergence de n=10 a été acceptée entre deux allèles d'un même locus et une couverture m=2 (4 lectures) a été choisie dans l'optique de faire un compromis entre confiance dans l'information conservée et perte de locus au sein de chaque individu. Cette couverture est néanmoins insuffisante pour avoir une bonne confiance dans la définition des différents allèles au sein de chaque locus mais le nombre de lectures par individu (souvent < 1 million) et le nombre important de locus sur le génome ne permettait pas une couverture suffisante pour de nombreux individus à un locus donné.

Lors de cette étape, plusieurs catalogues sont générés par individu dont un contenant la liste de tous les locus polymorphes avec les différents allèles identifiés et un autre contenant les séquences de chaque locus (polymorphe ou non) en identifiant les SNPs selon leur nature homozygote ou hétérozygote. Un simple compte permet à partir de ces fichiers de savoir combien de locus polymorphes et combien de SNPs sont présents par individu. Cette information permet de réaliser un tri des individus afin de ne conserver que ceux suffisamment informatifs pour rechercher les locus communs entre individus.

Le compte du nombre de locus se fait sur les fichiers de sortie avec l'extension .alleles.tsv. La commande utilisée consiste en un compte des noms d'allèles dans la troisième colonne du fichier sans prendre en compte les répétitions. La commande pipée est la suivante :

cut -f 3 X.alleles.tsv | uniq | wc -l | cut -f 1

Pour le compte du nombre de SNPs, le fichier utilisé est celui avec l'extension .snps.tsv. Ce fichier présente l'ensemble des locus identifiés pour l'individu donné où chaque ligne correspond à une base d'un locus, avec des locus s'enchaînant les uns après les autres. Chaque SNP est identifiée par la lettre E, il suffit alors de compter les E dans le fichier à l'aide de la commande grep suivante :

grep -c "E" X.snps.tsv

-Catalogue populationnel des locus – C-stacks

Une fois les catalogues individuels construits, ceux-ci sont comparés entre eux afin de construire un catalogue populationnel. Comme précédemment, la divergence n'est pas prise en compte lorsqu'on utilise des données assignées à un génome de référence puisque le programme ne se base que sur les coordonnées des piles le long du génome de référence. Dans le cas contraire, un nouveau paramètre de divergence entre individus (n) est à définir. Il est souvent défini comme légèrement supérieur à M surtout lorsque les populations sont supposées être génétiquement différenciées. Le script compare deux à deux les locus entre individus afin de définir si ceux-ci correspondent à un seul et même locus sur la base de ce paramètre. La commande utilisée pour l'utilisation de C-stacks est la suivante :

cstacks -b X -s A1 -s A2 ... -s An -o B -g -p N après p-stacks cstacks -b X -s A1 -s A2 ... -s An -o B –n n -p N après u-stacks avec : A1 -> An - nom des fichiers des catalogues individuels

B – nom du fichier de sortie

-b : batch ID ou numéro assigné à ce catalogue qui sera conservé par la suite

-g : pour génome, indique de la comparaison des locus est faite sur des données de 'mapping'

-p : permet une exécution du calcul en parallèle sur N cœurs du Cluster

-n : nombre de différences (n) maximum toléré pour considérer deux locus provenant de deux individus distincts comme appartenant à un même locus (n=10 pour *H. fuligineum*)

-Calcul des fréquences des SNPs entre populations : module Populations de Stacks

Dans un premier temps, un fichier des populations est créé, il s'agit d'un fichier texte contenant la liste de tous les individus avec leur appartenance à une population. Deux paramètres sont à déterminer au préalable pour que le programme puisse analyser les catalogues. Ces paramètres permettent de déterminer quels locus seront pris en compte dans l'analyse. Le premier paramètre est le nombre minimum (p) de populations dans lequel on veut que le locus soit présent pour être utilisé. En général, p est égal au nombre total de populations défini dans le fichier texte des populations. Le second paramètre détermine la fraction minimale (r) d'individus d'une population chez lesquels on retrouve un locus commun. Ce paramètre est très important car il définit la proportion de données manquantes dans le jeu de données qui sera utilisé par la suite pour réaliser les analyses génétiques. En général, il ne doit pas descendre en dessous de 80%. Le programme va aussi supprimer du catalogue populationnel les locus pour lesquels plus de 2 allèles sont trouvés chez au moins un individu.

Une fois les filtres effectués, le programme va calculer les fréquences des allèles à tous les SNPs/locus sélectionnés et différentes statistiques populationnelles telles que l'hétérozygotie observée et attendue, la diversité nucléotidique π , l'index de consanguinité Fis à chaque SNP, l'index de différenciation Fst entre paires de populations, etc ...

Ce programme propose en plus l'édition de différents formats de sortie des fréquences alléliques pour permettre l'utilisation d'autres programmes de génétique des populations : Structure, Genepop, etc. La commande utilisée pour ce programme est la suivante :

populations -b X -P A -M B -r C -p D -structure -write_random_snp -W C

avec : A - l'emplacement de l'ensemble des fichiers des programmes U ou Pstacks, C et S stacks

B - le nom du fichier texte des populations

C-liste (whitelist) restrictive de locus pour l'analyse

-b : batch ID

-r : fraction minimale d'individus porteur du locus au sein de chaque population

-p : nombre minimum de populations dans lesquelles le locus doit être présent pour être utilisé

--structure : exemple de format de sortie pouvant être demandé (--genepop ; --vcf ...) --write_random_snp : permet de ne conserver qu'un SNP par locus

Remarque : Si aucun fichier de populations et aucun paramètre populationnel ne sont donnés, le script « populations » génère des fichiers contenant l'ensemble des locus présents chez chaque individu sans exigence quant au partage de ces locus par plus d'un individu. Ce genre de résultats sera utilisé par la suite pour générer des tableaux de présence/absence des locus. Remarque 2 : l'option « write_random_snp » a été utilisée dans l'ensemble des catalogues de ce manuscrit à l'exception du cas d'*A. caudata* en raison d'une non-reproductibilité des résultats.

C - Structure génétique des populations

-Distance de Jaccard sur la base des locus communs

En raison du faible nombre de locus communs identifiés entre les populations du sud et du nord de l'EPR chez les espèces hydrothermales puis chez l'espèce antarctique *H. fuligineum* nous a conduit à exploiter l'information présence/absence des locus entre individus de même espèce. En effet, nos données ont révélé l'existence d'un très grand nombre de locus à l'échelle de l'individu, laissant supposer que le faible nombre de locus communs pouvait être dû à une trop forte divergence inter-allélique au sein de ces espèces (i.e. présence d'espèces cryptiques). La distance de Jaccard a initialement été définie pour comparer la composition floristique de différentes localités dans le Jura et les Alpes (Jaccard, 1901). L'indice de Jaccard S servant à calculer la distance de Jaccard est basé sur le rapport entre le nombre de taxons communs aux deux ensembles et le nombre total de taxons trouvés dans les 2 ensembles. La distance de Jaccard estime la dissimilarité des flores/faunes entre communautés en étant égal à 1-S. En génétique, cette méthode de comparaison a aussi notamment été utilisée dans le cadre de

données non-codominantes de type AFLP (Polymorphisme de longueur des fragments amplifies) (Wong *et al.*, 2001) ou RFLP (Polymorphisme de longueur des fragments de restriction) (Thormann *et al.*, 1994).

Le fichier de base qui sert au calcul de la matrice de distances est obtenu en utilisant le script populations de Stacks cité précédemment sans indiquer d'appartenance des individus à une population et, donc, sans appliquer aux locus les filtres se référant à cette information. Un fichier .vcf présentant la fréquence des allèles pour chaque locus à chaque individu est généré. Un simple traitement du fichier est réalisé afin de remplacer les données manquantes par des 0 et les autres par des 1 afin d'obtenir une matrice de présence absence des allèles pour tous les locus.

Une matrice de distances Jaccard est ensuite calculée et les individus sont ordonnés à partir de cette matrice dans un espace à 3 dimensions (Multidimensional scaling) à l'aide du script metaMDS sous R (Faith *et al.*, 1987) en utilisant la commande suivante :

B=metaMDS(A, B, distance=''jaccard'',k=3,try=100,binary=TRUE)

avec : A - matrice des locus en présence/absence

B – nom de la matrice de distances avec l'option distance= 'jaccard'

-k : nombre de dimensions dans lesquelles sont projetés les individus selon leurs distances de Jaccard avec les autres

-try : nombre de tests réalisés permettant de choisir l'estimation des distances dans les dimensions générant le moins d'écart entre les distances multidimensionnelles et celles traduites par la représentation

-binary : précise au script que l'absence d'un locus au sein d'un individu est codé par 0 et qu'une similarité doit se baser sur une co-présence et non une co-absence

Le contenu de la matrice est visualisé à partir de la variable **B\$points**. Ceci permet de créer un tableau dans lequel la première colonne contient le nom de chaque individu et les trois suivantes les coordonnées relatives des individus le long des 3 axes de la MDS. Une cinquième colonne est ensuite ajoutée manuellement pour définir la couleur des points (individus) en fonction de la population d'origine. La représentation graphique des relations entre individus est obtenue avec la commande qui suit :

s3d=scatterplot3d(Bter, type="p", axis=T, tick.marks=T, label.tick.marks=T, color=Bbis [,5], pch=19)

legend("topleft", bty="n", title="Species", legend=c("pop1","pop2",...,("popn"), fill=c("blue","green",...,"red"))

Avec : Bter – matrice de distances à laquelle a été soustraite la première colonne des identifiants Bbis – colonne contenant l'information population sous forme d'un code couleur -les autres options (type, tick.marks, label.tick.marks, topleft, bty, legend) définissent la taille et le type de points utilisés, le positionnement de la légende, etc...

-Analyse factorielle des correspondances (AFC) sur fréquences alléliques (Genetix v.4.01)

Une AFC (Benzécri, 1973) consiste dans une projection d'un grand nombre d'objets (individus/populations) dans un espace à 2 ou 3 dimensions à partir d'un ensemble de variables qualitatives (données génétiques) les caractérisant. Dans un premier temps, l'ensemble des données (i.e. fréquences génotypiques ou alléliques) est utilisé afin de représenter dans un espace à n (n dépendant du nombre de variables/locus considérés) dimensions un nuage de points dans lequel les coordonnées de chaque observation ou individus dépend directement de la valeur des variables présentes dans le tableau initial.

Dans un second temps, l'analyse détermine dans cet espace à n dimensions par le biais de régressions successives, l'axe qui capture la plus grande variance existante dans le nuage de points ainsi que la contribution relative et absolue de chaque locus et individu dans l'inertie de cet axe. Une fois cet axe identifié (valeur propre la plus élevée), l'analyse va chercher un deuxième axe perpendiculaire (définissant un plan avec le premier axe) permettant en complétant les informations déjà capturées par le premier axe de retranscrire le plus fidèlement possible les distances réelles entre les points (individus) dans notre nuage à n dimensions. Et l'analyse continue ainsi de suite jusqu'à avoir défini les axes correspondant aux n-1 variables étudiées.

La représentation 3D obtenue sur les 3 axes les plus explicatifs (% de variance expliquée) permet de condenser l'information pour étudier visuellement cette grande quantité de données et, de détecter des groupements d'individus ou des points aberrants. L'analyse est cependant extrêmement sensible aux données manquantes. Dans le cas de données d'un génotypage d'une espèce diploïde sur n locus, il faut prendre en compte que le génotype identifié à un individu et un site polymorphe donné traduit une double information, l'identité des deux allèles de l'individu. Cette particularité est prise en compte par le programme GENETIX (Belkhir *et al.*,

1996, 2004) qui permet de réaliser ce genre d'analyse à partir de fichiers reformatés à partir des fichiers .genepop fournis par le module populations de Stacks.

Genetix V.4.01 a été créé pour traiter des données microsatellite/allozyme, il ne peut pas analyser des jeux de données avec beaucoup de locus. Pour cette raison, dans les cas où le nombre de locus obtenus était supérieur à 1000 une sélection aléatoire des locus a été effectuée parmi les locus obtenus à l'aide de la commande R suivante :

B = A[,(sample(3:(ncol(A)), 1000))]

C = A[,1:2]

$\mathbf{D} = \mathbf{cbind}(\mathbf{C}, \mathbf{B})$

Avec : A – matrice initiale des fréquences contenant l'identité des individus pour chaque locus

- B matrice réduite à 1000 locus choisis aléatoirement avec la fonction 'sample'
- C colonnes des individus et de leur population d'appartenance
- D matrice finale limitée à 1000 locus aléatoirement choisis

-Classification des individus selon leur appartenance à des entités génétiques définies sur la base d'inférences Bayésiennes : Structure

Le programme Structure (Pritchard et al., 2000) utilise des données génotypiques multi-locus pour assigner les individus échantillonnés à des entités populationnelles définies a posteriori par une approche Bayésienne. Ce programme permet entre autres d'identifier des populations sur la base de leur structure génétique et non géographique en se basant sur l'équilibre de Hardy-Weinberg, et d'affecter une certaine proportion du génome de chaque individu à ces populations en fonction de leur génotype multilocus. Cette analyse est particulièrement pertinente pour l'étude des zones d'hybridation, l'identification de migrants de première ou seconde génération et/ou le mélange d'individus génétiquement différenciés au sein d'une même aire géographique. Le logiciel utilise la plupart des marqueurs génétiques couramment utilisés, y compris les SNPs, les microsatellites, les données RFLP et AFLP. Les fichiers nécessaires à son utilisation sont générés par le script populations de Stacks. L'analyse nécessite le paramétrage de plusieurs variables dont le nombre de classes (entités génétiques) K à identifier, le nombre de simulations nécessaire avant d'atteindre un état d'équilibre (burn -in) et le nombre de simulations utilisé dans le calcul et son pas d'échantillonnage. Pour l'ensemble de nos analyses, 10 000 simulations initiales ont été réalisées avant récolte des données, et 100 000 simulations supplémentaires ont été réalisées pour estimer les fréquences alléliques dans les 'groupes' génétiques et l'appartenance de chaque individu à chaque groupe. Des simulations avec des valeurs croissantes de K (1 à 6) ont été effectuées, et seul celle pour laquelle la valeur de Log(vraisemblance) était la meilleure a été retenu selon la méthode établie par Evanno et al. (2005).

-Assignation des individus à des populations selon leur génotype multilocus : GeneClass 2.0

Il est aussi possible de calculer des probabilités d'appartenance des individus à des populations prédéfinies sur la base de données génotypiques multi-locus à l'aide du programme GeneClass2 (Piry *et al.*, 2004) en partant d'un fichier au format .genepop sur la base des données de fréquences alléliques multi-locus. A partir d'un échantillonnage aléatoire des allèles dans chaque population, simule des génotypes multi-locus parentaux (1000 génotypes dans notre cas) et détermine ainsi une valeur seuil d'identité des individus à la population de référence à partir de ces génotypes simulés. Il va ensuite en se basant sur l'identité allélique des individus à tester, calculer une probabilité d'assignement sur la fonction de vraisemblance qui sera considérée comme significative si elle est supérieure à la valeur seuil obtenue pour cette population.

4 - Analyse de la diversité génétique des espèces

A - Estimation des diversités synonyme et non-synonyme : Reads2snp

Dans le cas des études visant à comparer la diversité génétique entre espèces proches mais vivant dans des environnements différents, une partie des données ddRAD de plusieurs individus de chaque espèce ont été assignées à des transcriptomes de référence réduits et filtrés afin d'identifier le polymorphisme synonyme et non synonyme au sein de chaque espèce. En effet même s'il existe des biais d'utilisation de certains codons (Akashi, 1995), les mutations synonymes sont pour la plupart insensibles à l'effet de la sélection alors que les mutations non-synonymes (remplacement d'acides aminés ou insertion de codons stop) peuvent être soumises à une forte pression de sélection, ainsi la comparaison des taux de fixation de ces deux types de polymorphisme constitue un bon moyen de décrypter les mécanismes influençant l'évolution

des séquences considérées (Yang & Nielsen, 2000). On s'attend ainsi par exemple dans le cas du modèle neutre à observer le même ratio piN/piS entre des lignées proches (Yang & Nielsen, 1998), ou à une augmentation de ce ratio dans le cas d'un relâchement de la sélection sur la fonction d'une protéine (Crandall & Hillis, 1997).

Pour ce faire, le script reads2snp développé par l'équipe de Nicolas Galtier (ISEM) permet à partir de données Illumina de type RNAseq issues de plusieurs d'individus d'une même espèce et d'espèces proches d'identifier la diversité nucléotidique d'une espèce à l'échelle du génome à partir des SNPs (« single nucleotide polymorphism » en anglais) trouvés dans les séquences (Gayral *et al.*, 2013 ; Yan *et al.*, 2017).

La commande utilisée pour ce programme est la suivante :

Reads2snp -bamlist A -bamref B

avec : A – un fichier texte contenant une liste de fichiers BAM et la correspondance de ces fichiers avec les individus utilisés

B – le génome ou transcriptome de référence utilisé pour la cartographie des lectures ayant fourni les fichiers BAM utilisés

Après avoir obtenu les données de polymorphisme, un second script dNdSpiNpiS appartenant aussi au pipeline développé par l'équipe de Nicolas Galtier permet de le catégoriser en polymorphisme synonyme ou non-synonyme en lui fournissant l'information sur le positionnement des SNPs dans les séquences codantes. Il est aussi possible de fournir à ce script des séquences provenant d'un 'outgroup' (espèce proche) afin d'estimer à la fois les piN et piS et les dN et dS, mais cette option n'est pas utilisée ici en raison de la trop grande divergence entre les espèces dont nous disposons. Dans ces analyses nous avons paramétré certaines variables pour réaliser le calcul des diversités uniquement sur les locus les plus informatifs (i.e. présents chez un large nombre d'individus). Les lectures obtenues étant cartographiées sur le transcriptome, le script filtre par défaut les informations provenant de séquences de référence (transcrits) que si celles-ci sont couvertes par les lectures à au moins 50% de leur longueur totale. Dans notre cas, les lectures proviennent de ddRAD et, donc constituent des locus de seulement 120 paires de bases très éparpillées sur le transcriptome. Nous avons donc reparamétré cette valeur à 0.001% de la couverture exigée. En revanche, nous avons défini le calcul des piN et piS sur une longueur minimale de 30 codons ce qui correspond à 90 paires de bases. Enfin, seuls les sites polymorphes retrouvés chez au moins la moitié des individus est prise en compte dans un premier temps.

La commande utilisée pour ce programme est la suivante :

dNdSpiNpiS -alignment_file=A -ingroup=sp -gapN_seq=0.99999 -min_nb_codon=30 gapN_site=10 -out=my_res

avec : A – nom du fichier de sortie de reads2snp contenant l'information sur le polymorphisme

B – sp est la façon dont reads2snp nomme l'espèce utilisée par défaut, cette option à remplir obligatoirement sert principalement à différencier l'espèce d'intérêt de l'espèce 'outgroup' lorsqu'une deuxième espèce est utilisée pour calculer les dN et dS.

-gapN_seq : proportion maximale de non-couverture tolérée le long des séquences de référence (transcrits)

-min_nb_codon : nombre minimum de codons exigé pour prendre en compte une séquence dans le calcul du piN et piS

-gapN_site : nombre minimum de copies (allèles) exigé pour un locus pour prendre en compte une séquence dans le calcul, la valeur maximale que peut prendre ce paramètre est 2 fois le nombre d'individus puisque que le script prend les deux allèles de chaque individu en compte

Remarque : Le nombre exigé et nécessaire de séquences pour une estimation au plus près de la réalité sera testé par la suite sous la forme de courbes de raréfaction en fonction du nombre d'individus analysés.

B - Calcul d'indices de diversité et tests de neutralité : Genepop/DNAsp

Les logiciels Genepop et DNAsp peuvent être utilisés en mode 'batch' sur un très grand nombre de locus et sont donc particulièrement adaptés pour regarder la distribution en fréquence de certains estimateurs du polymorphisme ou de certains tests de neutralité à l'échelle d'un génome. Le logiciel Genepop (Rousset, 1995, 2008) réalise locus par locus des tests exacts de Fisher pour tester des écarts à l'équilibre de Hardy-Weinberg et la différenciation génétique entre populations. Il permet également de calculer différents indices de différenciation (Fis et Fst) ainsi les déséquilibres de liaison entre locus. Le logiciel DNAsp (Librado & Rozas, 2009) permet quant à lui d'estimer la taille efficace des populations et le flux de gènes entre

populations et permet d'effectuer différents tests d'écart à la neutralité dans l'accumulation des mutations dans le polymorphisme (Fu et Li (1993), Kreitman et Aguadé(Hudson *et al.*, 1987), Tajima (1989), McDonald et Kreitman (1991), etc).

Ces deux programmes utilisent des fichiers au format .genepop ou des alignements au format .fasta qui font partie des sorties qu'il est possible de générer respectivement avec le script populations de Stacks ou dNdSpiNpiS.

5 - Identification des locus « outliers » (ou aberrants) sous sélection positive avec différents modèles de populations : Lositan & DetSel

Il est possible d'effectuer un criblage des locus ayant un comportement que l'on pourrait qualifier d'aberrant du point de vue de la différenciation génétique entre des populations choisies sous l'hypothèse neutre de la migration/dérive. En effet, en connaissant les fréquences alléliques à un locus donné, on peut connaître de quelle façon ce polymorphisme se distribue dans les populations selon le modèle de population utilisé en prenant ou non en compte les phénomènes démographiques. En utilisant l'information de la différenciation multi-locus et en se basant sur des hypothèses prédéfinies, il est possible d'estimer un intervalle dans lequel l'indice de différenciation entre populations peut varier et hors duquel les locus seront considérés comme ayant un comportement aberrant du point de vue de l'attendu migration/dérive. On parle alors de locus « outliers », dont le comportement ne peut s'expliquer qu'à travers l'action de la sélection naturelle.

Deux méthodes différentes ont été utilisées ici. La méthode de détection des Fst outliers développée par Beaumont & Nichols (1996) sur la base de la relation entre différenciation génétique attendue et hétérozygotie est implémentée dans le logiciel Lositan (Antao *et al.,* 2008). Ce modèle suppose que lorsqu'il y a de la sélection sur des locus en particuliers, leur nombre doit être assez faible et ne pas impacter la tendance générale observée au sein du jeu de données, ce qui permet de les identifier en raison de leur comportement exagérément différenciant au vu de leur hétérozygotie. Ainsi ce modèle est d'autant plus efficace que le nombre de locus utilisés est grand. Ce programme détermine une enveloppe de confiance autour des Fst produits par un grand nombre de locus en se basant sur un modèle en îles, c'est-à-dire dans lequel toutes les populations échangent un flux de gènes équivalent. DetSel (Vitalis, 2012) est un package utilisable sous R qui lui suppose un début de spéciation avérée entre deux sous-

populations et donc un flux de gènes réduit voire absent entre les populations. Pour calculer son enveloppe de confiance autour des valeurs de Fst multi-locus, DetSel simule un goulot d'étranglement qui précède la séparation des deux populations soeurs. Pour réaliser cette simulation il faut avoir une idée a priori sur plusieurs paramètres populationnels comme les tailles efficaces des populations actuelles ainsi que celle avant et au moment du goulot d'étranglement respectivement nommées Ne et No, ainsi que les temps écoulés en générations depuis le début du goulot d'étranglement (to) et après le début de l'isolement des populations (t) (Figure II-2).



Figure II-13 : Représentation du modèle d'isolement de deux populations, utilisé par DetSel pour détecter des locus « outliers » sous sélection positive avec les valeurs d'exemple pour les différents paramètres utilisés. A droite, la phylogénie associée obtenue à partir de 20 gènes et utilisée dans l'exemple. Traduit de Vitalis 2012.

6 - Bilan





Figure II-14 : Représentation des enchaînements de traitements de données et analyses bioinformatiques (encadrés) réalisés et des types de données obtenues au cours de cette thèse.

III - Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processus de spéciation chez les espèces du genre *Alvinella*

1 - Introduction

La perte et/ou la fragmentation d'habitat associées à la conversion des terres pour les activités humaines constituent la menace la plus grave pour la diversité biologique en milieu terrestre (Collinge, 1996). Cette utilisation des espaces génère une fragmentation des environnements qui, d'un point de vue écologique, perturbe les processus biologiques et physiques qui gouvernent le fonctionnement de l'écosystème (Young *et al.*, 1996). En effet, la fragmentation de l'habitat induit une baisse de connectivité entre les populations, une réduction de la taille efficace de celles-ci et par la même occasion une diminution de la diversité génétique des espèces (Johansson *et al.*, 2007).

Avant même que l'altération anthropique des systèmes naturels ne devienne un problème majeur dans l'érosion de la biodiversité marine et terrestre, les chercheurs se sont intéressés au rôle de la migration et de la dérive génétique sur les structures génétiques des espèces à travers la mise en place de modèles de populations allant d'un continuum d'individus sur une aire géographique donnée (Rousset, 1997) à des populations discrètes d'individus échangeant plus ou moins entre elles (Takahata 1983). En général, la différenciation génétique entre populations locales dépend d'un état d'équilibre entre la dérive génétique, qui tend à différencier les populations avec le temps et, la migration qui s'oppose à cette différenciation. Dans le cadre du modèle en îles, le degré de différenciation des populations dépend donc principalement du taux de migration (identique entre toutes les populations) et, de la taille efficace des populations (l'effet de dérive étant conditionné par celle-ci). Dans ce modèle, la structure génétique observée ne dépend pas de la distribution géographique des populations. A l'inverse, le modèle théorique en pas japonais (Stepping stones model : Kimura & Weiss, 1964) a été proposé afin de prendre en compte la variabilité géographique du taux de migration entre populations (les populations n'échangent qu'avec leurs plus proches voisins) et son effet sur la diversité génétique de chaque population. Ce modèle de flux de gènes en pas japonais se base sur le principe d'isolement par la distance (Wright, 1943) appliqué à des populations fragmentées (Figure III-1). Ces modèles théoriques peuvent également en prenant en compte les processus d'extinction/recolonisation dans les populations s'intéresser alors à l'origine des individus qui recolonisent les nouveaux habitats (Whitlock & McCauley 1990, McCauley *et al.* 1991). D'autres modèles basés sur le modèle en îles et ne prenant pas en compte la dérive, mais comparant l'effet relatif d'un flux de gènes asymétrique à celui d'une barrière géographique (flux de gènes symétriquement limité dans ce cas) ont quant à eux permis de caractériser le type de rupture de fréquence allélique que l'on observerait dans l'une et l'autre des situations (Nagylaki, 1976). En revanche, ces modèles déterministes basés sur l'équilibre migration/dérive ne prennent pas en compte l'effet de la sélection naturelle sur la distribution de la diversité génétique au sein et entre ces populations.



Figure III-15: Représentation du modèle en Stepping stones à une et deux dimensions (Kimura & Weiss, 1964)

Ainsi un autre équilibre à prendre en compte est celui qui existe entre sélection et migration. C'est ainsi que Barton en 1979 a pris en compte l'impact qu'avait le taux de recombinaison sur le polymorphisme neutre lié génétiquement à un locus sous sélection (Figure III-2). Dans nombre de cas, la différenciation génétique des populations à une échelle locale n'est malheureusement pas prédite par la distribution géographique des individus. De nombreux processus micro-évolutifs peuvent expliquer les patrons de différenciation génétique d'une espèce alors même que ces populations sont reliées entre elles par des flux de gènes. On peut citer, par exemple, la mise en place de contacts secondaires entre populations avec de l'hybridation locale entre formes parentales (Muths *et al.* 2010), le succès de reproduction variable entre individus (sweepstakes reproductive success : Hedgecock & Pudovkin 2011), la dispersion collective de larves et leur recrutement simultané en certaines localités (Broquet *et al.* 2013) ou le choix d'habitat par les larves/post-larves (Bierne *et al.* 2003), ou encore la sélection diversifiante des individus/larves par l'habitat (David & Jarne 1997, *cf.* review Eldon

et al. 2016). En termes d'hétérogénéité génétique à micro-échelle spatiale, de nombreux modèles ont cherché à déterminer l'effet de la sélection diversifiante au sein d'une population. Il a ainsi été montré que dans des conditions particulières où un environnement présente deux niches induisant des pressions de sélection différentes, un polymorphisme stable peut être maintenu dans le temps (Levene, 1953). Et si on prend en compte un cas d'enzyme avec additivité des propriétés (Gillespie, 1977), il est alors possible de prédire les proportions et l'intensité de la sélection dans chaque niche nécessaire au maintien d'un polymorphisme stable dans le temps (Smith & Hoekstra, 1980).



Figure III-16 : Représentation des clines de fréquences neutres dont l'intensité de la rupture de courbe est corrélée à celle d'un locus lié, avec X la distance à la barrière, t le temps et u la fréquence de l'allèle neutre considéré (Barton, 1979).

Dans le cas où ces habitats sont isolés l'un de l'autre, il a été montré que la simple dérive sur du polymorphisme neutre peut aussi amener à la différenciation des populations simplement par le biais du tri aléatoire fait au sein de chacune et maintenant alternativement telle ou telle version de chaque gène (Takahata, 1983). L'existence de différentes niches écologiques (et/ou habitats) dans le temps où l'espace est donc un moyen de maintenir du polymorphisme dans les populations lorsque l'on applique un coefficient de sélection différent selon les différents génotypes en présence et même entre populations par simple effet de dérive. Chaque niche se caractérise par des pressions de sélection différentes, et la proportion relative de chaque niche dans le temps où l'espace est un paramètre à prendre en compte dans le maintien ou non des différentes formes alléliques au sein de la population. Si ces deux niches sont séparées dans l'espace, les différents allèles seront donc sélectionnés localement ce qui implique un différentiel des fréquences alléliques entre les habitats pour une population donnée (Hedrick,

1985). Mais des différences de fréquences alléliques entre populations peuvent aussi résulter d'un contact entre des populations génétiquement différenciées en générant des patrons intermédiaires le long d'un cline géographique suivant l'axe de connectivité de ces populations. Ces clines peuvent alors s'aligner le long d'un gradient environnemental entre les populations sources et mimer un cline d'adaptation.

Enfin à l'inverse, la migration en mettant en contact des fonds génétiques différents peut générer de nouvelles combinaisons alléliques dans la zone de contact. La dérive et la sélection opérant localement au niveau de ces zones de mélange vont alors pouvoir permettre l'augmentation en fréquence de certaines combinaisons d'allèles, mais également conduire à la disparition de certaines formes alléliques mal-adaptées ou en fréquence trop faible dans la population. Ainsi l'établissement et le maintien de différenciations génétiques adaptatives au sein d'espèces résulte d'un équilibre entre migration, sélection et dérive (Yeaman & Otto, 2011) dans un environnement le plus souvent hétérogène. L'enjeu est donc pour une situation donnée de réussir à distinguer l'effet de ces différentes forces évolutives. Ce n'est qu'en distinguant les effets respectifs de ces trois forces que l'on peut expliquer les patrons de différenciation génétique observés à un instant donné dans un environnement donné.

La fragmentation de l'environnement n'est pas nécessairement le résultat de l'activité humaine. De nombreux environnements sont naturellement fragmentés et constituent un véritable laboratoire à ciel ouvert pour l'étude des mécanismes de dispersion ainsi que les effets des barrières à la migration (distances entre habitats) sur les échanges entre populations. Ces habitats fragmentés sont souvent associés au milieu terrestre avec les archipels, les oasis en milieu désertique (Bradford *et al.*, 2003), les réseaux fluviatiles et lacustres (Rieman & McIntyre, 1995), l'environnement montagnard (Bleich *et al.*, 1990) mais, également en milieu marin avec les lagunes, les baies envasées ou encore, les milieux réduits associés aux sources hydrothermales, aux zones de suintements froids ou aux carcasses de cétacés (Smith, 1989). Dans ces environnements, il est alors possible d'estimer les contributions respectives de l'isolement géographique et de l'adaptation locale dans le partitionnement de la diversité génétique des espèces.

L'existence des sources hydrothermales résulte de la dynamique tectonique de la lithosphère entraînant la formation de la nouvelle croûte océanique à l'axe des dorsales. Elle est directement dépendante de la profondeur de la chambre magmatique sous la croûte océanique et du niveau de fissuration du plancher océanique à l'axe de la dorsale. Pour cette raison, les zones d'hydrothermalisme se distribuent de façon agrégative le long des dorsales et constitue donc un bon modèle de dispersion uni-dimensionnel (Figure III-2) (Chevaldonné *et al.*, 1997). La

distance séparant deux sources hydrothermales peut aller de quelques dizaines de mètres à plusieurs centaines de kilomètres (Plouviez et al., 2009). Les émissions de fluide hydrothermal ont une concentration élevée en divers composés réduits et en sulfures avec des enrichissements métalliques et de fortes concentrations en CO₂, H₂ voire en méthane constituant un environnement particulièrement sélectif vis-à-vis de l'installation de la faune profonde (Tunnicliffe et al., 1998). Les espèces inféodées à cet environnement présentent donc une longue histoire évolutive d'adaptations physiologiques et constituent une faune endémique très spécialisée et peu diversifiée (Tunnicliffe 1991). Cet environnement est en plus sujet à de fortes variations thermiques. Des mesures réalisées en différents points d'une cheminée, y compris à l'intérieur de la cheminée et sur la paroi de celle-ci, ont montré des variations de l'ordre de plus de 100°C au sein des conduits de cheminée sur 10 jours et de 15°C en l'espace de 10 min à la surface de la cheminée (Tivey et al., 2002). D'autres mesures au pied de cheminées ont montré des températures oscillant entre 2 et 23°C au cours de la journée (Chevaldonné et al., 1991). Au niveau des moulières qui se situent en général en périphérie des émissions hydrothermales, ce sont des températures plus basses mais variant tout de même entre 2 et 12°C qui ont été observées (Mullineaux et al., 2003). Enfin des suivis de température à la surface de colonies d'Alvinella pompejana, qui construit son tube sur les parois des cheminées hydrothermales, ont montré une amplitude de température de plus de 10°C en 1h (Chevaldonné et al., 1991 ; Le Bris & Gaill, 2007) et des variations spatio-temporelles comprises entre 5 et 50°C.

En plus de la nature agrégative de l'environnement hydrothermal et de son extrême variabilité dans le temps, les émissions ont une durée de vie limitée. En effet, l'expansion continue de la croûte océanique, la dynamique intrinsèque des flux de chaleur entre la chambre magmatique et la croûte océanique et les remaniements tectoniques récurrents le long des dorsales génèrent de nouveaux réseaux de failles favorisant la circulation hydrothermale mais provoquent aussi l'extinction de sites préexistants (Watremez *et al.* 1990, Jollivet *et al.* 1999). L'arrêt des émissions au niveau d'une cheminée peut aussi survenir par précipitation des sulfures polymétalliques à l'intérieur du conduit provoquant le colmatage de celui-ci (Haymon & Macdonald, 1985). La durée de vie d'un site hydrothermal peut aller de quelques années à plusieurs dizaines d'années mais les champs regroupant de nombreuses émissions sur plusieurs kilomètres sont quant à eux relativement pérennes sur plusieurs centaines, voire milliers d'années selon le taux d'accrétion des dorsales (Shank *et al.*, 1998).

La nature éphémère et fragmentée de l'environnement hydrothermal a très vite généré un fort intérêt dans le monde scientifique qui a cherché à comprendre comment la faune très spécialisée et endémique de cet environnement pouvait se disperser et coloniser de nouveaux sites (Lutz *et*

al., 1984). On s'attend en effet dans ce genre d'environnement à rencontrer des espèces ayant une croissance rapide, une maturité précoce et de bonnes capacités de dispersion pour recoloniser très vite les zones d'émissions nouvelles (Lutz, 1988 ; Vrijenhoek, 1997). Une étude sur plusieurs espèces de bivalves, mollusques et annélides a pu montrer un isolement par la distance entre les populations hydrothermales du Pacifique Oriental (EPR) (Vrijenhoek et al., 1998). Il semble cependant que cet isolement par la distance soit principalement dû à la présence de barrières physiques à la dispersion comme certaines failles transformantes ou zones de moindre activité hydrothermale (Plouviez et al. 2009). Ces études plus récentes basées soit sur la composition faunistique des sites le long de l'EPR (Matabos et al., 2011) soit sur le séquençage du gène mitochondrial Cox-1 chez plusieurs espèces emblématiques de l'EPR (Won et al. 2003, Hurtado et al. 2004, Plouviez et al., 2009, 2010, 2013) ont en effet montré l'existence de barrières aux flux de gènes, l'une au niveau de l'équateur et l'autre au niveau de la microplaque de l'île de Pâques. Ces barrières sont séparées entre elles par des zones de transition où les lignées géographiques sont remises en contact et peuvent s'hybrider localement (Matabos et al. 2008, Plouviez et al. 2010, 2013) (Figure III-3). Cette différenciation nord/sud à travers la barrière équatoriale est particulièrement marquée chez l'annélide Alvinella pompejana qui occupe la partie la plus chaude des cheminées hydrothermales (Desbruyères et al. 1998). L'étude menée par Plouviez et al. (2010) montre en effet que certains locus nucléaires comme la PGM-1 sont fortement différenciés de part et d'autre de la barrière avec des allèles distincts et divergents laissant penser à un isolement en allopatrie des populations de cette espèce.

Alvinella pompejana, ou ver de Pompéi, (Desbruyères & Laubier, 1980) est l'une des espèces emblématiques du milieu hydrothermal. En effet, le nom de cette espèce provient d'une part de celui du bathyscaphe Alvin ayant permis son observation première au niveau des sources hydrothermales de 21°N/EPR (Corliss *et al.* 1980), et *pompejana* fait référence à la ville antique de Pompéi en Italie célèbre pour avoir été recouverte par les productions de cendres d'une éruption du Vésuve. La première description d'*Alvinella pompejana* (Desbruyères & Laubier, 1980) considérait l'espèce comme possédant deux formes ontogéniques successives, l'une juvénile et l'autre mature (forme épitoque) (Figure III-4). Six ans après, les premières études sur le polymorphisme enzymatique de cette espèce ont amené à scinder cette espèce en deux espèces bien différentes sans aucun allèle partagé (Autem *et al.* 1986), et la forme initialement décrite comme juvénile est désormais nommée *Alvinella caudata*. Les espèces du genre Alvinella sont caractérisés par une association avec des bactéries filamenteuses sur leur partie dorsale à l'intérieur de leurs tubes (Desbruyères & Laubier, 1986). Ces animaux présentent de
nombreuses adaptations permettant la vie dans ces eaux chargées en métaux lourds et composés sulfurés. En effet même si son rôle est discuté, l'épibiose permet probablement de réaliser un recyclage du carbone au sein du tube de l'animal, on observe aussi une détoxification des composés sulfurés par oxydation au niveau des branchies et des mécanismes permettant l'excrétion des métaux lourds (Desbruyères *et al.*, 1998).



Figure III-3: Dorsale du Pacifique oriental (EPR: East Pacifique Rise) avec la position des sites hydrothermaux échantillonnés à 13°N/EPR et à 9°50N lors de la campagne PHARE 2002 et Mescal 2010/2012 ainsi que les champs hydrothermaux échantillonnés dans la



Figure III-4: *Alvinella pompejana* (à gauche) et *Alvinella caudata* (à droite) présentées comme les formes matures et juvéniles d'*Alvinella pompejana* (Desbruyères & Laubier, 1980) et, révisé par la suite par Autem et al. (1986) comme des espèces à part entière.

Parce qu'elles font partie des premières espèces décrites et en raison d'un haut niveau de variation génétique ces espèces sont rapidement devenues un sujet d'étude privilégié pour

comprendre la dynamique de dispersion et recolonisation des sites dans cet environnement. Peu d'informations sont disponibles quant aux traits d'histoire de vie de ces espèces (Tunnicliffe, 1998). La possibilité pour ces polychètes térébellomorphes de sortir de leur tube pour s'accoupler tête bêche et la morphologie particulière de leurs spermatozoïdes aux flagelles atrophiés (Jouin-Toulmond et al. 2002) suggèrent une possible fécondation interne ou tout du moins à l'intérieur du tube des femelles (Storch & Gaill, 1986 ; Chevaldonné & Jollivet, 1993). De plus, la taille des ovocytes matures de ces deux espèces (autour de 200 µm) laisse penser à un développement lécithotrophe des larves menant à des larves erpochètes comme chez Paralvinella grasslei (Chevaldonné et al., 1997, Pradillon et al. 2002, Faure et al. 2007) ce qui favoriserait ainsi le recrutement local. La reproduction semble asynchrone mais les observations in situ montrent que les œufs ont une flottabilité négative (Pradillon, 2002) et sont émis dans la colonne d'eau après fécondation avec une synchronisation des pontes entre femelles à partir des ovocytes matures accumulés dans les oviductes (Faure et al. 2007). Ainsi ce mode de reproduction cumulé à une forte propension à la rétention de la descendance devrait conduire à un isolement génétique des populations avec la distance géographique : le temps de durée de vie larvaire permettant l'échange d'un seul migrant par génération entre 2 sites étant estimé à 8 jours (Chevaldonné et al., 1997). Or, il a été montré au nord de l'EPR que les populations d'Alvinella sont génétiquement homogènes, ce qui suggèrerait :

(1) que malgré les observations histologiques et comportementales les larves peuvent effectivement disperser sur de grandes distances,

(2) que la dynamique spatiale des champs hydrothermaux et, à l'intérieur même de ces champs, celle des sites hydrothermaux eux-mêmes, permet des échanges entre populations lorsque les champs et les sites se rapprochent à la faveur des remaniements tectoniques le long de la dorsale (Jollivet *et al.* 1999).

Cette espèce vit également dans un environnement spatialement hétérogène. C'est d'ailleurs l'une des premières espèces à coloniser les cheminées nouvellement formées (Pradillon *et al.*, 2005b) lorsqu'elles constituent des diffuseurs à anhydrite (sulfate de barium/calcium) particulièrement poreux présentant des températures supérieures à 100°C. A ce titre, le ver de Pompéi est considéré comme l'une des espèces les plus thermophiles du système hydrothermal (Chevaldonné *et al.* 1992, Cary *et al.* 1997, Ravaux *et al.* 2013), entre autres parce qu'en colonisant les parois des cheminées hydrothermales, il est au plus près des émissions hydrothermales non diluées et donc très chaudes (et chargées de composés toxiques). Les conditions thermo-chimiques varient beaucoup entre cheminées selon leur âge, et l'espèce est exposée à une large gamme de températures, les conditions thermo-chimiques d'une cheminée mature à conduits en sphalérite/chalcopyrite dans lesquelles s'établit par la suite la population pour se reproduire étant très différentes de celles d'un diffuseur à anhydrite. Une étude du polymorphisme de la phosphoglucomutase (PGM) a d'ailleurs montré l'existence d'un polymorphisme adaptatif entre colonisateurs et reproducteurs dans laquelle les isoformes enzymatiques n'ont pas la même thermostabilité dans les différentes populations du nord de l'EPR, avec une corrélation positive entre la fréquence de l'une de ces isoformes et la température mesurée à l'entrée des tubes des vers (Piccino *et al.*, 2004). A ce titre, le ver de Pompéi constitue un modèle biologique particulièrement intéressant pour étudier le rôle de la sélection environnementale sur la structure génétique de l'espèce et la part respective des processus de dérive et de sélection dans les patrons de différenciation génétique à l'échelle locale.

Une étude plus poussée du polymorphisme de la phosphoglucomutase d'A. pompejana a permis de caractériser d'une part la divergence entre les différentes isoformes, mais aussi de quantifier l'effet adaptatif des mutations non synonymes responsables du polymorphisme de charge. Ainsi, il a pu être montré par la réalisation d'une généalogie des allèles que les principales isoformes ont été maintenues sans recombinaison sur une très longue période de temps qui prédate la séparation des populations du ver de part et d'autre de la barrière équatoriale. En effet, alors que trois isoformes principales sont identifiées par l'étude du polymorphisme de charge, quatre lignées alléliques très divergentes sont présentes avec deux formes alléliques spécifiques retrouvées au nord et 2 autres retrouvées au sud. Dans cette généalogie, les allèles du sud apparaissent dérivés des allèles nord et le polymorphisme entre allozymes thermosensible et thermorésistant est conservé de part et d'autre de la barrière aux flux de gènes, datée à environ 1-2 millions d'années par Plouviez et al. (2009). Des mesures de stabilité et d'activité sur les protéines recombinantes donnent aussi de sérieuses pistes quant à l'existence d'un polymorphisme 'adaptatif' suggérant que de chaque côté de la barrière, l'isoforme la plus fréquente serait plus active mais aussi plus sensible aux augmentations de la température. Ceci suggère ainsi que le maintien de ce type de polymorphisme puisse suivre un modèle de sélection différentielle à 2 niches tel que proposé par Levene (1953) et Gillespie (1985) en partie fonction de l'habitat thermique rencontré par les vers (cf Hedrick 1987).

Toutes ces observations restent à l'échelle d'un gène et ne permettent pas de prendre en compte l'importance de cette adaptation à la température à l'échelle du génome de l'espèce. Cela pose la question de leur fréquence et de leur localisation dans le génome ainsi que de leur histoire évolutive en termes d'adaptation locale et d'isolement géographique.

Ces polymorphismes sont-ils liés par une histoire commune provenant de l'histoire démographique des populations ou datent-ils de la colonisation de cet environnement hyper-variable ?

A l'inverse ces polymorphismes ont-ils été acquis progressivement et de manière indépendante au cours de l'évolution de cette espèce pour vivre dans le pôle le plus chaud de l'environnement hydrothermal, ce pôle constituant un refuge particulièrement efficace contre la prédation et une position privilégiée pour assurer le fonctionnement optimal de son épibiose (Gaill *et al.*, 1987) ? Les divergences observées entre les allèles nord et sud montrent également que cette espèce est impactée par la barrière équatoriale aux flux de gènes, et montre aussi une adaptation à la température. Mais au-delà de l'exemple, l'évolution de ce gène montre qu'un seul et même gène porte les marques simultanées d'une différenciation géographique et d'une différenciation locale liée à l'habitat. Comme nous l'avons dit précédemment, les espèces du genre *Alvinella* constituent un modèle intéressant pour étudier le rôle de l'adaptation locale dans les processus de différenciation génétique au sein de l'espèce et posent plusieurs questions :

Etant donné le rôle structurant majeur de la barrière équatoriale aux flux de gènes sur la structuration génétique des populations à l'échelle du génome mitochondrial et de quelques gènes nucléaires, quel effet a-t-elle sur les flux de gènes entre les populations nord et sud à l'échelle du génome ?

L'isolement entre localités et les différences thermiques inter-sites jouent-ils un rôle dans la différenciation locale des populations ? Il s'agira ici de savoir si la différenciation géographique entre les populations nord et sud est plus ou moins forte selon le type d'habitat.

Quelles sont les modalités d'action de la sélection diversifiante et comment opère-t-elle sur le génome d'*Alvinella pompejana* et son espèce sœur *A. caudata* ?

Pour répondre à ces différentes questions, nous allons nous intéresser à ces deux espèces sœurs et syntopiques, toutes 2 strictement inféodées au pôle chaud des cheminées hydrothermales du Pacifique Oriental et analyser le polymorphisme de marqueurs échantillonnés aléatoirement le long du génome à travers une approche ddRAD en utilisant un assemblage du génome du ver de Pompéi et les transcriptomes des 2 espèces comme références. Par cette technique il sera alors envisageable d'identifier les gènes divergents entre populations isolées et ceux intervenant dans l'adaptation locale et potentiellement mettre en évidence des redondances (Andrew & Rieseberg, 2013). Il sera aussi possible de mettre en évidence le rôle de la sélection quant aux possibilités d'introgression entre fonds génétiques différents lorsque l'on dispose de plusieurs réplicats environnementaux de mise en contact de ces différents fonds génétiques (Bierne et al., 2003).

2 - Matériel et Méthodes

-Echantillonnage des individus

Les espèces *A. pompejana* et *A. caudata* ont été récoltées sur les mêmes cheminées (espèces syntopiques) le long de l'EPR et les effectifs d'individus ayant servi à la constitution des banques ddRAD sont présentés dans le Tableau III-1. Au final, les individus d'*A. pompejana* proviennent de 5 latitudes différentes et, pour les sites 13°N/EPR et 9°50'N/EPR où les animaux ont été récoltés en fonction de leur habitat thermique, une distinction est faite entre les individus provenant de cheminées chaudes ou froides. Dans le cas d'*Alvinella caudata*, des échantillons n'étaient disponibles que pour trois des cinq sites précédents. La distinction chaud/froid a également été faite pour les individus de 9°50'N de cette espèce. Les échantillonnages sur l'EPR sud ont été effectués lors de la campagne Biospeedo en 2004 sur le N/O L'Atalante, les prélèvements out été réalisés grâce au sous-marin habité Nautile et son bras télémanipulateur. Les prélèvements sur les cheminées de 13°N ont quant à eux été réalisés en 2002 lors de la campagne Phare avec le ROV Victor6000 («Remote Operated Vehicle » en anglais pour véhicule téléguidé) et son navire porteur le N/O L'Atalante. Enfin les échantillons obtenus sur les sites P-Vent et Bio9 de 9°50'N ont été obtenus lors de la campagne Mescal en 2012 avec l'aide du sous-marin habité Nautile et son navire porteur N/O L'Atalante.

espèce	localisation sur l'EPR	effectif	
A'wnella pompejana	12°N	CHD	13
	13 1	FRD	10
	9°N	CHD	22
		FRD	22
		10	
		11	
	21,33°S		22
Alvinella caudata	9°N	CHD	11
		FRD	14
	17°S		12
	21°S		13

Table III-1 : Effectifs des individus ayant servi à la production de banques RAD pour chaque espèce selon la localité et le type d'habitat thermique (EPR nord).

L'ensemble des techniques associées à la construction des banques séquencées, les traitements bioinformatiques des données et l'analyse génétique des données produites pour obtenir les résultats présentés dans cette partie sont décrites dans le chapitre 2.

3 - Résultats

A - Evaluation de la qualité des banques

Après un séquençage Illumina sur deux lignes d'un séquenceur Hiseq 2500 par le service de génomique dédié de l'Université McGill de Québec, les données nous ont été envoyées sous la forme de fichiers issus d'un démultiplexage des lectures par index (12 fichiers compressés contenant les données de séquençage appariées (R1 & R2)) et nettoyées sur la base de leurs scores de qualité (phred scores). Dans un premier temps une vérification du nombre de lectures obtenues par index a été réalisée en fonction du nombre d'individus associés à chaque index afin d'avoir une idée de la couverture des individus et leur proportion relative dans la composition de la banque pour le séquençage (Figure III-5). Ce graphique montre une corrélation positive significative entre le nombre de lectures obtenues et le nombre d'individus à un index donné, indiquant que le mélange des librairies a dans l'ensemble bien été effectué de façon équimolaire à quelques exceptions étant donné que certains points s'écartent tout de même de la tendance générale.



Figure III-5 : Représentation graphique du nombre de lectures obtenues en fonction du nombre d'individus pour les 10 index illumina utilisés.

Par la suite, un démultiplexage des individus au sein des index grâce aux barcodes a été réalisé avec le script process_rad_tag. Différents filtres ont été appliqués pour contrôler la qualité des séquences. Après filtrage, le nombre de lectures retenues pour chaque individu ainsi que le nombre de séquences perdues à chaque filtre nous donne un aperçu de la qualité de nos constructions dans l'absolu (Figure III-6). Ainsi on constate que 67,2% des séquences sont conservées, l'essentiel de la perte de lectures lors du filtrage est dû à des inserts trop petits menant à des séquences contenant une partie de la

séquence des adaptateurs (reads containing adapter sequence) qui représentent 23,9% du total. On peut donc considérer qu'à l'exception d'une sélection de la taille des fragments un peu trop décalée vers les 150 pb lors de la construction des banques ddRAD, les librairies sont assez homogènes avec un nombre de lectures bien proportionné entre individus.

Barcode	Filename	Total	No RadTag	Low Quality	Retained
CTGGTT-ATCACG	BCN1	1552906	49081	343	1020494
AAGATA-ATCACG	BCN2	2380096	42382	706	1590649
ACTTCC-ATCACG	BCN3	2652612	40350	634	1787775
TTACGG-ATCACG	BCN4	5028702	94333	1323	3150841
AACGAA-ATCACG	AAC1	709694	11979	207	619309
ATTCAT-ATCACG	BCN5	258228	18049	70	207092
CCGACC-ATCACG	AAF1	4376362	31085	941	3248215
CATCAA-ATCACG	AAC9	2980548	28940	679	2465591
GCCTGG-ATCACG	AAC10	4519282	41183	963	3480768
TGCTTG-ATCACG	AAC11	1184142	235131	237	892396



Figure III-6 : Exemple de répartition des lectures après filtration pour les différents barcodes et index (individus).

Le nombre total (pour *A. pompejana* et *A. caudata*) de séquences conservées après filtrage est visualisé sous la forme d'un histogramme représentant le degré de couverture en lectures des individus séquencés (Figure III-7). Cette couverture est relativement faible avec une valeur modale d'environ 2 millions de lectures par individu, une grande variance de cette couverture selon les individus et une proportion non-négligeable d'individus très peu couverts (30%).



Figure III-7 : Diagramme de classes de la couverture individuelle en lectures après filtrage.

Lors du mélange des individus pour la construction de la banque, des mesures individuelles de la concentration d'ADN, par QuBit, ont été réalisées afin d'avoir une représentation des individus en quantité similaire, on s'attendait donc à observer une distribution normale le plus resserrée possible autour d'une valeur moyenne. On peut donc voir que dans notre cas il y a un excès d'individus plus et moins couverts qu'attendus à la moyenne.

Bien que le script Process_Rad_Tag permette un démultiplexage des individus en prenant en compte l'appariement des données R1 et R2 (séquences « forward » et « reverse » de part et d'autre du fragment séquencé), les étapes suivantes du logiciel Stacks ne permettent pas l'exploitation de données appariées. Nous avons donc décidé de concaténer les données obtenues (i.e. les lectures R1 et R2 appariées et non appariées) puisqu'il n'est pas possible de concaténer les R1 et R2 après analyse et filtrage des lectures par le pipeline Stacks. En effet, l'attribution de numéros d'identité à chaque locus lors de la construction des catalogues ne permet pas de fusionner des catalogues *a posteriori* puisqu'un même numéro correspond alors à deux locus, l'un provenant du premier catalogue et l'autre du second.

B - Cartographie des lectures individuelles sur génome de référence de *A*. *pompejana*

Une fois les lectures démultiplexées, process_rad_tag nous fournit 4 fichiers par individu. Ces fichiers correspondent aux lectures 1 et 2, appariés et non-appariés. Nous avons réalisé une première cartographie de contrôle des lectures d'*A. pompejana* appariés sur le génome de référence obtenu par le Consortium *Alvinella* (coord. A. Claridge-Chang, A_Star* Institute, Singapour). Ce génome comprend environ 3000 scaffolds de taille comprise entre 50 Kb et 800 kb couvrant à peu près les deux tiers des 450 Mb du génome.

Après filtrage, les lectures retenues sont composées exclusivement d'une séquence provenant de l'individu séquencé, les adaptateurs ayant été éliminés. Pour ce faire, les séquences brutes sont normalement tronquées au niveau du site de restriction PstI (CTGCAG) et MseI (TTAA). Or nous avons pu constater pour les individus d'A. pompejana que le site PstI était remplacé occasionnellement par le site NsiI (ATGCAT) qui diffère du site Pst1 par la première et la sixième base du motif reconnu. Cet artéfact dans la digestion du génome a donc résulté en un nombre de locus potentiels plus grand qu'attendu par l'analyse *in silico* qui nous avait permis de choisir les enzymes de restriction adaptées à la génération de 20 000 locus environ. Lors de cette analyse, le nombre de fragments obtenus après digestion était prédit en extrapolant les résultats obtenus pour une séquence aléatoire présentant le même niveau de GC que celui de notre génome de référence. De plus dans certains cas où des lectures se chevauchent (Figure III-8), nous avons pu constater que les deux bases en question ne sont pas séquencées identiquement selon que l'on soit en début ou en fin de séquence. Les séquences de référence sur le génome au niveau de ces positions ambigües présentent souvent l'un ou l'autre des deux sites de restriction suggérant qu'il ne s'agit pas là d'une erreur de séquençage selon le brin séquencé.



Figure III-8 : Exemple de séquençage différentiel du site Pst1 selon le sens du séquençage appliqué au fragment : en fin de lecture (NsiI) et début de lecture (PstI).

Une vérification du nombre de lectures appartenant à chaque catégorie (PstI et NsiI) a été faite par le biais d'une recherche des séquences respectives de ces deux enzymes dans les fichiers R1. La proportion de lectures présentant le site NsiI n'excède pas 1% du total des séquences quel que soit l'individu considéré (Figure III-9). Quelles qu'en soient les causes, pour éviter de générer un polymorphisme artificiel dans nos jeux de données, nous avons tronqué les cinq premières bases et les 11 dernières pour toutes les lectures (site de restriction et site de barcode). Pour *Alvinella pompejana*, la cartographie des lectures sur le génome a été réalisée avec le jeu de données tronquées.



Figure III-9 : Pourcentages du nombre de sites NsiI par rapport au nombre de sites PstI obtenus pour chaque individu d'*A. pompejana*.

Quelle que soit l'espèce considérée, la cartographie a été effectuée avec les R1 et R2. Ceci nous a permis d'appliquer un filtre supplémentaire prenant en compte l'information PstI/NsiI (ne conservant que l'un ou l'autre) et nous assurer ainsi plus de confiance quant à la qualité des locus trouvés. Dans le cas d'Alvinella pompejana (Ap) la cartographie des lectures a pu être faite sur le génome de référence du ver issu d'un individu échantillonné à 9°50N (EPR nord) pouvant donner lieu à un nombre plus important de mésappariements de lectures dans le cas des individus du sud de l'EPR qui présentent une divergence comprise entre 1 et 5% avec les individus du nord pour environ 4% des gènes (D. Jollivet, données non publiées). Pour Alvinella caudata (Ac), il a fallu faire un choix quant à la référence à utiliser pour la cartographie puisque nous ne disposons que de son transcriptome. Afin d'avoir un comparatif nous avons à la fois cartographié les lectures d'Alvinella pompejana sur son génome et son transcriptome de référence après avoir effectué une recherche d'ORF (Open Reading Frame en Anglais pour Cadre de lecture ouvert) et fait de même pour Alvinella caudata en prenant le transcriptome de référence d'Alvinella caudata et le génome d'Alvinella pompejana. Pour les alignements, nous avons choisi deux modalités d'alignement des lectures selon le degré de divergence de l'espèce considérée avec le génome de référence Ap, l'alignement global « End-to-end » (Ap sur Ap) ou l'alignement « Local » (Ac sur Ap) à l'aide du programme Bowtie2 sous Galaxy.

L'alignement « End-to-end » tente d'aligner les lectures sur la totalité de leur longueur et est plus souvent utilisé quand les séquences sont très proches (lectures d'*Alvinella pompejana* sur son propre génome) et de taille comparable. Dans le cas de l'alignement « Local », une méthode heuristique rapide permet d'effectuer des calculs de scores adaptés qui évitent de pénaliser les régions non-homologues et ne calculent le score que sur les régions conservées. Différents tests ont été effectués afin de déterminer la taille optimale de la graine (motif alignable) lors de l'alignement « Local ». Les résultats de ces différents alignements sont présentés dans le Tableau III-2.

Table III-2 : Bilan quantitatif des cartographies de lectures pour un individu de A. pompejana						
et A. caudata selon la séquence de référence utilisée (Les deux colonnes en couleur						
correspondent au paramétrage choisi pour les deux espèces et les pourcentages sont calculés						
par rapport au nombre total de séquence dans chaque cas)						

Espèce	Alvinella pompejana			Alvinella caudata						
Séquence de reference	génome A.p. "End-to- end"	Transcrits A.p.	CDS A.p.	génom e A.p. "End- to- end"	génom e A.p. "Local" par défaut (I=22)	génom e A.p. "Local" (I=20)	génom e A.p. "Local" (I=19)	génom e A.p. "Local" (I=18)	génom e A.p. "Local" (I=15)	transcrits A.c.
Nb total de lectures	1310580	1310580	1310580	437254	437254	437254	437254	437254	437254	437254
Lectures	1310580	1310580	1310580	437254	437254	437254	437254	437254	437254	437254
appariées	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)	(100%)
				410925	309703	231030	227199	249014	398513	
	400407	1104111	122879	(93,98	(70.83	(52.84	(51.96	(56.95	(91.14	415985
0 Match	(30,55%)	(84,25%)	(93,76%)	%)	%)	%)	%)	%)	%)	(95,14%)
					83112	138558	143563	134971		
	7480057	164718	81576	24170	(19.01	(31.69	(32.83	(30.87	27120	12650
1 Match	(57,08%)	(12,57%)	(6,22%)	(5,5%)	%)	%)	%)	%)	(6.20%)	(2,89%)
					44439	67666	66492	53269		
	162116	41751		2159	(10.16	(15.48	(15.21	(12.18	11621	8619
> 1 Match	(12,37%)	(3,19%)	214 (0,02%)	(0,49%)	%)	%)	%)	%)	(2.66%)	(1,97%)

C - Création des catalogues de locus : Stacks

Les différents fichiers de sortie générés nous permettent l'acquisition de deux informations : le nombre de SNPs (*single-nucleotide polymorphism ou polymorphisme à l'échelle des nucléotides*) qui sont obtenus par comparaison entre individus et le nombre de locus polymorphes pour chaque individu. Dans les graphiques III-10 et III-11, présentant pour chaque individu le rapport entre le nombre de SNPs et le nombre de locus polymorphes on peut voir que cette valeur est relativement constante entre les individus et se situe autour de 2,5 et 2,75 SNPs par lecture pour A. pompejana et A. caudata respectivement. Ces valeurs pourraient être plus faibles qu'attendues par la divergence des gènes nord/sud dans le cas d'A. pompejana. Nous avons néanmoins pu sélectionner les individus conservés pour la suite de l'analyse sur la base du nombre de SNPs, en ne conservant que ceux pour lesquels au moins 10 000 SNPs ont été identifiés (voir graphique).



Figure III-10 : Nombre de snps et ratio SNPs/locus polymorphes par individu d'A. caudata.



Figure III-11 : Nombre de SNPs et ratio SNPs/locus polymorphes par individu d'A. pompejana.

D - Identification des structures de populations avec *a priori* sur l'appartenance des individus à des localités géographiques ou à un habitat thermique

-Constitution des jeux de données

Dans un premier temps, les individus ont été regroupés sur la base de la géographie. Cette première distinction nous a donc amenés à définir 5 populations pour *Alvinella pompejana*, deux au Nord de l'EPR (13°N et 9°50'N) et trois au sud (7°25'S, 18°25'S et 21°33'S). De même pour *Alvinella caudata*, une population au Nord de l'EPR (9°50'N) et deux au sud (18°25'S et 21°33'S) ont été définies pour cette première analyse de la différenciation génétique des individus échantillonnés. Nous avons aussi regroupé les individus sur la base de leur habitat thermique sur les échantillons de l'EPR nord pour lesquels des données physico-chimiques étaient disponibles (données N. Le Bris, comm. Pers. ; Matabos *et al.* 2008).

Le logiciel Populations nous permet de filtrer nos données en se basant sur des critères à l'échelle des populations. Ainsi après lui avoir fourni ce que l'on appelle une « population map » (carte des populations) qui est une liste de nos individus avec une attribution à chaque population, nous n'avons conservé que les allèles présents au moins chez 80% d'individus dans toutes les populations afin de ne pas dépasser un total de 20% de données manquantes. Le nombre de locus communs entre les populations définies pour l'analyse est présenté dans la Figure III-12 sous la forme de diagrammes de Venn.

On constate ainsi pour *A. pompejana* que le nombre de locus communs entre populations varie entre 9084 et 21986 lorsque les populations proviennent du nord ou du sud de l'EPR. En revanche ce nombre passe sous le millier de locus lorsque les 5 populations nord et sud sont prises en considération. De même pour *A. caudata*, on passe de plus 4000 locus communs au nord ou au sud à une centaine de locus communs lorsque que l'on compare les populations de part et d'autre de l'équateur. Cette différence peut s'expliquer en partie du fait que le nombre d'individus ou de populations augmente dans la comparaison nord et sud mais étant donné l'intensité du phénomène, cette décroissance pourrait être due à une différenciation génétique plus importante entre nord et sud ayant conduit à l'exclusion de nombreux locus associés aux régions divergentes du génome (introns, séquences non-codantes, gènes à fort taux d'évolution).



Figure III-12 : Nombre de locus communs entre populations et habitats pour *A. pompejana* (a) et *A. caudata* (b)

-Présence/absence des locus entre individus (polymorphisme sur les sites de restriction utilisés dans la construction des librairies)

Dans un premier temps, pour prendre la totalité de l'information disponible en termes de locus cartographiés sur le génome, nous avons utilisé les données de présence/absence de l'ensemble des locus identifiés pour chaque espèce tous individus confondus en utilisant la distance de Jaccard entre individus. Cette méthode permet de s'émanciper de la divergence entre les populations et le faible nombre de locus communs. Cette analyse a donc été faite en utilisant l'information provenant de 2 200 942 locus pour *Alvinella pompejana* et 2 101 123 locus pour *Alvinella caudata* (Figure III-13). Il faut noter ici que ces locus peuvent dans certains cas correspondre à différents allèles d'un même locus pour lesquels la divergence de séquence à artificiellement amené à les considérer séparément lors de la construction des catalogues. Ainsi ce nombre ne traduit en aucune mesure le nombre de locus identifiés pour chaque individu mais

un cumul de tous les locus obtenus chez chaque individu. Le positionnement des individus correspond à la distance génétique calculée sur le nombre de locus communs entre les individus pris 2 à 2 sans *a priori* d'appartenance à une population.



Figure III-13 : Distribution des individus dans un espace à 3 dimensions (MDS pour multidimensional scaling) sur la base de matrices de distances obtenues avec la métrique de Jaccard entre paires d'individus (indice basé sur la proportion de locus présents chez les deux individus à la fois) pour (a) *A. pompejana* sur 2 200 942 de locus et (b) pour *A. caudata* sur 2 101 123 de locus.

Dans le cas d'*A. pompejana*, un regroupement des individus selon leur origine géographique est observable. Les individus provenant de 9°50N sont nettement séparés des autres localités de l'EPR et l'analyse permet également une séparation des individus selon l'habitat thermique des cheminées. Le reste des individus est partitionné en deux groupes légèrement chevauchants et composés, d'une part, de l'ensemble des individus provenant du sud de l'EPR sans distinction géographique apparente et, d'autre part, des individus provenant de 13°N sans effet de la température de l'habitat cheminée. Enfin deux individus de 21°S se regroupent avec à ceux de 9°50'N.

Dans le cas d'A. *caudata* la distinction des individus nord / sud est également observable, de même que celle des individus issus des cheminées chaudes et froides. On constate aussi une séparation des individus originaires de 18°25'S et 21°33'S. Le groupe d'individus prélevés sur les cheminées froides de 9°50'N est positionné de façon intermédiaire entre le groupe chaud de 9°50N et celui de 18°25'S. Enfin un individu prélevé dans l'environnement 'froid' de 9°50'N se regroupe avec les individus de 18°25'S.

-Différenciation génétique des populations nord et sud

Une des premières analyses réalisées a été l'estimation de la différenciation génétique entre populations (à partir de l'estimateur θ du Fst de Weir & Cockerham 1984) et tester la significativité de cette différenciation à l'aide d'un test exact de Fisher sur les différences des fréquences génotypiques entre les populations sur l'ensemble des locus communs avec le script Genepop v 4.2.2 (Rousset, 2008). Ces valeurs et les tests de différenciation sont présentés dans la Figure III-14. Seules les comparaisons Nord/Sud et 9°50'N/13°N pour *A. pompejana* présentent des valeurs Fst associées à un test exact de différenciation significatif, avec respectivement des valeurs de 0,1764 et 0,0063. Aucune autre comparaison ne présente de différenciation significative pour l'ensemble des locus considérés. Si l'on regarde chez *A. pompejana* plus précisément les lots de populations génétiquement différenciées, la proportion de locus associés à cette différenciation est respectivement de 15,05 et 1,73 % pour la comparaison Nord/Sud et la comparaison 13°N/9°50'N, respectivement.



Figure III-14 : Valeurs de Fst (Weir & Cockerham 1984) obtenues entre populations et habitats pour (a) *A. pompejana* et (b) *A. caudata*. La significativité des valeurs est calculée au seuil de 5% (***) et présentée avec le pourcentage de locus ayant un test exact de Fisher significatif au seuil de 5%.

Après cette analyse en présence/absence, les données ont été filtrées afin de ne conserver que les locus partagés entre les groupes d'intérêt avec le programme populations de Stacks pour obtenir les génotypes multilocus de nos individus. Dans un premier temps, les analyses génétiques ont été réalisées en opposant les individus récoltés au nord et au sud de l'EPR toutes localités et tous habitats confondus (Figure III-15). Dans un deuxième temps, des analyses ont été restreintes aux populations situées au nord de l'EPR puis aux populations situées au sud de l'EPR avec un nombre beaucoup plus élevé de locus (Figures II-16 et II-17). Enfin dans un troisième temps, les individus prélevés au nord de l'EPR (9°50'N et/ou 13°N) ont été regroupés selon le type de cheminées (diffuseurs 'haute température' ou cheminées plus froides) pour les deux espèces (Figure III-16) afin de tester l'hypothèse d'une différenciation génétique des individus selon l'habitat thermique. Pour chaque niveau de comparaison, les graphiques obtenus avec AFC 3D et le logiciel Structure sont mis en parallèle.

Chapitre III – Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processus de spéciation chez les espèces du genre *Alvinella*



Figure III-15 : Analyse factorielle des correspondances (AFC 3D Genetix) et histogramme de proportions de chaque génome chez les individus d'*A. pompejana* sur 851 locus communs entre les populations de l'EPR nord et l'EPR sud obtenu à l'aide du logiciel Structure avec K=2 (log(likelihood)=-12398,0).

Dans le cas d'*A. pompejana*, l'analyse portant sur les individus des 5 champs hydrothermaux et basée sur 851 locus communs (Figure III-15) montre une nette différenciation de part et d'autre de l'EPR que ce soit avec l'AFC 3D ou avec l'analyse Structure. Les trois axes de l'AFC expliquent respectivement 4,88, 3,36 et 3,00 % de la variance observée au niveau des fréquences alléliques dans le jeu de données. Cette première analyse suggère que la différenciation nord/sud des populations repose sur un faible nombre de locus ou un faible différentiel des fréquences alléliques à ces locus. La séparation des populations nord et sud est entièrement expliquée par le premier axe de l'analyse alors que l'axe 3 montre un début de différenciation génétique entre 9°50'N et 13°N, d'une part et entre 21°33'S et 18°25'S d'autre part même s'il existe une zone de recouvrement entre individus. L'analyse structure montre quant à elle une séparation des individus en 2 unités génétiques correspondant respectivement au nord et au sud. L'augmentation du nombre de classes ne permet pas de distinguer un troisième fond génétique. De plus on peut constater que les deux individus provenant de 7°S qui sont majoritairement assignés au fond génétique du sud avec des traces d'introgression

issues du fond génétique du nord à hauteur de 10%. L'habitat thermique des *Alvinella* n'ayant pas été défini pour les individus de l'EPR sud, aucune analyse de différenciation par l'habitat n'a été effectuée à cette échelle d'espace.



Figure III-16 : AFC et diagrammes Structure (K=2 et K=3) obtenus pour *A. pompejana* sur 11 069 locus communs entre les champs 7°25'S, 18°25'S et 21°33'S de l'EPR sud. Le troisième graphe de Structure correspond à un agrandissement du graphe 2.

Lorsque que les analyses sont effectuées de part et d'autre de l'EPR, 11 069 locus communs aux localités du sud, 13 798 locus communs entre 9°50'N et 13°N et 13 412 locus communs entre les deux habitats thermiques associés au populations nord ont été identifiés. On y retrouve respectivement 30,1, 78,7 et 83,2% des locus qui étaient déjà communs entre les populations nord et sud.

L'analyse menée sur les trois populations d'*A. pompejana* au sud de l'EPR confirme l'hypothèse d'une introgression du fond génétique nord chez les deux individus de 7°25'S (Figure III-16). Les trois premiers axes de l'AFC expliquent 12,51, 5,83 et 5,60% de la variance observée dans ce jeu de données avec une nette séparation des deux individus de 7°25'S des autres individus du sud. L'analyse structure permet quant à elle de distinguer l'introgression des 2 individus de 7°25'S par le fond génétique nord (à hauteur de 10 et 17%), ce qui est du même ordre de grandeur que ce que l'on observait dans l'analyse précédente. De plus, bien que la valeur de vraisemblance soit meilleure pour K=2, lorsque la simulation est faite avec K=3, un troisième fond génétique en faible proportion (inférieure à 5%) apparaît chez les individus de 21°33'S et pourrait correspondre à un flux de gènes limité issu des individus d'*A. pompejana* de la dorsale Pacifique-Antarctique situés plus au sud comme cela a été montré par Jang *et al.* (2016).

Quand on s'intéresse aux populations du nord (Figure III-12), les analyses comparant les types d'habitat ou les champs hydrothermaux 9°50'N et 13°N ont permis de récupérer respectivement 13 412 et 13 798 locus communs. En revanche, Structure et l'analyse factorielle des correspondances n'ont pas permis de mettre en évidence de structure génétique à cette échelle. Afin de vérifier l'éventualité d'un effet croisé de la sélection par l'habitat et de la différenciation des localités, la même analyse a été faite en catégorisant les individus en 4 échantillons : les individus 9°50'N sur habitat chaud et ceux sur habitat froid, et les individus 13°N sur habitat chaud et ceux sur habitat froid, et les individus 13°N sur habitat chaud et ceux sur habitat froid, et les individus en groupes bien définis. En revanche, l'analyse Structure permet de distinguer deux fonds génétiques (Figure III-17), l'un de ces deux fonds étant minoritaire (<3%) et presque exclusivement trouvé chez les individus présents dans l'habitat 'froid'. Un test de Chi2 sur la fréquence de ce fond entre habitats indique que cette introgression allélique affecte plus spécifiquement et de manière significative l'habitat 'froid' de 9°50'N.



Figure III-17 : Proportion des fonds génétiques du sud et du nord d'A. *pompejana* sur les 9 594 locus communs entre les 4 échantillons (localité x habitat) au nord avec K=2 (log(likelihood)= -109037,5).

Ainsi les analyses effectuées de façon indépendante sur l'EPR nord ou l'EPR sud s'appuient sur un plus grand nombre de locus. Néanmoins, ces analyses ne sont pas discriminantes à cette échelle spatiale en raison d'un grand nombre de locus quasi-monomorphes (allèle alternatif sous forme de singleton ou en fréquence très faible). En raison de la faible couverture sur nos locus, ces allèles rares (et non informatifs) peuvent être la résultante d'erreurs de séquençage ou à des excès de singletons dus à un effet démographique et/ou sélectif passé. Pour pouvoir affiner notre analyse sur la différenciation génétique des populations, une sélection de locus a été faite au sein des catalogues en prenant l'hétérozygotie comme premier critère (en se limitant à des locus avec des allèles en fréquence intermédiaire afin d'éviter les erreurs de séquençage) et en triant ensuite ces locus sur la base de la valeur de Fst (selon l'estimateur θ de Weir & Cockerham 1984) afin de mieux discriminer les populations considérées.

L'utilisation de ces 2 critères a abouti à la conservation des 500 premiers locus parmi les 13412 communs aux deux habitats du nord et les 13798 communs aux deux localités du nord. Une fois cette sélection faite, une deuxième sélection a été faite sur ces 2 fois 500 locus pour ne conserver que ceux partagés entre les populations nord et sud. Cette sélection a permis de conserver 23 et 31 locus présents au sud et différenciant respectivement les individus selon le

type d'habitat et leur position géographique au nord. Il faut noter que seuls deux locus sont communs à ces groupes de locus.

Les résultats de l'analyse Structure générée sur la base des locus discriminants en fonction du type d'habitat au nord mais communs avec les populations sud sont présentés dans la Figure III-18. Tout d'abord on retrouve la différenciation entre nord et sud. Ces données montrent aussi qu'une proportion non négligeable d'individus du nord présente un fond génétique spécifique du sud sur les locus impliqués dans la différenciation par l'habitat. On retrouve aussi un individu présentant une signature nord sur le champ 21°33'S pour ces mêmes locus. Ainsi la présence d'un fond génétique sud au nord semble moins conditionnée par le type d'habitat à 13°N, qu'à 9°N où elle apparaît plus fréquente dans l'habitat « chaud » si l'on considère les individus présentant plus de 30% d'allèles sud au nord. Etant donné les résultats obtenus précédemment (Figure III-17) il semble qu'il y ait effectivement une introgression préférentielle selon l'habitat et la localité, mais que celle-ci ne se fasse pas identiquement selon les locus considérés.



Figure III-18 : Histogramme des proportions des fonds génétiques nord et sud chez les individus *A. pompejana* obtenues par le logiciel Structure sur 23 locus communs entre le nord et le sud différenciant les types d'habitat au nord (K=2, log(likelihood)=-662,8)

Les résultats de la même analyse réalisée avec les 31 locus différenciant les localités 13°N et 9°50'N sont présentés dans la Figure III-19. Cette analyse permet cette fois-ci de distinguer trois fonds génétiques spécifiques : l'EPR sud, 9°50'N et 13°N. A 9°50'N, les individus présentant un mélange de fonds génétiques sont les mêmes que ceux détectés comme « hybrides » entre les fonds nord et sud avec les locus impliqués dans la différenciation par l'habitat. Cette correspondance n'est pas retrouvée avec les individus issus de 13°N. En revanche, les individus issus des habitats « froids » ne sont pas des individus présentant des

mélanges d'allèles nord/sud dans les 2 localités, ces individus présentant leur propre signature géographique. Ces résultats suggèrent que l'habitat « chaud », qui constitue en général un nouvel environnement à coloniser pour l'espèce pourrait recevoir plus de migrants issus d'origine géographique différente que l'habitat « froid » ou la compétition intra-spécifique est sans doute plus grande. Il est aussi possible que la proportion d'habitats « chauds » soit plus grande au sud, expliquant ainsi l'établissement préférentiel de ce fond génétique dans les habitats « chauds » du nord.



Figure III-19 : Histogramme des proportions des fonds génétiques chez les individus *A*. *pompejana* obtenues par le logiciel Structure sur 31 locus communs entre le nord et le sud différenciant 9°50'N de 13°N (K=3, log(likelihood)=-899,9). A noter que seulement 2 locus sont communs avec les locus impliqués dans la différenciation par l'habitat.

Ces mêmes jeux de données ne permettent pas de distinguer les individus du nord en fonction des types d'habitats avec une AFC, mais montrent une nette séparation des individus 13°N et 9°50N (Figure III-20). A partir des trois premiers axes qui expliquent respectivement 9,91, 8,32 et 7,16% de la variance observée, trois groupes y sont identifiables et correspondent aux trois fonds génétiques identifiés dans l'analyse Structure correspondante. Chaque axe explique une partie de la différenciation génétique entre localités, séparant ainsi les individus de 13°N, de ceux du sud et enfin de ceux de 9°50'N. On retrouve dans cette analyse les tendances identifiées précédemment, soit une position intermédiaire des individus d'habitats « chauds » entre leur localité et le sud EPR ainsi qu'une proximité plus marquée de 13°N avec le sud que dans le cas de 9°N.

Chapitre III – Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processus de spéciation chez les espèces du genre *Alvinella*



Figure III-20 : AFC obtenue pour *A. pompejana* sur 31 locus communs entre le nord et le sud différenciant les localités au nord.

La même approche a été utilisée pour déterminer la structure des populations chez *A. caudata*. Dans le cas d'*A. caudata* l'étude basée sur les locus communs entre les populations de l'EPR nord et l'EPR sud ne permet de récupérer que 137 locus communs. Le résultat de l'AFC générée à partir de ces données est présenté dans la Figure III-21. Les trois premiers axes expliquent respectivement 9,66, 5,62 et 4,79% de la variance des fréquences alléliques entre les 3 populations échantillonnées (9°50'N, 18°25'S et 21°33'S). L'axe 1 sépare les populations nord et sud, et place les individus de 9°50'N issus de l'habitat froid en position intermédiaire. Les axes 2 et 3 quant à eux permettent de séparer les 2 populations géographiques de l'EPR sud. L'analyse Structure sur les 3 populations est présentée sur la Figure III-22) a été réalisée en prenant l'ensemble des SNPs disponibles, une sélection aléatoire d'un seul SNP par locus ne permettant pas une bonne reproductibilité des résultats en raison du trop petit nombre de locus identifiés.



Figure III-21 : Distribution des individus *A. caudata* issus des populations nord et le sud de l'EPR sur les 3 axes de l'analyse factorielle des correspondances à partir de 137 locus communs.



Figure III-22 : Histogramme des proportions des 2 fonds génétiques obtenus par le programme Structure (K=2) sur les individus d' *A. caudata* à partir de 137 locus communs entre le nord et le sud de l'EPR ($\log(likelihood)=-6527,2$)

L'analyse Structure montre la coexistence de 2 fonds génétiques, l'un majoritairement trouvé au nord et l'autre majoritairement trouvé au sud avec des individus intermédiaires (potentiellement « hybrides ») à 9°50'N présentant un mélange des fonds génétiques avec entre 40 et 60% de fond génétique du sud. Ces individus « hybrides » sont majoritairement trouvés dans l'habitat froid de 9°50'N (7 individus sur les 12 séquencés). Lorsque les comparaisons sont faites entre types d'habitats au nord ou entre localités au sud, le nombre de locus augmente. En ce qui concerne la comparaison entre types d'habitats à 9°50'N à partir de 4 018 locus communs, ni l'analyse Structure ni l'AFC ne permettent de distinguer des groupes d'individus. Lorsque la comparaison est faite sur les localités du sud, l'analyse Structure montre la présence d'un fond génétique supplémentaire à hauteur de 30% chez deux individus de $18^{\circ}25$ 'S à partir des 4 882 locus communs (Figure III-23). Bien que le score de maximum de vraisemblance soit meilleur pour K=2, l'analyse avec 3 classes permet de distinguer le fond génétique identifié chez les deux individus de $18^{\circ}25$ 'S à hauteur de 35% et un fond génétique supplémentaire chez un individu de $21^{\circ}33$ 'S à hauteur de 10% mais cela reste anecdotique.



K = 2

Figure III-23 : Histogrammes des proportions de fonds génétique obtenues pour chaque individu par le logiciel Structure chez *A. caudata* sur 4.882 locus communs entre les 2 localités du sud (K=2, log(likelihood)=-24887,7 et K=3, log(likelihood)=-24654,5).

Comme pour A. pompejana, un test a été fait afin de ne conserver que les locus les plus informatifs (hétérozygotie et Fst), mais il n'a pas été possible d'augmenter la résolution de nos analyses.

E - Identification des locus 'outlier' en Fst entre populations et effet de la sélection diversifiante

-A. pompejana

La distribution des valeurs de Fis mono-locus au sein des populations et de Fst mono-locus entre populations pour les 851 locus identifiés entre l'EPR nord et l'EPR sud (Figure III-24c et II-24d) ont une valeur modale égale à zéro avec plus de 70% des Fis présentant une valeur nulle. Pour ce qui est des valeurs de Fst, seule une quinzaine de locus ont des valeurs élevées significativement différentes. On observe tout de même un nombre non négligeable de locus ayant des Fis proches de 1 ou -1 et cette tendance s'accentue avec les locus identifiés comme étant sous sélection positive par Lositan (Figure III-24 a) et par DetSel (Figure III-24 b). En

effet les locus outliers présentent une distribution normale avec une longue queue de distribution.

Ces distributions suggèrent donc que la majorité des locus est à l'équilibre de Hardy-Weinberg et ne présente donc ni excès, ni déficit en hétérozygotes avec un très faible nombre de locus impliqués dans de la différenciation nord/sud. De façon logique, les locus identifiés comme sous sélection sont aussi ceux présentant les valeurs de Fst les plus élevées puisque que le modèle assume une différenciation des populations. Ces outliers sont répartis aléatoirement en termes de déséquilibres en hétérozygotes et couvrent toute la gamme des valeurs de Fis. Or ces valeurs de Fis sont des valeurs moyennées entre les deux populations, pour cette raison une vérification a été faite afin de connaître le comportement de ces locus de part et d'autre de l'EPR (Figure III-24b).



Figure III-24 a : Résultat de l'analyse Lositan représentant l'ensemble des 851 locus communs d'A. *pompejana* entre nord et sud selon leurs valeurs d'hétérozygotie et de Fst et identifiant l'intervalle de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée.



Figure III-24 b : Résultat de l'analyse DetSel représentant l'ensemble des 851 locus communs d'A. *pompejana* entre nord et sud selon leurs valeurs prises pour deux variables calculées par le programme et identifiant l'intervalle de confiance sous l'hypothèse d'une spéciation en cours définissant les locus neutres et ceux sous sélection positive.



Figure III-24 c : Distribution des 851 locus communs *d'A. pompejana* entre nord et sud selon leurs valeurs de Fis moyen et de Fst entre EPR nord et EPR sud avec identification des locus identifiés comme 'outliers' sous l'hypothèse d'un modèle en îles (Lositan, en rouge) ou d'isolement avec migration (Detsel, en jaune). Le diagramme de droite représente le résultat



des tests de permutation (10000) réalisés avec Genetix sur les valeurs de Fst des locus « outliers » Fst identifiés par Lositan (la colonne rouge correspondant à la mesure réelle).

Figure III-24 d : a. Distribution des 851 locus communs *d'A. pompejana* entre nord et sud selon leurs valeurs de Fis dans chaque population avec identification des locus identifiés comme 'outliers' sous l'hypothèse du modèle en îles (en rouge) ou du modèle d'isolement avec migration (en jaune). b. Distribution des p-values calculées pour chaque locus dans chaque population en fonction des valeurs de Fis avec une valeur seuil de Fis (0,5) pour laquelle l'essentiel des p-values est inférieur à 0,1. c. Tableau traduisant la distribution des locus dans le graphique (a.) selon qu'ils sont ou non dans l'intervalle de confiance des p-values (-0,5 < Fis < 0,5).

La distribution des p-values nous a permis de voir que seules les valeurs de Fis supérieures à 0,5 sont significatives. Ainsi on constate que 90,03% des locus ont une distribution neutre dans les deux populations à la fois. Très peu (0,14%) de locus présentent des valeurs de Fis significativement supérieures à 0 dans les deux populations, ce qui réfute l'hypothèse d'effets Walhund entre populations (qui aurait en plus induit une recrudescence de locus à fort Fst). On peut aussi noter une proportion importante de locus présentant des valeurs significativement négatives dans l'une des deux ou les deux populations représentant 1,45, 2,31 et 1,59% respectivement au nord, au sud ou dans les deux populations. Ce cas de figure correspond à un excès d'hétérozygotie dans l'une des populations ou partagé dans les deux, ce qui pourrait être soit dû à des locus dupliqués non filtrés, soit à de la sélection balancée. Dans les cas où cet excès d'hétérozygotes n'est observable que dans l'une des deux groupes on peut rejeter l'éventualité d'une paralogie étant donné qu'il est peu probable que la duplication ne survienne

que dans l'un des deux groupes. Lorsque l'on regarde plus particulièrement les locus « outliers » Fst, tous sont distribués le long des Fis = 0 dans l'une ou l'autre des populations. Nous avons voulu observer plus précisément la façon dont le polymorphisme se distribuait dans les deux groupes pour les locus « outliers » sous Detsel. L'analyse des fichiers populationnels nous a permis d'identifier l'identité génotypique des individus pour ces mêmes locus (Figure III-25).



Figure III-25 : Fréquences génotypiques des groupes pour les 8 locus identifiés comme 'outliers' par DetSel (modèle d'isolement avec migration). Bleu = Génotype homozygote AA, Rouge = Génotype homozygote BB, violet = Génotype hétérozygote AB.

Cette analyse nous permet de voir que dans la majorité des cas, les deux formes homozygotes de nos polymorphismes bi-alléliques sont respectivement spécifiques du nord et du sud de l'EPR. On ne retrouve l'allèle minoritaire que sous forme hétérozygote au sein de chaque population. Et pour 2 de ces 8 locus, on ne retrouve au nord presque que des individus hétérozygotes. En fonction de l'hétérozygotie observée deux explications sont possibles. Dans le cas où les individus hétérozygotes sont en fréquence faible dans les populations, il est probable qu'il s'agisse d'un allélisme diagnostique avec un très faible échange d'allèles à travers la barrière, l'allèle « introgressé » est alors en trop faible fréquence pour apparaître à l'état homozygote. A l'inverse dans le cas où les hétérozygotes sont majoritaires dans au moins l'une des populations (souvent au nord apparemment), on peut envisager (1) soit qu'il y ait eu duplication d'un côté ou de l'autre de la barrière et que cette duplication ait été ensuite bloquée par la barrière, (2) qu'il y ait eu un polymorphisme ancestral mais que l'un des 2 allèles ait été fixé par dérive dans l'une des 2 populations et, (3) qu'il y ait un maintien adaptatif des hétérozygotes dans l'une des deux populations ou qu'une des 2 populations soit une zone d'hybridation entre le sud et le nord.

Pour *A. pompejana*, l'analyse de la différenciation génétique des populations de l'EPR nord (13°N et 9°50'N) avec 13 798 locus et celle des populations de l'EPR sud (18°25'S et 21°33'S) avec 10 561 locus (Figures II-26 b et II-26 c) montre des situations similaires du point de vue de la distribution des valeurs de Fis moyen et de Fst que prennent les locus. En revanche, on observe une proportion nettement plus importante de locus identifiés comme 'outliers' dans le cas de la comparaison 13°N/9°50'N. Dans les deux cas, les valeurs de Fst sont centrées sur zéro et ne dépassent pas 0,5 alors que certains allèles étaient fixés de part et d'autre de la barrière équatoriale dans l'analyse précédente avec des Fst de 1. Cette distribution est encore plus resserrée autour de zéro entre les localités du sud de l'EPR, confirmant l'absence de différenciation entre 21°33'S et 18°25'S dans les analyses globales présentées dans la Figure III-14.

En ce qui concerne les valeurs moyennes de Fis entre populations nord et entre populations sud, la plupart des locus présente des valeurs proches de zéro avec cependant une proportion non négligeable de valeurs plus élevées proches de 1 traduisant un déficit en hétérozygotes et potentiellement un effet Walhund. Dans un cas comme dans l'autre, si l'on regarde plus spécifiquement les locus 'outliers', la distribution des valeurs moyennes de Fis reste la même, avec cependant une augmentation de la proportion des locus à Fis = 1. Les résultats des analyses Lositan ayant permis d'identifier les locus outliers sont présentés en Figure III-26 a et représentent les individus selon leurs valeurs de Fst et d'hétérozygotie ainsi que l'enveloppe de confiance permettant de déterminer quels locus sont identifiés comme sous sélection positive ou balancée. Pour mieux comprendre comment se distribue le polymorphisme des différents locus, les valeurs de Fis propres à chaque population sont présentés sous forme de 'biplots' dans la Figure III-26b. Que ce soit au nord ou au sud, la grande majorité des locus présente des valeurs de Fis non-significativement différentes de zéro dans les deux populations comparées (91,57% 13°N/9°50'N et 95,71% 18°25'S/21°33'S), ce qui correspond à des locus à l'équilibre de Hardy-Weinberg dans chaque population.

Dans le cas de la comparaison 13°N/9°50'N, on retrouve cependant 2,63% et 4,36% des locus (environ 7% cumulés) présentant un déficit en hétérozygotes à 13°N ou à 9°50'N, et une fraction de locus avec des excès d'hétérozygotes dans les 2 populations (0,75%). Si l'on regarde plus précisément les locus identifiés sous sélection positive, on remarque qu'ils se distribuent sur les deux axes correspondant à des Fis nuls dans l'une ou l'autre des populations. Comme dans le cas de la comparaison Nord/Sud, ce cas de figure correspond à la fixation d'un des deux allèles dans l'une des deux populations. La densité des locus 'outliers' est néanmoins beaucoup

plus forte dans le cas d'un Fis nul (ou quasi-nul) à 13°N avec des déficits ou des excès d'hétérozygotes à 9°50'N. Cette différence de densité entre les deux axes pourrait s'expliquer par une arrivée plus importante d'allèles en provenance des populations du sud de l'EPR.

Dans le cas de la comparaison 18°25'S/21°33'S, la distribution des valeurs de Fis est beaucoup plus homogène avec un nombre de locus 'outliers' beaucoup plus faible qu'au nord. Les locus 'outliers' sont néanmoins distribués principalement sur l'axe correspondant à un Fis nul à 18°25'S, suggérant là encore qu'une arrivée d'allèles en provenance de la dorsale Pacifique-Antarctique située plus au sud, pourrait affecter plus spécifiquement la population de 21°33'S en passant à travers la barrière de la microplaque de l'île de Pâques.



Figure III-26 a : Résultats des analyses Lositan représentant pour *A. pompejana* les 13 798 locus communs entre les localités du nord (en haut) et des 10 561 locus communs entre les localités du sud (en bas) selon leurs valeurs d'hétérozygotie et de Fst et identifiant les intervalles de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée



Figure III-26 b : Distribution des 13 798 locus communs entre les localités du nord (en haut) et des 10 561 locus communs entre les localités du sud (en bas) selon leurs valeurs de Fis moyen et de Fst entre populations d'*A. pompejana* avec identification des locus 'outliers' sous l'hypothèse du modèle en îles avec Lositan (rouge). Les diagrammes de droite représentent le résultat des tests de permutation (10000) réalisés avec Genetix sur les valeurs de Fst des locus « outliers » Fst (la colonne rouge correspondant à la mesure réelle).

Chapitre III – Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processus de spéciation chez les espèces du genre *Alvinella*



Figure III-26 c : 1a. Distribution des 13 798 locus communs d'A. pompejana entre 13°N et 9°50'N selon leurs valeurs de Fis dans chaque population avec identification des locus 'outliers' sous l'hypothèse du modèle en îles avec Lositan (rouge). 1b. Distribution des p-values calculées pour chaque locus dans chaque population à l'aide d'un test exact de Fisher sur les distributions génotypiques et avec la valeur seuil de Fis (0,5) pour laquelle la majorité des p-values sont inférieures à 0,1. 1c. Tableau traduisant la distribution des locus dans le graphique selon qu'ils sont ou non dans l'intervalle de confiance des p-values.

2a. Distribution des 10 561 locus communs *d'A. pompejana* entre 18°25'S et 21°33'S selon leurs valeurs de Fis dans chaque population avec identification des locus 'outliers' sous l'hypothèse du modèle en îles (positive sel). 2b. Distribution des p-values calculées pour chaque locus dans chaque population à l'aide d'un test exact de Fisher sur les distributions génotypiques et représentation de la valeur seuil de Fis (0,75) pour laquelle la majorité des pvalues sont inférieures à 0,1. 2c. Tableau traduisant la distribution des locus dans le graphique selon qu'ils sont ou non dans l'intervalle de confiance des p-values. L'analyse des Fis et des Fst a été également effectuée avec 13 412 locus en regroupant les individus selon le type d'habitat thermique d'*A. pompejana* avec les données environnementales qui avaient pu être obtenues au niveau des populations de 9°50'N et 13°N (Figures II-27 b et II-27 c). Les distributions des valeurs de Fst et de Fis sont sensiblement identiques à celles précédemment observées entre localités, mais avec des valeurs Fst plus faibles que dans les autres niveaux de comparaison et une distribution encore plus resserrée autour de zéro.

L'analyse des valeurs de Fis au sein de chaque population (Figure III-27 c) montre là encore une forte proportion de locus à valeurs de Fis qui ne sont pas significativement différentes de zéro (92,34%) dans les deux types d'habitat. On remarque également que 3,65% et 2,68% de locus présentent des valeurs de Fis positive spécifiquement dans l'habitat chaud ou dans l'habitat froid, respectivement, traduisant un déficit en hétérozygotes. Sachant que la différenciation entre habitats semble affectée par des effets différents (introgression, sélection locale, migration longue distance) selon que l'on soit à 13°N ou à 9°50'N (Figures III-18 et III-19), il est probable qu'en regroupant des individus de ces deux localités sur la base de leur habitat, on puisse générer un effet Wahlund. Ainsi les populations de 9°50'N et 13°N possèdent leur propre signature allélique avec des échanges limités. Ces déficits en hétérozygotes majoritairement observés dans l'habitat chaud (Figure III-27 c) pourraient être expliqués par l'action d'une sélection diversifiante sur l'un des 2 allèles dans cet habitat et la fixation de l'allèle alternatif dans l'autre habitat.



Figure III-27 a : Résultat de l'analyse Lositan représentant pour *A. pompejana* les 13 412 locus communs entre les types d'habitat selon leurs valeurs d'hétérozygotie et de Fst et identifiant l'intervalle de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée


Figure III-27 b : Distribution des 13 412 locus communs entre types d'habitat du nord d'*A. pompejana* selon leurs valeurs de Fis moyen et de Fst entre populations avec identification des locus 'outliers' sous l'hypothèse du modèle en îles avec Lositan (en rouge). Le diagramme de droite représente le résultat des tests de permutation (10000) réalisés avec Genetix sur les valeurs de Fst des locus « outliers » Fst (la colonne rouge correspondant à la mesure réelle).

Chapitre III – Rôle de l'adaptation locale et des barrières physiques à la dispersion dans le processus de spéciation chez les espèces du genre *Alvinella*



Figure III-27 c : a. Distribution des 13 412 locus communs *d'A. pompejana* entre types d'habitat du nord de l'EPR selon leurs valeurs de Fis dans chaque groupe avec identification des locus 'outliers' sous l'hypothèse du modèle en îles avec Lositan (en rouge). b. Distribution des p-values calculées pour chaque locus dans chaque population à partir d'un test exact de Fisher avec la représentation de la valeur seuil de Fis (0,5) pour laquelle la majorité des p-values sont inférieures à 0,1. c. Tableau traduisant la distribution des locus dans le graphique selon qu'ils sont ou non dans l'intervalle de confiance pour les p-values.

-A. caudata

La filtration des données RAD chez *A. caudata* après cartographie sur le génome d'*A. pompejana* nous a permis d'obtenir 137 locus communs entre les populations de l'EPR nord et l'EPR sud (Figure III-28 b et c). Dans ce cas, le Fst moyen calculé (0,2752) sur l'ensemble des locus n'est pas significatif probablement en raison du faible nombre de locus utilisé. L'analyse des locus communs au nord et au sud permet tout de même de mettre en avant une structure géographique des populations. Pour cette raison, l'analyse DetSel a été réalisée en plus de l'analyse Lositan (Figure III-28 a), mais cette première n'a pas permis d'identifier de locus « outliers ». Les distributions des 137 valeurs de Fis et de Fst sont centrées sur des valeurs proches de zéro, avec tout de même des queues de distribution très étalées. Dans le cas de l'histogramme de classes des locus 'outliers', on retrouve une situation assez comparable à celle trouvée chez *A. pompejana* entre localités nord et sud avec un lot de locus présentant des valeurs de Fis positives élevées traduisant un déficit en hétérozygotes. Cette dominance de locus à Fis positifs dans les locus « outliers » se retrouve également dans les comparaisons entre localités

du sud (Figure III-30 a) ou entre les 2 types d'habitat thermique définis sur les cheminées de l'EPR nord (Figure III-29 a). Lorsque l'on regarde la distribution des Fis entre nord et sud (Figure III-28 c), on peut voir que 98,55% des locus ne présentent pas de valeurs significativement différentes de zéro dans les deux populations et qu'aucun locus ne présente de valeurs de Fis négatives au sud. On retrouve tout de même 1,01% de locus avec des valeurs de Fis significativement négatives au nord et au sud, ce qui correspond à un excès d'hétérozygotes communs aux deux populations suggérant de la sélection balancée ou des locus dupliqués non filtrés par Stacks.



Hétérozygotie

Figure III-28 a : Résultat de l'analyse Lositan représentant l'ensemble des 137 locus communs d'A. *caudata* entre nord et sud selon leurs valeurs d'hétérozygotie et de Fst et identifiant l'intervalle de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée.



Figure III-28 b : Distribution des 137 locus communs d'*A. caudata* entre l'EPR nord et sud selon leurs valeurs de Fis moyen et de Fst entre populations avec identification des locus 'outliers' sous l'hypothèse du modèle en îles (en rouge : sélection diversifiante et en vert : sélection balancée).



Figure III-28 c : a. Distribution des 137 locus communs *d'A. caudata* entre nord et sud selon leurs valeurs de Fis dans chaque population avec identifications des locus identifiés comme outliers sous l'hypothèse de populations en îles (en rouge : sélection diversifiante et en vert : sélection balancée).

b. Distribution des p-values calculées pour chaque locus dans chaque population en fonction des valeurs de Fis elles-mêmes et représentation de la valeur seuil de Fis (0,5) pour laquelle l'essentiel des p-values sont inférieures à 0,1. c. Tableau traduisant la distribution des locus dans le graphique (a.) selon qu'ils passent ou non dans l'intervalle de confiance pour les p-values.

Dans le cas de la comparaison entre types d'habitats au nord pour *A. caudata* (Figures III-29 a, b et c), la distribution des valeurs de Fst sur 4 018 locus est majoritairement centrée autour de zéro avec tout de même quelques locus présentant des valeurs plus fortes allant jusqu'à 0,35. La distribution des valeurs de Fis est également centrée sur zéro avec une proportion non négligeable de locus à valeurs positives ou négatives. Les valeurs de Fis des locus 'outliers' se distribuent de façon opposée à l'attendu neutre observé sur l'ensemble du jeu de données. Les locus 'outliers' sont en effet subdivisés en un groupe avec des valeurs situées entre -0,25 et -0,5 (excès d'hétérozygotes) et un groupe avec des valeurs entre 0,2 et 0,6. Ces résultats suggèrent que les locus impliqués dans de la différenciation par l'habitat présentent tous des déséquilibres en hétérozygotes que l'on ne retrouve pas dans les locus « neutres ». L'analyse en biplot des Fis par habitat (Figure III-29 c) montre que 2,60% et 2,02% des locus ont des valeurs de Fis significativement positives (déficit en hétérozygotes) dans l'habitat chaud ou dans l'habitat froid, respectivement. Ceci peut être le résultat d'une contre sélection des hétérozygotes dans l'un des deux habitats thermiques ou à l'inverse une sélection positive pour

un allèle donné dans l'un des habitats. On retrouve encore 1,73% de locus présentant un excès d'hétérozygotes dans les deux populations potentiellement associées à des locus dupliqués non filtrés ou de la sélection balancée. On observe également qu'un grand nombre de locus « outliers » ont un Fis égal à zéro dans l'habitat froid et pourraient représenter un polymorphisme maintenu par sélection balancée au sein de l'habitat chaud. En comparaison avec *A. pompejana* (Figures III-27), on retrouve nettement plus de locus identifiés comme étant sous sélection balancée avec l'habitat (degré de différenciation génétique inférieur à celui attendu par le modèle en îles). Ces locus semblent présenter des valeurs de Fis significativement négatives soit au sein des deux habitats soit au sein de l'habitat « froid ». Il est donc probable que ce soit ce partage de polymorphisme par sélection balancée qui maintienne globalement une faible différenciation entre habitats, mais il semble aussi que cet effet de l'habitat sur la différenciation des populations soit tout de même plus marqué chez *A. caudata* que chez *A. pompejana*.



Figure III-29 a : Résultat de l'analyse Lositan représentant l'ensemble des 4018 locus *d'A. caudata* communs entre types d'habitats au nord selon leurs valeurs d'hétérozygotie et de Fst et identifiant l'intervalle de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée.



Figure III-29 b : Distribution des 4 018 locus communs d'*A. caudata* entre types d'habitat sur l'EPR nord selon leurs valeurs de Fis moyen et de Fst entre populations avec identification des locus 'outliers' sous l'hypothèse du modèle en îles (sélection diversifiante en rouge et sélection balancée en vert).



Figure III-29 c : a. Distribution des 4 018 locus communs *d'A. caudata* entre types d'habitat sur l'EPR nord selon leurs valeurs de Fis dans chaque population avec identification des locus

'outliers' sous l'hypothèse du modèle en îles (vert : sélection balancée et rouge sélection diversifiante). b. Distribution des p-values calculées pour chaque locus dans chaque population avec un test exact de différenciation définissent une valeur seuil de Fis à 0,75 pour laquelle la majorité des p-values sont inférieures à 0,1. c. Tableau traduisant la distribution des locus dans le graphique selon qu'ils passent sont ou non dans l'intervalle de confiance pour les p-values.

Enfin, la même analyse réalisée sur les 4 882 locus communs d'A. caudata entre 2 localités 18°25'S et 21°33'S (Figures III-30 a, b et c) montre une distribution des valeurs moyennes de Fis décalée vers les valeurs positives suggérant un effet Wahlund dans l'une ou l'autre des 2 localités. Dans ce cas, il est possible que l'une des 2 zones géographiques corresponde à une zone de contact entre entités génétiques différentes ou que notre échantillonnage mélange des individus d'habitats différents ayant une signature génétique différente au sein d'une même localité, mais ceci n'est pas retrouvé sur les analyses Structure ou AFC si ce n'est trois individus présentant des traces de fonds génétiques différents. La distribution des valeurs de Fst est quant à elle centrée sur 0 avec tout de même une queue de distribution atteignant des valeurs de plus de 0,5. L'analyse des valeurs de Fis par localité nous permet de comprendre un peu mieux comment se comportent les locus 'outliers' avec des valeurs positives de Fis. Dans un premier temps, on peut voir qu'en réalité peu de ces locus présentent des valeurs de Fis réellement significativement différentes de zéro. Mais on observe tout de même environ 14% des locus avec des valeurs de Fis significativement positives à 18°25'S et 21°33'S. Ceci pourrait indiquer un effet Wahlund beaucoup plus important au niveau de la population de 18°25'S. Il est peu probable que l'on soit en présence de locus diagnostiques discriminant les deux localités. Tout comme dans le cas de la comparaison entres habitats (Figures III-29), on retrouve encore un grand nombre de locus identifiés comme étant sous sélection balancée, et cumulé au fait que 1,13% des locus présentent un déficit significatif en hétérozygotes, cela pourrait expliquer que la différenciation soit limitée entre localités.



Figure III-30 a : Résultat de l'analyse Lositan représentant l'ensemble des 4882 locus *d'A. caudata* communs entre localités au sud selon leurs valeurs d'hétérozygotie et de Fst et identifiant l'intervalle de confiance sous l'hypothèse d'un modèle en îles définissant les locus sous sélection positive ou balancée.



Figure III-30 b : Distribution des 4 882 locus communs d'*A. caudata* entre les localités 18°25'S et 21°33'S selon leurs valeurs de Fis moyen et de Fst entre populations avec identification des locus 'outliers' sous l'hypothèse du modèle en îles (positive sel).



Figure III-30 c : a. Distribution des 4 882 locus communs *d'A. caudata* entre les localités $18^{\circ}25$ 'S et $21^{\circ}33$ 'S selon leurs valeurs de Fis dans chaque population avec identification des locus 'outliers' sous l'hypothèse du modèle en îles (positive sel). b. Distribution des p-values calculées pour chaque locus dans chaque population en fonction du test exact de Fisher et représentation de la valeur seuil de Fis (0,7) pour laquelle la majorité des p-values sont inférieures à 0,1. c. Tableau traduisant la distribution des locus dans le graphique selon qu'ils sont ou non dans l'intervalle de confiance pour les p-values.

F - Identification des gènes associés aux locus sous sélection positive dans l'analyse de la différenciation entre les populations de l'EPR nord et sud.

Les locus identifiés comme étant sous sélection positive (ou liés physiquement à des locus sous sélection) par Lositan et/ou Detsel entre les populations nord et sud ont été cartographiés sur le génome d'*A. pompejana*. Les régions flanquantes ces locus (2000 bp encadrant le locus sur le génome) ont été récupérées et blastées (tblastx) sur les banques de données (GenBank, EMBL) afin d'identifier le ou les gène(s) associés à ces locus. Le résultat de ces analyses est présenté dans le Tableau III-3 pour *A. pompejana* où, sur 112 locus identifiés comme étant sous sélection positive, 20 ont pu être identifiés, et pour *A. caudata* où, sur 11 locus sous sélection, 2 gènes ont pu être également identifiés. La liste détaillée des gènes ou fonctions identifiées pour les séquences des 20 locus reconnus pour *A. pompejana* et des 2 locus d'*A. caudata* sont présentés dans le Tableau III-4.

Dans les deux cas, les fonctions moléculaires principalement touchées par les gènes concernés sont des enzymes ayant une fonction catalytique, notamment des transférases et des phosphatases et des protéines ayant des fonctions de transport et/ou liaison avec d'autres molécules. Enfin quand on regarde les processus biologiques impliqués, les deux locus d'*A. caudata* sont impliqués dans des processus ontogéniques tout comme 16 des 20 locus pour *A. pompejana. A. pompejana* présente aussi une tendance générale plutôt axée sur les processus impliqués dans le métabolisme de l'animal et les processus de régulation biologique impliquant chacun 16 des locus utilisés.

Table III-3 : Nombre de gènes impliqués dans les principaux processus cellulaires et fonctions métaboliques identifiés par Blast2Go pour les quels ont été assignés les 20 et 2 gènes identifiés chez *A. pompejana* et *A. caudata*

		A. pompejana	A. caudata
Processus biologiques	processus cellulaires	20/20	2/2
	métabolisme	16/20	1/2
	régulation	16/20	1/2
	processus développementaux	12/20	2/2
Fonction	liaison	12/20	1/2
moléculaire	activité catalytique	16/20	2/2

	Nombre de sequences	Espèce
	correspondantes	concernée
β 1-3 galactosyl-transférase	1	A caudata
nucléoside triphosphate phosphatase	1	A. CUUUUUU
acétyl transférase	1	
nucléoside triphosphate phosphatase	2	
DNA polymérase	1	
synaptotagmine	1	
isocitrate déshydrogénase	1	
AMPc phosphodiesterase	1	
sous unité alpha coatomer	1	ğ
c-GMP kinase dépendante	1	eja
transporteur d'acide L-ascorbique	1	du
adhésion cellulaire	1	od
ficoline	2	4
dopey	1	
stérol o acyltransferase	1	
F-box	1	
transfert de groupes phosphoreux	1	
3 déshydro sphinganine réductase	2	
gld1	1	

Table III-4 : Récapitulatif des correspondances trouvées pour les gènes associés à de la sélection positive diversifiante entre les populations nord et sud pour *A. caudata* et *A. pompejana*.

G - Diversités des polychètes *Alvinella* de part et d'autre de la barrière équatoriale

La cartographie des lectures obtenus sur les ORFs des transcrits issus de l'assemblage du transcriptome d'un individu du nord de l'EPR a permis de caractériser les diversités synonyme et non-synonyme pour les espèces *A. pompejana* et *A. caudata* au nord et au sud de l'EPR. Afin d'avoir des résultats comparables entre populations pour chaque espèce, le même nombre d'individus a été utilisé pour chaque sous population. Au final, 17 individus d'*A. pompejana* au nord et au sud ont donc été utilisés pour cette analyse et 19 dans le cas d'*A. caudata* (Figure III-31). Ces données ont également permis pour *A. pompejana* au nord et au sud et pour *A. caudata* au nord d'estimer le D de Tajima et les indices de diversité associés pour un certain nombre de transcrits (Figure III-32).



Figure III-31 : Données de polymorphisme synonyme et non-synonyme au nord et au sud de l'EPR pour *A. pompejana* et *A. caudata* générées à partir de données cartographiées sur les transcriptomes respectifs de chaque espèce, dans la partie codante des gènes. Nombre de contigs obtenus pour chaque groupe d'individus ayant permis le calcul des diversités synonyme et non-synonyme.



Figure III-32 : Boîtes à moustaches présentant la distribution des valeurs du D de Tajima, le nombre moyen de différences entre séquences et le nombre de sites ségrégeant pour les *A*. *pompejana* au nord et au sud de L'EPR (bleu) et pour *A. caudata* au nord (vert). Les boites à moustaches indiquent du centre vers l'extérieur, la valeur médiane, les quartiles ainsi que le premier et le dernier décile. Les croix indiquent les valeurs moyennes et les points les valeurs non comprises dans les boîtes à moustaches. Pour chaque population, le nombre de séquences utilisées est indiqué et pour chaque estimateur aucune différence significative n'a été identifiée à l'aide du test de Tukey (seuil de significativité 0,1).

On observe une tendance pour une diversité plus faible au sein des populations du sud chez les deux espèces sur la diversité synonyme. Pour *A. pompejana* les diversités synonymes et nonsynonymes sont plus faibles dans les populations de l'EPR sud, et pourraient être dues à un effet démographique au sud alors que dans le cas d'*A. caudata* l'absence de variations sur le non-synonyme et le recouvrement des écart-types sur le synonyme laissent penser à des tailles efficaces similaires de part et d'autre de la barrière équatoriale. On peut aussi constater que, toutes localités confondues, *A. pompejana* présente des valeurs de diversité plus faibles qu'*A. caudata*, ce qui pourrait être la marque d'un effet plus fort de la sélection sur cette espèce placée au plus près de émissions sur les cheminées. Dans les deux cas, le nombre de contigs ayant été obtenu et utilisé pour l'analyse est plus faible dans les populations du sud, une diminution atteignant 50% dans le cas d'*A. pompejana* au sud. Dans l'ensemble, les valeurs moyennes du D de Tajima sont négatives, en moyenne inférieures à -1. On ne retrouve cependant aucune différence significative quant aux distributions des valeurs de D de Tajima et pour les mesures de diversités associée entre les populations considérées (Figure III-32).

4 - Discussion

A - Limites méthodologiques

-Limitations des analyses génétiques avec des marqueurs RAD lorsque la couverture des locus est faible

La couverture des locus en nombre de lectures sur les séquences de référence ou lors de la construction des catalogues des locus entre individus est un paramètre qui permet de filtrer les données afin de vérifier si le polymorphisme identifié est réel et non le résultat d'erreurs de séquençage. En effet, les erreurs de séquençage sont statistiquement rares et distribuées aléatoirement entre lectures. Elles ont donc beaucoup plus de chance d'apparaître sous la forme de singletons dans le jeu de données et avec une couverture élevée et un filtre sur la fréquence des allèles rares, on peut limiter le nombre de mutations artéfactuelles dans le jeu de données. Dans notre cas, la couverture exigée pour la construction de nos catalogues a été volontairement abaissée à 2 lectures par locus étant donné le faible nombre de locus récupérés avec une couverture supérieure, beaucoup d'individus ayant un nombre de lectures inférieur à 1 000 000. Une erreur de séquençage se caractérise par le fait qu'elle ne correspond pas à une réalité biologique et que statistiquement elle n'est pas partagée par deux individus différents (tout comme une mutation apparue récemment : 1/2N). Le polymorphisme généré par les erreurs de séquençage conduit donc à brouiller le signal de la différenciation entre populations quand on se base sur les fréquences alléliques pour identifier des sous-groupes en augmentant le nombre de locus polymorphes de façon arbitraire (locus monomorphes deviennent polymorphes par la présence de singletons artéfactuels). Pour pallier à ce problème, il est alors nécessaire de filtrer le jeu de données sur l'hétérozygotie ou sur la fréquence des allèles rares.

Un autre risque lié à une couverture faible des locus est celui de voir diminuer le nombre de locus en commun entre populations. En effet lorsque l'on augmente le nombre d'individus dans chaque population considérée et que la couverture individuelle est trop faible (en raison de la faible couverture des locus), on augmente par la même occasion de façon considérable la

probabilité qu'un locus donné soit perdu au sein d'une population lors du filtrage sur la base de sa présence/absence dans chaque population.

-Construction de catalogues de locus sur la base de similitudes et divergence entre allèles

Lors de la construction des catalogues individuels, une divergence maximale est définie afin d'éviter tout risque d'assimiler de la paralogie à de l'allélisme et ainsi de générer encore une fois du polymorphisme artéfactuel et potentiellement surnuméraire (plus de 2 allèles à un locus donné chez un individu). Or les allèles d'un même locus qui présentent un nombre de différences supérieur au seuil du filtre seront alors considérés comme des locus différents. Ceci risque alors de limiter l'identification de lignées alléliques différentes maintenues dans une population par sélection diversifiante ou balancée, mais on risque surtout de ne pas pouvoir identifier des lignées alléliques maintenues entre populations isolées et ayant donc divergé plus que le nombre autorisé de différences entre lectures. Ceci pourrait expliquer en partie pourquoi de façon surprenante, toutes les localités sont différenciées les unes des autres avec la distance de Jaccard (à l'exception de la différenciation 18°25'S/21°33'S pour A. pompejana) alors que cette différenciation est très peu perceptible avec les analyses multivariées sur fréquences génotypiques voire inexistante avec l'analyse Structure. En effet dans le cas ou des lignées alléliques auraient divergé suffisamment pour être considérées comme des locus par notre analyse, le contraste apporté par l'information de l'allélisme serait perdu au sein des analyses l'utilisant (Structure, AFC) et gagné par l'analyse de présence/absence.

-Limites à l'analyse avec des populations peu différenciées

La différence de résolution observée entre le programme Structure et les AFC peut s'expliquer par la faible différenciation génétique entre localités sur un très grand nombre de locus très peu polymorphes. Il a en effet été montré que les programmes d'assignation d'individus à des populations (groupes) basés sur des inférences bayésiennes plutôt que sur des *a priori* d'assignations présentaient des difficultés dans l'identification de structures hiérarchiques et ce d'autant plus lorsque les populations échangent beaucoup (Latch *et al.*, 2006). En effet, dans des populations simulées suivant un modèle en nombre infini d'îles pour une valeur de Fst inférieure à 0,03, le programme ne parvient pas à identifier l'ensemble des groupes ayant été préalablement définis ; et lorsque la valeur de Fst est inférieure à 0,01, Structure n'est plus en mesure de différencier les individus en 2 groupes. Or, dans notre étude, aucune paire de localités ne présente de valeurs de Fst excèdant 0,01, à l'exception de celle (Fst=0,014) calculée entre les localités de 18°25'S et 21°33'S pour A. pompejana. L'étude de Latch et collaborateurs (2006) a aussi mis en évidence que la sensibilité de ce type d'analyse dépend du nombre d'individus et de la quantité d'information (ici le nombre de locus), on peut donc penser que le manque de résolution observé dans nos analyses vient pour une part de la taille de nos échantillons et du nombre de locus communs non informatifs récupérés indépendamment de part et d'autre de la barrière équatoriale. Cette absence de données informatives est partiellement confirmée par le fait que le logiciel Structure est capable de retrouver une structure géographique des populations dès lors que les locus sont sélectionnés sur la base de leur hétérozygotie et d'un tri décroissant des valeurs de Fst. Ces données suggèrent que la diversité génétique des Alvinella est sans doute plus faible que ce que nous attendions (i.e. beaucoup de locus monomorphes au seuil de 95%) et que la faible couverture des locus a conduit à sur-dimensionner le nombre de locus polymorphes avec des erreurs de séquencage non filtrés.

Ainsi les biais générés par notre méthode d'analyse et de construction des jeux de données doivent être pris en compte dans l'interprétation des résultats et la discussion qui vont suivre.

B - Rôle de la barrière équatoriale dans l'évolution des populations d'*Alvinella* : spéciation en cours ou remise en contact secondaire des populations nord et sud ?

Hamilton et May (1977) ont montré avec leur modèle que même dans un environnement plus ou moins continu, il est plus intéressant pour une espèce d'investir une majorité de sa descendance dans la dispersion afin qu'elle s'établisse à bonne distance des génotypes parentaux. Et si cet habitat est fragmenté, cette capacité de dispersion doit être suffisante pour couvrir la distance moyenne entre habitats propices à l'établissement de la descendance. Si l'habitat de l'espèce présente en plus la particularité d'être transitoire, il existe en plus pour l'espèce un risque important d'extinction locale voire globale si celle-ci ne possède pas les capacités de dispersion adéquates. Ceci explique que dans de tels scénarios, le rapport bénéfice/ investissement soit élevé et d'autant plus lorsque le taux de renouvellement de ces habitats est élevé (McPeek & Holt 1992). En plus de la dispersion, une autre adaptation possible à cette distribution fragmentée et transitoire de l'habitat est la dormance (Levin et al., 1984). Ces deux possibilités ne se substituent pas l'une à l'autre et peuvent même se trouver co-favorisées lorsque l'on observe une synchronicité des habitats dans leurs transitions (Snyder 2006). Dans le milieu marin, ces deux processus peuvent être encore plus étroitement associés si la phase de dormance se superpose à la phase dispersive, et permet aux larves de maintenir un délai avant de se métamorphoser tant que les conditions d'installation ne sont pas propices (Pradillon et al. 2001). Instabilité environnementale et fragmentation ont donc un rôle très important dans l'évolution du polymorphisme des espèces et sur la distribution de cette variabilité génétique dans l'espace. D'une façon générale, la fragmentation conduit le plus souvent à une augmentation de la différenciation génétique entre populations locales par effet de dérive (Lande 1988, Nunney & Campbell 1993) et l'extinction locale, à une réduction en taille de la métapopulation entraînant une érosion de la diversité génétique (Gilpin 1991). Lorsque la distance de dispersion est inférieure à la distance moyenne entre deux habitats, comme c'est souvent le cas des environnements insulaires et cavernicoles (Juan et al. 2010), on observe souvent que la différenciation génétique des populations augmente avec la distance géographique et suit alors un modèle en pas japonais (« stepping stones model » Kimura & Weiss 1964) dès lorsqu'un équilibre démographique est atteint. Dans le cas où l'espèce peut disperser sur de grandes distances, l'évolution de la structure génétique de l'espèce dépend en grande partie du ratio entre le taux d'extinction locale et la migration, et plus spécifiquement entre le nombre de migrants échangés entre populations établies et le nombre de propagules colonisant de nouveaux habitats et l'on observe alors des exemples d'homogénéisation génétique des populations locales pour des taux d'extinction élevés (Slatkin 1977). En effet, comme le soulignent Harrisson et Hastings (1996), l'évolution d'une espèce vivant dans un environnement à fort taux de renouvellement tend à favoriser la dispersion sur de longues distances tout en réduisant sa taille efficace et donc son polymorphisme. Or la diminution de la taille efficace et du polymorphisme conjugués induisent une faible différenciation génétique des populations locales. Or nos analyses révèlent un effet non négligeable de l'isolement entre localités et de la sélection locale à l'échelle des populations dans la différenciation entre cellesci, ce qui laisse penser que dans le cas des Alvinellas, cette relation entre taux de renouvellement de l'environnement et dispersion n'est pas valable.

Une barrière aux flux de gènes a déjà été mis en évidence au niveau de l'Equateur sur la dorsale Est Pacifique par une étude du polymorphisme nucléotidique de la cytochrome oxydase I (*Cox-1*) chez 7 espèces emblématiques de l'environnement hydrothermal dont un bivalve, trois gastéropodes patelliformes et trois polychètes dont le ver de Pompéi *Alvinella pompejana* (Plouviez et al., 2009). L'échantillonnage comprenait les mêmes sites que ceux de cette thèse, plus un site plus au nord (21°N) et trois localités intermédiaires entre 7°25'S et 18°25'S au sud. En estimant la divergence entre populations nord et sud et leurs tailles efficaces, cette étude a permis de dater la séparation des faunes nord/sud à 11,6 Ma pour le gastéropode Lepetodrilus elevatus et à 1,3 millions d'années pour les autres espèces, cette date correspondant à peu près à la mise en place d'un système de failles transformantes dans la région de 4°S (Gofar/Discovery). Une seconde étude multigénique a été entreprise ensuite sur A. pompejana, en utilisant des données de polymorphisme enzymatique, ainsi que des données de polymorphisme nucléotidique de trois gènes codant pour deux enzymes et une globine (PGM-1 ou phosphoglucomutase, SAHH ou S-adénosylhomocysteine hydrolase et la GlobX globine X ; Plouviez et al., 2010). Dans ce cas, l'isolement des populations nord et sud a été estimé plus précisément à 1,8 millions d'années. Cette étude a d'une part, confirmé les conclusions sur la mise en place d'une barrière physique à la dispersion à l'Equateur mais a aussi pu mettre en évidence une semi-perméabilité de cette barrière selon les locus considérés avec la présence d'une zone de contact secondaire à 7°25'S sur l'EPR. En effet, bien que les lignées mitochondriales nord et sud soient arrêtées complètement par la barrière équatoriale, Plouviez et al. (2010) ont montré d'une part, un effet clinal des fréquences allozymiques dans les populations, centré à 7°25'S et, d'autre part, que la SAHH présente un polymorphisme partagé entre populations alors que les locus PGM et GlobX ont des allèles diagnostiques de part et d'autre de la barrière, avec la présence d'un petit nombre d'individus ayant des allèles originaires du nord à 7°25'S. Etant donné le rôle structurant majeur de la barrière équatoriale sur les flux de gènes au niveau du génome mitochondrial et des gènes nucléaires testés, plusieurs questions restent posées :

Quelle est la proportion du génome impactée par cette barrière en termes de différenciation génétique nord/sud ?

Les échanges à travers la barrière sont-ils plus spécifiquement associés à de l'introgression adaptative ?

La différenciation observée correspond-t-elle à un isolement avec une spéciation en cours ou assiste-t-on à un relâchement de la barrière avec remise en contact secondaire ?

-Impact de la barrière équatoriale aux flux de gènes sur la différenciation des populations de part et d'autre de celle-ci

Que l'on s'intéresse à *A. pompejana* ou *A. caudata*, la composition allélique des individus (présence/absence des allèles) permet déjà de différencier les populations selon leur signature nord et sud. Chez *A. caudata*, les individus sont clairement regroupés selon leur origine géographique vis-à-vis de la barrière équatoriale, à l'exception d'un individu avec une signature sud-EPR échantillonné au nord de la barrière. Chez *A. pompejana* la distinction des individus selon leur affiliation nord/sud suit une modalité plus complexe. Les populations nord et sud sont bien séparées mais le degré d'apparentement des individus en fonction de la distance de Jaccard ne suit pas forcément leur positionnement géographique. Logiquement cette divergence se retrouve dans le nombre de locus partagés entre populations de part et d'autre de l'EPR. En effet, dans le cas d'*A. pompejana* on obtient jusqu'à 16 fois moins de locus communs lorsque les populations comparées se trouvent de part et d'autre de l'EPR et à 35 fois moins de locus communs chez *A. caudata* sans que pour l'une des deux espèces on puisse corréler ces différences d'abondance à des différences de couverture entre sous populations.

Si l'on regarde les résultats obtenus sur la base des locus communs entre les populations nord et sud pour les deux espèces (851 pour A. pompejana et 137 pour A. caudata), nos calculs de Fst bien que non significatifs pour A. caudata révèlent des valeurs élevées avec respectivement 0,1764 et 0,2752 pour A. pompejana et A. caudata. En effet, bien que la différenciation génétique des populations nord et sud ne soit pas significativement différente de zéro, chez A. caudata 9,24% des locus sont impliqués dans cette différenciation, et ce nombre passe à 15% chez A. pompejana. Ces données à l'échelle du génome valident les travaux de Plouviez et al. (2010) et donc que la barrière au flux de gènes n'affecte pas l'ensemble du génome de façon identique. Ainsi, alors que nos analyses ainsi que celles de Jang (Jang et al., 2016) et Plouviez (Plouviez et al., 2010) sur A. pompejana montrent que du polymorphisme nucléaire est partagé entre le nord et le sud de l'EPR, on constate une différenciation très marquée des lignées mitochondriales de part et d'autre de la barrière équatoriale aux flux de gènes (Plouviez et al., 2009 ; Plouviez et al., 2010 ; Jang et al., 2016). Les réseaux produits par ces deux dernières études à partir du polymorphisme de la cytochrome oxydase (mtCOI) sont présentés dans la figure III-33. Les analyses multivariées (AFC) et d'assignation des individus à des groupements génétiques (Structure) montrent aussi l'existence de deux entités génétiques au nord et au sud de l'EPR, y compris pour *A. caudata*. Cependant l'assignation de certains individus laisse supposer une perméabilité encore plus grande de cette barrière chez *A. caudata*, avec plus d'individus introgressés par des allèles du sud à 9°50'N. Chez *A. pompejana*, on observe que la proportion d'allèles issus du sud de l'EPR est très faible (<2-3% du génome des individus introgressés) à 9°50'N mais les deux individus échantillonnés à 7°25'S présentent environ 15% d'allèles avec une signature du nord, ce qui indique un flux migratoire unidirectionnel du nord vers le sud. Cette unidirectionalité du flux avait déjà été montré avec la distribution des formes alléliques nord et sud de la PGM-1 chez *A. pompejana* (Plouviez *et al.* 2010). Chez *A. caudata*, cette directionalité du flux apparaît opposée avec une fraction très importante des individus du nord présentant jusqu'à 80% d'allèles avec une signature du sud.

Chez A. caudata, les résultats du logiciel Structure suggèrent la possibilité d'une remise en contact secondaire dans la région de 9°50'N, cette remise en contact étant elle plus spécifiquement observée à 7°25'S chez A. pompejana (présentes données de Structure et résultats de Plouviez et al. 2010, et Jang et al. 2016). L'analyse des séquences de la cytochrome oxydase I d'A. caudata effectuée par S. Plouviez montre une divergence nord/sud plus élevée que celle trouvée chez A. pompejana allant dans le sens d'une séparation plus ancienne des deux lignées géographiques d'A caudata. Aux vues de nos données et de notre échantillonnage des populations, il semble que la zone de contact entre les fonds génétiques de A. caudata se situe donc à 9°50N et plus particulièrement au niveau des cheminées froides. Les données de Structure suggèrent également que les individus d'A. pompejana pourraient présenter un faible pourcentage d'allèles introgressés en provenance du sud sur les cheminées froides de 9°50'N même si les locus incriminés ne sont pas ceux impliqués dans la différenciation entre les populations échantillonnées à 13°N et 9°50'N. Cette analyse comparative entre les deux espèces suggère donc que le flux migratoire à travers la barrière pourrait être plus spécifiquement orienté du nord vers le sud pour A. pompejana et du sud vers le nord pour A. caudata. Cette dynamique pourrait alors expliquer la juxtaposition des deux espèces le long des parois des cheminées hydrothermales après une longue histoire évolutive en allopatrie, l'une des espèces provenant du sud et l'autre du nord de l'EPR suite aux nombreuses réorganisations tectoniques, notamment la dorsale mathématicienne ayant donné lieu aux microplaques de Bauer et de Mendoza, entre 12 et 18 millions d'années (Mammerickx & Klitgord 1982, Eakins & Lonsdale 2003). Néanmoins, cette hypothèse nécessite un échantillonnage beaucoup plus exhaustif des populations d'A. caudata et notamment du site de 7°25'S.

Nos analyses montrent tout de même que chaque localité géographique présente des spécificités et se distingue plus ou moins les unes des autres, bien qu'il ait été montré que les larves d'*A*.

pompejana possèdent des exigences thermiques et d'exposition aux sulfures nécessitant un développement embryonnaire à l'extérieur de la colonie parentale (Pradillon et al., 2005a). Ceci peut s'expliquer, d'une part par le fait que leur durée de dispersion ne semble pas leur permettre de couvrir des distances suffisantes se déplacer d'un champ hydrothermal à un autre (Chevaldonné et al., 1997), et, d'autre part, que les exigences larvaires ne peuvent être satisfaites qu'en s'écartant de la colonie parentale sans pour autant quitter le site actif (Pradillon et al., 2005a).

-L'EPR sud : zone de transition entre l'EPR et les populations de la dorsale Pacifique-Antarctique (PAR)

Au niveau de l'EPR sud, la différenciation entre localités est plus ou moins marquée selon l'information utilisée, i.e. présence/absence des locus ou fréquences alléliques des locus communs des localités 18°25'S et 21°33'S. Aucune différenciation n'a pu être observée entre localités quelle que soit l'approche utilisée. En effet, dès l'analyse sur la base de la présence/absence de locus, l'ensemble des individus du sud montre une relative homogénéité de leur diversité tout en se différenciant des autres populations du nord EPR. Toutefois, l'analyse des données génétiques au sud de l'EPR sur 11 069 locus pour A. pompejana et 4 882 locus pour A. caudata avec le logiciel Structure suggère la présence d'un troisième fond génétique à 21°33'S pour les deux espèces en forçant l'analyse avec K=3. Ce résultat fait écho à l'étude de Jang et collaborateurs en 2016 sur A. pompejana utilisant un nouveau lot d'individus provenant de la ride Pacifique-Antarctique (PAR, pour Pacific-Antarctic Ridge). Ces auteurs identifient en effet trois fonds génétiques distincts sur la base de locus mitochondrial Cox-1 et onze locus nucléaires. Ces fonds génétiques correspondent respectivement au nord et au sud de l'EPR et la ride Pacifique-Antarctique. Toutefois, les diagrammes Structure d'A. pompejana basés sur les 851 locus communs nord/sud ou les 11 069 communs entre localités du sud suggèrent que les deux individus de 7°25'S puissent être des migrants de seconde génération en provenance du nord de l'EPR. D'autre part, l'augmentation du nombre de classes à K=3 dans l'analyse Structure suggère que les individus légèrement introgressés à 21°33'S le sont avec un fond génétique différent de celui de l'EPR nord. Cette observation permet de relativiser les conclusions de Jang et collaborateurs (2016) suggérant que les individus du sud EPR sont doublement introgressés par des allèles provenant du nord EPR et de la dorsale Pacifique-Antarctique sans distinction de localités. Ainsi dans notre cas,

l'introgression d'allèles se limite aux localités proches des populations sources. Il est possible que certains allèles issus du nord EPR ou de la PAR balayent sur une plus large portion du sud EPR, mais ces derniers sont alors trop peu nombreux pour être détectés dans nos analyses.

Cette mise en évidence d'un fond génétique introgressant légèrement les populations du sud EPR est intéressante puisqu'elle permet de mettre en évidence un cas de subdivision des populations avec un fond génétique placé en position intermédiaire et bénéficiant d'apports multiples de la part des fonds génétiques l'entourant. Il serait intéressant de disposer d'un échantillonnage incluant des individus de PAR puisque, selon l'étude de Jang *et al.* (2016), c'est justement cette population qui constitue la source majoritaire de flux de gènes pour le sud EPR. Il serait ainsi peut-être possible d'identifier la ou les zones de contact, et de mettre en évidence une zone d'hybridation entre les populations de la PAR et l'EPR sud. La structure génétique des deux espèces pourrait être alors plus complexe avec le relâchement d'une seconde barrière aux flux de gènes située plus au sud au niveau de la microplaque de l'île de Pâques (Hurtado *et al.* 2004). En effet, le relâchement de cette barrière a conduit à la formation d'une zone d'hybridation chez le bivalve *Bathymodiolus thermophilus* à 23°S (Johnson *et al.* 2013) et des queues d'introgression pouvant s'étendre jusqu'à la barrière équatoriale selon les espèces considérées (Plouviez *et al.* 2013, Jang *et al.* 2016).

De plus, Jang et al. (2016) confirment les dates d'émergence de ces deux barrières précédemment proposées par Plouviez *et al.* (2013) entre les différentes moulières de *B. thermophilus*, avec un premier isolement de populations d'alvinelles lors de la mise en place de la microplaque de l'île de Pâques, il y a 4,2 millions d'années et l'émergence de la barrière équatoriale, il y a 0,79 million d'années. Depuis la remise en contact des populations sud EPR et PAR et du flux plus ou moins unidirectionnel du nord vers le sud à travers la barrière équatoriale, les populations du sud EPR devraient être les plus hétérogènes du point de vue de leur composition allélique. Or nos résultats ne montrent pas de différences significatives de diversité en sud EPR et PAR révélée par Jang et al. sur quelques locus ne traduise pas une réalité à l'échelle du génome.



Figure III-33 : Réseaux d'haplotypes de mtCOI d'*A. pompejana* d'après a, Plouviez *et al.*, 2010 (bleu – individus du nord EPR ; orange – individus du sud EPR); b, Jang *et al.*, 2016 (bleu – individus du nord EPR ; vert – individus du sud EPR ; rouge – individus du PAR).

C - Identification des gènes et fonctions sous sélection diversifiante

L'existence de la barrière équatoriale est confirmée par nos données de polymorphisme le long du génome pour les deux espèces d'Alvinella, mais cette différenciation géographique n'affecte qu'une petite partie de leur génome. En effet, un très grand nombre de locus ne montre pas de différenciation à travers cette barrière, soit parce qu'ils sont presque monomorphes, soit parce qu'ils ne présentent pas de différences de fréquences alléliques entre les deux zones géographiques. A l'inverse, les lignées mitochondriales montrent une démarcation beaucoup plus nette entre les populations nord et sud (Plouviez et al., 2009), ce qui laisse penser que des mécanismes génétiques de verrouillage empêchant la remise en contact des fonds génétiques ou générant de la spéciation sont peut-être à l'œuvre. Parmi ces mécanismes on peut envisager qu'il y ait des incompatibilités cyto-nucléaires puisque mitochondrie et noyaux ne montrent pas la même structure génétique (Plouviez et al., 2010), l'existence d'incompatibilités pré- et postzygotiques limitant ou empêchant la naissance ou la reproduction d'hybrides est aussi envisageable puisque l'on ne retrouve pas d'hybrides de première génération dans nos analyses. Si des incompatibilités pré et post-zygotiques sont en place et que l'isolement reproducteur est achevé et maintenu, les migrants qui parviennent à passer la barrière ne peuvent se reproduire avec le fond génétique local et, on devrait alors assister à un maintien des deux fonds génétiques dans les zones de contact, ce maintien générant un effet Wahlund (mélange géographique d'individus génétiquement différenciés). Ces mélanges se traduisent par une proportion importante de locus présentant des valeurs positives et élevées de Fis traduisant un déficit en hétérozygotes. Or ce n'est pas le cas chez les *Alvinella*. Comme nous l'avons vu dans le cas plus marqué d'*A. pompejana*, seuls 15% des locus identifiés montrent de la différenciation des populations à travers la barrière.

Nous avons donc entrepris une identification des gènes associés aux locus « outliers » dans la différenciation génétique nord et sud des populations afin de savoir si certains gènes impliqués dans les mécanismes de l'isolement reproducteur ou dans la contre-sélection d'hybrides pouvaient avoir évolué sous sélection positive de part et d'autre de la barrière. Cette recherche par Blast nous a permis d'identifier 19 des 112 locus retenus chez *A. pompejana* et, chez *A. caudata*, 2 des 11 locus potentiellement sous sélection positive ou liés à des gènes sous sélection.

Les deux gènes identifiés pour A. caudata (β 1-3 galactosyl-transférase et nucléoside triphosphate phosphatase) interviennent dans des processus du développement embryonnaire. Dans une étude de Porter et Johnson (2002), les auteurs montrent que les gènes du développement, et notamment ceux qui ont un grand nombre d'interactions dans les réseaux de gènes ont un rôle important dans les mécanismes de spéciation. Ainsi de par ce grand nombre d'interactions, le développement en plus de contraindre les processus évolutifs va aussi être une voie privilégiée pour la spéciation même dans des situations où le flux de gènes est important. Le rôle des galactosyl-transférases dans la reconnaissance spermatozoïde/ovocyte a déjà été mis en évidence chez les mammifères et plus particulièrement chez la souris (Swanson et al., 2003 ; Miller et al., 1992). Les nucléosides-transférases, qui sont aussi identifiées par deux fois au niveau des gènes 'outliers' d'A. pompejana, ont été identifiées comme ayant un rôle clé dans l'acquisition ou la perte de capacité à parasiter chez des nématodes à travers des pertes de fonction ou transferts horizontaux, intervenant par la même occasion dans la colonisation de nouvelles niches écologiques et la formation de nouvelles espèces (Zarlenga & Mitreva, 2016). Pour A. pompejana, 16 des 19 locus identifiés ont un rôle métabolique pouvant avoir un rôle important vis-à-vis de l'environnement aéro-anaéobique du ver (e.g. isocitrate déshydrogénase) et participent pour beaucoup aux processus de régulation biologique. Parmi ces gènes, cinq ont attiré notre attention. En plus des nucléosides transférases, l'AMPc phosphodiestérase est une protéine dont certains allèles sont associés à la perte du flagelle chez certaines espèces de bactéries du genre Brucella en relation avec leur mode de vie. Ce genre comprend en effet plusieurs espèces dont certaines sont parasites, et l'hypothèse amenée par Chain et al. (2005) est que cette perte de flagelle pourrait être une réponse adaptative au passage de la vie extracellulaire à intracellulaire de cette bactérie. Cette étude est intéressante si on prend en compte le fait que des observations histologiques des gonades d'A. pompejana et plus précisément l'observation de cellules à noyaux condensés et présentant un grand nombre de mitochondries laisse penser qu'il s'agit de spermatozoïdes et que leur caractère mobile été perdu (Storch & Gaill, 1986). Les travaux sur l'ultrastructure des spermatozoides d'A. pompejana et A. caudata montrent que ces espèces ont des spermatozoides avec un flagelle atrophié (Jouin et al., 2002) pouvant être mis en lien avec le mode de reproduction très particulier de ces organismes où la fécondation se fait par voie interne au niveau de leur spermathèque lors de la ponte (Pradillon & Gaill, 2003). Ce caractère jouant un rôle clé dans cette stratégie reproductive de l'espèce, il est probable que des incompatibilités génétiques au niveau de ce gène puissent entraîner des dysfonctionnements en termes de motilité du spermatozoïde et d'empêcher la rencontre gamétique dans la spermathèque de la femelle. L'acétyl-transférase (Watanabe et al., 2008) et la Fbox (Rieseberg & Blackman, 2010) ont quant à elles été identifiées chez des plantes comme responsables respectivement des souches hybrides non viables dans un cas et une stérilité des mâles hybrides dans l'autre cas. L'acétyltransférase est également identifiée comme gène candidat responsable d'incompatibilités nucléo-cytoplasmiques entre petit-pois de culture et sauvages (Bogdanova et al., 2015). Or, les lignées mitochondriales nord et sud d'A. pompejana présentent une divergence de presque 2% (Plouviez et al., 2010) et pourraient générer ce genre d'incompatibilités. Enfin, le dernier gène ayant attiré notre attention est le gène gld1. Il est caractérisé chez les nématodes comme ayant une action clé dans la différenciation des lignées germinales en mâle ou femelle, l'initiation de la méiose et potentiellement l'apparition de tissus tumoraux (Francis et al., 1995). On peut donc penser que ce gène joue un rôle de verrou dans la formation d'hybrides entre les populations nord et sud en empêchant chez les hybrides le développement normal de la lignée germinale les rendant stériles.

La fonction dans lesquelles ces gènes sont impliqués ne constituent pas une preuve en soi, mais ce sont de véritables indices nous permettant de supposer que la mise en place d'une barrière aux flux de gènes il y a 0,79 - 1,23 millions d'années a été suffisamment durable pour permettre la mise en place de mécanismes impliqués dans l'isolement reproducteur. Effectivement des incompatibilités pré- et post-zygotiques sont possibles puisque que l'on y retrouve à la fois des gènes liés au développement et d'autres impliqués dans la reconnaissance gamétique. Une autre éventualité est l'existence d'incompatibilités nucléo-cytoplasmiques entre les individus du nord et du sud pour les deux espèces *d'Alvinella* puisque ces populations montrent une forte

différenciation de leurs patrimoines mitochondriaux sans mélange de ces lignées de part et d'autre de la barrière.

-Contribution de la géographie et de l'habitat thermique sur la distribution du polymorphisme des deux espèces d'Alvinella au niveau de l'EPR nord

La différenciation entre nord et sud est donc en grande partie générée par une barrière physique aux flux de gènes, mais nous disposons d'indices montrant un renforcement de celle-ci par la mise en place d'une barrière endogène. De l'isolement par la distance a déjà été identifié pour plusieurs espèces de bivalves, mollusques et annélides entre les populations hydrothermales de l'EPR (Vrijenhoek *et al.*, 1998). La question se pose alors quant à l'éventuelle différenciation qu'il peut y avoir entre les localités les plus proches pour nos deux espèces. Etant donné que la barrière équatoriale n'affecte pas de la même manière nos deux espèces en termes de différenciation, on peut supposer que les deux espèces n'ont vraisemblablement pas les mêmes potentiels dispersifs. Ces différences de capacités dispersives doivent donc être perceptibles entre les localités les plus proches que ce soit au nord ou au sud de l'EPR.

Alvinella pompejana :

Les différentes analyses à l'échelle de l'EPR nord sur 13 000 locus ont permis de mettre en évidence un fond génétique nord relativement uniforme sur l'ensemble des individus et une fraction très faible de locus apparentés à un deuxième fond génétique probablement associé à la lignée sud, plus spécifiquement présent dans l'habitat froid du nord de l'EPR. Cette introgression allélique bien que faible est significativement plus importante chez les individus des cheminées froides du champ de 9°50'N. A l'inverse, l'analyse Structure basée sur les 51 locus les plus discriminants entre 9°50'N/13°N (selon le seuil d'hétérozygotie et valeurs les plus élevées de Fst) et communs entre les populations nord et sud permet de distinguer trois fonds génétiques distincts qui discriminent les populations 13°N, 9°50'N et sud EPR. Cette analyse permet aussi d'identifier des individus « hybrides » suggérant un flux de gènes partiel du fond génétique de 13°N chez les individus présents cette fois-ci sur l'habitat chaud de 9°50'N, et du sud EPR dans l'habitat 'chaud' de 13°N.

On retrouve tout de même une distribution particulière des valeurs de Fis prise par les 13 798 locus de l'EPR nord où 7% des locus communs aux deux localités présentent un déficit significatif en hétérozygotes à 13°N ou à 9°50'N. Ces cas de figure correspondent à un fort allélisme diagnostique des deux localités. Lorsque les locus outliers de la différenciation entre

13°N et 9°50'N sont seulement pris en compte, la grande majorité des locus ne présente aucun déséquilibre en hétérozygotes à 13°N mais des excès ou des déficits à 9°50'N pouvant être dus à une introgression d'allèles en provenance de l'EPR sud à 9°50'N, ou à un flux de gènes préférentiel dans le sens 13°N vers 9°50'N.

L'analyse Structure basée sur les 23 locus différenciant les individus issus des deux habitats thermiques et communs entre les populations nord et sud permet de distinguer cette fois-ci deux fonds génétiques distincts, avec l'un présent exclusivement au nord de l'EPR et l'autre quasiment exclusif du sud EPR. On retrouve des individus « hybrides » au nord dans les deux types d'habitat pour la localité 13°N mais uniquement dans l'habitat chaud à 9°50'N. On retrouve en plus un déficit en hétérozygotes dans l'un des deux habitats (3,65% dans l'habitat chaud et 2,68% dans l'habitat froid) qui pourrait, là encore, traduire une sélection par l'habitat ou l'une des deux formes alléliques est contre-sélectionnée dans l'un des deux habitats, notamment l'habitat 'chaud'.

Il est possible en effet que les individus échantillonnés au sud et exposés à des conditions plus chaudes (taux d'accrétion plus élevé) présentent des adaptations favorisant une introgression d'allèles du sud au sein des habitats les plus « chauds » sur l'EPR nord. De plus le contraste que l'on observe entre ces différentes analyses utilisant des catalogues de locus différents et révélant une introgression tantôt favorisée dans un habitat ou dans l'autre, révèle aussi que la sélection sur la lignée sud s'opère différentes selon les gènes considérés.

Alvinella caudata :

Chez *A. caudata* comme chez *A. pompejana*, on observe une fraction de locus présentant un excès d'hétérozygotes dans les deux populations de l'EPR nord, ceux-ci pouvant être le fruit d'introgression adaptative avec sélection balancée en faveur des hétérozygotes ou de duplications de gènes non filtrés. Chez *A. caudata*, cependant on observe un schéma inverse à la distribution des hétérozygotes observée chez *pompejana* dans les deux types d'habitats. En effet, alors que l'on devine une différenciation génétique des individus vivant sur les deux types de cheminées avec l'information apportée par les locus communs entre les populations nord et sud, celle-ci disparaît lorsque l'on se focalise sur les 4 018 locus communs entre habitats. Comme nous l'avons déjà vu, il est très probable qu'un grand nombre d'erreurs de séquençage et de polymorphisme non informatif ait été maintenu dans notre jeu de données. Dans ce cas le filtrage des données sur l'hétérozygotie et les valeurs élevées de Fst ne permet pas d'augmenter la résolution des analyses AFC ou Structure. Malgré cela on observe tout de même avec les 137 locus communs au nord et au sud une introgression exclusive du fond génétique sud au sein de

celui du nord au niveau des habitats froids comme dans le cas d'A. pompejana. La distribution des valeurs moyennes de Fis est centrée sur zéro mais on retrouve une proportion non négligeable de locus présentant des valeurs élevées proches de 1. L'analyse des valeurs de Fis prises par ces locus dans chaque population séparément permet de voir que ces déficits en hétérozygotes sont en fin de compte spécifiques à l'une ou l'autre des populations (2,60% dans l'habitat chaud et 2,02% dans l'habitat froid). Cependant, à l'inverse d'A. pompejana, la plupart des déséquilibres en hétérozygotes sont plus spécifiquement associés aux locus « outliers » de la différenciation des individus par l'habitat (la distribution des Fis dans les locus « outliers » n'est pas centrée sur zéro mais s'écarte de zéro). Cette distribution particulière des locus « outliers » de la différenciation par l'habitat renforce donc notre hypothèse d'une introgression adaptative des allèles du sud chez l'espèce A. caudata. Cette différenciation plus marquée des individus d'A. caudata en fonction de l'habitat thermique pourrait également expliquer le polymorphisme observé sur la longueur des branchies chez cette espèce. En effet, on retrouve une plus forte proportion d'individus à branchies courtes dans les habitats plus froids (D. Jollivet, comm. pers.), et une étude plus spécifique entre le phénotype branchial et le génotype multilocus est désormais nécessaire pour mieux comprendre la part de la sélection environnementale dans la distribution des différentes lignées géographiques.

La différenciation de nos deux espèces entre le nord et le sud de l'EPR semble donc dépendante de gènes verrous empêchant l'hybridation entre fonds génétiques selon le type d'habitat thermique rencontré et de la localité. Il s'agit donc d'un cas complexe où la différenciation se fait simultanément sur la base de la géographie via l'isolement par la distance, et sur la base de l'habitat avec de la sélection différentielle jouant localement sur de l'introgression allélique en provenance de l'EPR sud. Les zones hybrides peuvent être vues comme des arènes dans lesquelles les recombinaisons entre allèles sont testées fonctionnellement dans chaque population pour des combinaisons de mutations ayant émergées en allopatrie (Payseur, 2010). Il semble que ces deux types d'environnement constituent des arènes dans lesquelles les pressions de sélection s'opérant sur les hybrides ou tout du moins les combinaisons mutationnelles sont différentes.

Dans ce genre de situation, l'idéal serait de produire des hybrides et estimer la valeur sélective des types parentaux et des hybrides dans différentes conditions thermo-chimiques rencontrées par l'espèce pour rechercher les incompatibilités pré- ou post-zygotiques (Malfant *et al.*, 2017) et, ainsi faire la part entre les composantes endogènes et exogènes de cette barrière aux flux de gènes. Puis de cartographier ces locus « outliers » sur un génome plus complet afin d'identifier des relations de proximité entre ces locus et faire la part des choses entre ce qui est de

l'adaptation locale et ce qui n'est que le résultat de l'isolement géographique (Fraïsse *et al.*, 2014).

-Un gradient latitudinal des habitats thermiques lié au taux de renouvellement des sites ?

Le taux d'accrétion de la dorsale Est Pacifique augmente du nord au sud et, conduit à un gradient latitudinal de la fréquence des habitats thermiques allant d'habitats globalement plus « chauds » et plus éphémères au sud vers des habitats plus « froids » et pérennes au nord (Francheteau & Ballard, 1983). Or plusieurs espèces montrent des traces d'une barrière aux flux de gènes probablement due à une succession de failles transformantes qui perturbent les courants de fond dans cette zone. Bierne et collaborateurs ont montré en 2011 que les barrières endogènes générées par des incompatibilités intrinsèques aux espèces (incompatibilités pré- et post-zygotiques) ont tendance à s'aligner avec le temps sur les barrières exogènes liées soit à des zones de moindre migration, soit à des gradients d'habitats. Dans le cas d'*A. pompejana*, les données sur l'allélisme de la PGM-1 montre l'existence d'un polymorphisme adaptatif lié à la fréquence des habitats chauds et froids (Piccino *et al.*, 2004). On peut donc penser qu'un cline environnemental lié à une augmentation latitudinale des habitats chauds du nord au sud avec les contraintes physiologiques qu'il entraîne puisse être calé sur la zone de moindre migration au niveau de la barrière équatoriale (Plouviez et al., 2010), ce qui pourrait amener à une distribution de fonds génétiques (Figure III-34).



Figure III-34 : Pourcentage des fonds génétiques (PAR : ride pacifique antarctique ; SEPR : sud de la ride du pacifique oriental ; NEPR : nord de la ride du pacifique oriental) composant les populations d'*A. pompejana* le long de l'EPR séparées par des barrières aux flux de gènes (traits pointillés rouge dont l'intervalle entre les traits traduit la perméabilité). Le tout est mis en parallèle avec le taux d'accrétion de la dorsale (données extraites de Francheteau & Ballard, 1983) marqué par les points bleus et extrapolés entre ces mesures (trait fin continu bleu).

On retrouve donc dans ce scénario, un taux d'accrétion croissant de 21°N à 20°S, qui constitue un proxy du taux d'extinction et de la fréquence des habitats chauds. On y retrouve les trois fonds génétiques d'*A. pompejana* correspondant aux trois provinces PAR, sud EPR et nord EPR avec, au sein de l'EPR nord, les deux fonds génétiques mis en évidence dans ce travail qui pourrait être associés aux deux lignées mitochondriales précédemment identifiées au nord (Plouviez *et al.* 2013, Jang *et al.* 2016) et pour lesquels un séquençage du Cox-1 nous permettrait de vérifier cette hypothèse.

Au final, les lignées mitochondriales diagnostiques du nord et du sud EPR (Plouviez *et al.,* 2010) et la différenciation génomique que nous avons mise en évidence entre ces lignées laisse penser que les flux de gènes sont fortement limités par des incompatibilités nucléocytoplasmiques constituant une barrière endogène alignée sur une zone de très faible migration potentiellement couplée à un cline d'habitats.

5 - Conclusion

Alvinella pompejana et *A. caudata* sont deux espèces que l'on retrouve tout le long de la dorsale Est Pacifique (EPR) et cette aire de répartition couvre aussi une partie de la ride du Pacifique-Antarctique. La dynamique temporelle de l'environnement hydrothermal ainsi que son organisation dans l'espace en font un cas d'étude privilégié pour comprendre le rôle des mécanismes de dispersion et d'adaptation dans la différenciation locale des populations.

Notre étude a confirmé l'effet structurant de la barrière équatoriale aux flux de gènes entre les populations nord et sud pour nos deux espèces. Cette limitation des échanges n'est pas absolue et semble être en partie liée aux capacités de dispersion des espèces, ce qui suggère des causes en partie physiques (distances entre populations trop grandes, courants divergents, perte des larves au niveau des failles transformantes, durée de vie larvaire, ...) à ces échanges restreints amenant à de la spéciation par allopatrie. Bien qu'il y ait une relative perméabilité de cette barrière, les fonds génétiques du nord EPR et du sud EPR se mélangent difficilement. Différents mécanismes génétiques sont envisageables au vu de nos résultats.

Etant donnés les gènes impliqués dans la différenciation nord/sud des populations, des incompatibilités pré- et post-zygotiques sont possibles puisque que l'on y retrouve à la fois des gènes liés au développement et un identifié chez une autre espèce comme intervenant dans la reconnaissance gamétique. Une autre éventualité est l'existence d'incompatibilités nucléo-

cytoplasmiques entre les individus du nord et du sud pour *A. pompejana* puisque ces populations montrent une forte différenciation de leurs patrimoines mitochondriaux sans mélange de ceux-ci dans les populations.

On observe aussi une différenciation des individus occupant des types d'habitats thermiquement différents. Bien que cette différenciation soit faible et ne porte que sur un petit nombre de locus, on constate néanmoins que pour nos deux espèces, une introgression préférentielle des allèles sud dans l'habitat le plus froid des cheminées hydrothermales. Cette sélection différentielle n'opère donc pas spécifiquement sur du polymorphisme existant initialement dans la population nord, mais agit comme un filtre permettant on non l'établissement d'un patrimoine génétique extérieur dont l'arrivée est déjà en grande partie limitée par une barrière géographique et potentiellement par des incompatibilités nucléocytoplasmiques. Néanmoins, une partie de la différenciation génétique retrouvée entre habitats thermiques est associée à des locus diagnostiques du fond génétique nord et à des différences génétiques plus spécifiquement trouvées entre 9°50'N et 13°N. Il est donc possible que dans le nord EPR, l'hétérogénéité génétique locale des populations puisse résulter à la fois d'une sélection par l'habitat et de barrières endogènes régulant le flux de gènes en provenance du sud EPR, voire même avec des populations situées plus au nord (21°N ou bassin de Guaymas). De même que les différences génétiques observées entre habitats, la différenciation entre localités du sud suit un peu le principe de l'effet de bord. Pour les deux espèces, les populations 18°25'S et 21°33'S se distinguent par des « pools » de gènes introgressés d'origine différentes et correspondant très probablement à de la migration longue distance issue des fonds génétiques différents du nord EPR et PAR. Ainsi, la différenciation entre populations d'Alvinella se fait sur la base d'un polymorphisme préexistant partitionné géographiquement sur la base d'évènements d'isolement répétés et conditionnés par l'histoire tectonique récente de la dorsale. Cependant, ces flux de gènes entre localités pourraient être modulés selon la nature des habitats et leur fréquence dans les différentes zones géographiques.

IV - Structure, connectivité et diversité des populations d'*Harmothoe fuligineum* en Antarctique

1 - Introduction

A - Température et adaptabilité des espèces

La température est l'un des paramètres majeurs contrôlant la distribution des espèces sur le globe. Ce paramètre affecte à la fois la tolérance physiologique et les interactions écologiques, en particulier pour les espèces ectothermes, ne pouvant pas compenser les variations de température. A l'heure actuelle, la réalité du réchauffement climatique est entendue unanimement dans la communauté scientifique, et les modèles visant à prévoir l'évolution de divers paramètres environnementaux tels que la température, la pluviométrie, la sécheresse ou le taux de CO₂ se multiplient (Cox *et al.*, 2000 ; Hansen *et al.*, 2000 ; Dai, 2013). Une partie de ces modèles se base sur différents scénarios définis par l'IPCC (comité intergouvernemental sur le changement climatique) prenant en compte des mesures quantitatives des populations humaines à des échelles régionales (développement économique, efficacité énergétique, disponibilité énergétique, productivité agraire, contrôle local de la pollution, etc ...).

En corrélation avec ce réchauffement, de nombreuses espèces ont déjà commencé une migration vers les pôles ou vers de plus hautes altitudes (Parmesan and Yohe, 2003), avec un déplacement atteignant 50 km par décennie en moyenne pour l'océan (Helmuth *et al.*, 2006). Ainsi certaines espèces marines littorales se sont déplacées et on a notamment observé l'entrée d'espèces tropicales dans la Méditerranée (Lejeusne et al., 2009). De même, sur la côte Ouest des Etats Unis, la limite entre deux espèces de patelles du genre *Lottia*, l'une Nord et l'autre Sud, a migré de 75 km vers le Nord en l'espace de 20 ans. Cette différence d'aire de répartition est de plus corrélée à la stabilité thermique d'une enzyme étudiée comme indicateur de l'adaptation thermique (Dong et Somero, 2009). Dans l'Atlantique Nord-Est, la température moyenne des eaux de surface a augmenté de 0,23°C par décennie entre 1977 et 2007 (Goikoetxea et al 2009). Au cours de la même période dans cette région, la moule *Mytilus edulis* et la telline *Macoma balthica* ont migré, respectivement, de 1000 et 300 km vers le Nord (Berge et al 2005 ; Jansen et al 2007). Dans ce contexte, la capacité de dispersion des espèces aura une grande importance dans le changement de leurs aires de répartition (Malcolm *et al.*, 2002 ;

Hijmans & Graham, 2006 ; Pearson, 2006). Ces déplacements dans le temps ou l'espace consistent en fait à l'exploitation d'une nouvelle niche favorable et à l'abandon d'une ancienne devenue quant à elle défavorable (Pauls *et al.*, 2013).

Si les espèces ayant des aires de répartition continue sur une large gamme de latitudes peuvent modifier leur distribution, des barrières physiques (masses de terres, profondeur non adaptée) ou biotiques (prédation, compétition, parasitisme) peuvent s'opposer à leur déplacement. Dans le cas des espèces de haute altitude ou des espèces polaires, ce sont les zones climatiques ellesmêmes qui vont disparaître. Dans ce cas, deux scénarios sont possibles : soit la connectivité avec d'autres populations de la même espèce porteuses d'allèles plus adaptés aux conditions chaudes permet le passage de ces allèles vers la population qui fait l'expérience d'une augmentation de température (disparition d'allèles d'adaptation locale froide), soit la capacité de survie de l'espèce dépendra de son adaptabilité aux nouvelles conditions, et donc de son polymorphisme pré-existant ou de l'apparition de nouvelles mutations (Barrett & Schluter, 2008). Pour ce qui est de l'apparition de nouvelles mutations, le taux de mutation et la taille efficace de la population sont les paramètres clé puisqu'ils traduisent la probabilité qu'un évènement mutationnel arrive dans un laps de temps donné. Ce taux de mutation peut varier selon les gènes et leur position le long des chromosomes (Taddei et al., 1997) et dépend également de l'efficacité des mécanismes de réparation de l'ADN (Sniegowski et al., 1997). Cependant, le temps de génération est aussi très important pour que la fréquence allélique change dans la population et pour que l'adaptation soit efficace (voir Peck, 2011 pour une revue). Ainsi, pour une espèce marine avec un temps de génération de quelques années, une mutation ne pourra se répandre dans la population (augmentation de fréquence) que sur plusieurs décennies au moins. Cette échelle de temps n'est pas compatible avec la vitesse de réchauffement actuel, d'autant plus que la température affecte potentiellement tout le génome et que la probabilité d'apparition de mutations bénéfiques sur chaque locus est très faible.

L'adaptabilité aux conditions changeantes peut aussi s'envisager sous l'angle de nouvelles combinaisons d'allèles favorisant une meilleur 'fitness'. Si on utilise la représentation de l'environnement comme un paysage adaptatif (Wright, 1982), les mutations peuvent induire un déplacement au sein de ce paysage dont l'altitude traduit la valeur sélective de l'individu pour une combinaison génétique donnée dans cet environnement. Selon l'intensité des forces de sélection en cours, le paysage va être plus ou moins abrupte et rendre plus ou moins facile un déplacement d'un pic adaptatif à un autre. Ainsi un moyen d'accélérer le déplacement au sein de ce paysage est l'acquisition d'un polymorphisme existant dans une autre population par flux génique. Ces associations entre fonds génétiques distincts vont donc littéralement permettre des

sauts au sein de ce paysage dont le résultat sera aléatoire. Une étude s'est penchée sur la connectivité entre populations comme réponse rapide à la variation environnementale (Kremer *et al.*, 2012) et a ainsi montré que bien que cette connectivité puisse engendrer des mal-adaptations, sous certaines conditions elle peut favoriser une réponse adaptative de celle-ci aux changements climatiques. En effet dans le cas du changement climatique, des pics adaptatifs préexistants disparaissent et de nouveaux apparaissent, ainsi ces croissements entre fonds génétiques peuvent permettre à une espèce d'atteindre plus rapidement un nouveau pic adaptatif qui aurait nécessité un temps plus important avec l'apparition de nouvelles mutations.

B - Histoire géologique et évolution de la température en Antarctique

L'Antarctique, et en particulier la péninsule Antarctique, sont étudiés dans le cadre du réchauffement climatique en cours puisque c'est au sein de cette province entre autres qu'on y observe les plus fortes hausses de température (Vaughan *et al.*, 2003). En effet, le réchauffement global n'affecte pas toutes les régions du globe de façon homogène. Ainsi, alors que la hausse moyenne globale de température depuis 1950 est d'environ 0,5°C, la Péninsule Antarctique a, quant à elle, connu une hausse de 1,5°C (Houghton *et al.*, 2001).

Au cours de son histoire, le continent Antarctique a cependant connu de plus importantes variations climatiques mais sur de plus longues échelles de temps. Depuis environ 70 millions d'années le continent Antarctique occupe sa position géographique actuelle. Avant cette période, il se trouvait sous des latitudes avec un climat tempéré. En étudiant les rapports isotopiques de l'oxygène (∂^{18} O) benthique, Lisiecki et Raymo (2005) ont montré que les températures n'ont pas diminué graduellement mais sont restées tempérées jusqu'à environ 40 millions d'années (Eocène, Figure IV-1A). A cette période, la concentration de CO2 atmosphérique chute et la température globale moyenne diminue. Plus tard au cours de l'Eocène, le passage de Drake entre la pointe Sud de l'Amérique du Sud et la Péninsule Antarctique s'ouvre, permettant ainsi la mise en place du courant circum-polaire antarctique (CCA ; Barker & Thomas, 2004). Ce courant, combiné avec les températures froide associées aux faibles concentrations atmosphériques de CO₂, permettent la première glaciation du continent Antarctique à l'entrée dans l'Oligocène (33,9 Millions d'années ; Figure IV.1A). Avant la fin de l'Oligocène, l'Antarctique fond et reste ainsi jusqu'à environ 15 millions d'années (Miocène), où la reglaciation de l'Antarctique a lieu (Lisiecki et Raymo, 2005). Depuis, bien que la température varie, le continent reste dans l'ensemble gelé (Figure IV-1A et B). Nous avons cependant des cycles glaciaires importants qui voient alterner des périodes

d'extension et rétraction de la glace de mer. Entre 1 et 3 millions d'années, ces cycles se sont alternés à une fréquence de 41 000 ans et au cours du dernier million d'années, le cycle a connu une fréquence d'environ 100 000 ans (Lisiecki et Raymo, 2005).



Figure IV-17 : Evolution de la température moyenne en Antarctique depuis 65 millions d'années (A) et vue plus précise pour les dernières 5 millions d'années. Températures estimées à partir de données isotopiques ¹⁸O de carottes de sédiment. D'après Lisiecki et Raymo (2005).

Ces reconstructions sur la période la plus récente sont confirmées par des mesures isotopiques de l'oxygène piégé dans des carottes de glace permettant de remonter à environ -420 000 ans (Figure IV-2 ; Petit *et al.*, 1999). A partir de cette reconstruction, on peut voir que la température de surface en Antarctique n'a jamais excédé 3°C au cours de cette période. Cette série temporelle permet aussi de mettre en évidence les variations de température engendrées par 4 aires glaciaires qui se sont succédées ces 420 000 dernières années. De par l'isolement du continent, cette succession d'aires glaciaires et l'extension de la glace qu'elles ont engendrée à profondément influencé la diversité génétique des espèces en générant des isolements et des remises en contact répétés favorisant la formation d'espèces cryptiques (Rogers, 2007). Lors de l'extension de la glace, les espèces ont pris refuge dans des zones à la limite de la glace comme les îles subantarctiques. La séparation physique résultante a permis l'émergence d'espèces par vicariance et leur mise en contact secondaire après la fonte des glaces. Ainsi, en dépit de
l'apparente homogénéité de l'environnement antarctique, on peut observer une structure géographique marquée pour la distribution des espèces, la sympatrie d'espèces proches (et parfois cryptiques) et la possibilité de formation d'hybrides.



Figure IV-2 : Reconstitution de différents paramètres environnementaux de l'air continental sur plus de 400 000 ans en Antarctique à partir de prélèvements de glace (Petit et al., 1999).

Chez une espèce de crinoïde présentant une apparente distribution circumpolaire Antarctique, une étude moléculaire a permis de mettre en évidence (i) l'existence de deux espèces cryptiques vivant en sympatrie, (ii) des histoires évolutives différentes avec l'une des deux espèces présentant une grande diversité de sous populations, (iii) des niveaux d'endémisme et de connectivité entre populations différents pour les deux espèces (Hemery *et al.*, 2012). A l'inverse d'autre espèces montrent quant à elles une diversité génétique marquée par les glaciations passées, mais aucune structuration des populations autour de l'Antarctique (Raupach *et al.*, 2010). Ainsi il apparait capital d'avoir une idée claire de la structuration des populations d'une espèce Antarctique avant d'envisager une quelconque étude impliquant des mesures de diversité.

C - Variabilité contemporaine de la température en Antarctique

L'apparente homogénéité de l'Antarctique cache aussi des variations contemporaines. Malgré une relative stabilité de la température sur l'ensemble des eaux côtières du continent Antarctique avec une très faible saisonnalité de la température des eaux de surface, des différences existent entre les provinces (Figure IV-3). L'amplitude thermique annuelle (estimée à partir de données couvrant la période de décembre 1981 à avril 2005) au sein même de l'Antarctique peut être inférieure à 0,4°C pour la côte à 180°W (Mer de Ross), et atteindre 3,2°C au niveau de la Péninsule. Pour ces mêmes zones, la température moyenne annuelle passe de 0°C à presque 4°C, respectivement, entre 1981 et 2005. Cette différence est d'autant plus marquée si l'on compare le continent aux îles sub-antarctiques situées au niveau du front polaire délimité par le courant circumpolaire antarctique (Barnes *et al.*, 2006).

Bien que faible, cette variabilité thermique entre localités suggère que les espèces ayant une distribution circum-Antarctique et occupant aussi les îles périphériques pourraient avoir un potentiel adaptatif supérieur aux espèces dont les distributions sont plus restreintes. D'autre part des espèces présentant une structuration en rapport avec ces différentes localités pourraient montrer de l'adaptation locale, et posséder la capacité de coloniser d'autres régions plus froides après leur réchauffement. Elles pourraient aussi s'hybrider avec les populations déjà présentes dans ces zones.



Figure IV-3: A) Clines d'amplitude annuelle de la température des eaux côtières autour de l'Antarctique (10-20m de profondeur) estimées à partir de données couvrant la période de Décembre 1981 à Avril 2005 ; B) Température moyenne des eaux de surface de l'océan austral estimées à partir de données couvrant la période de Décembre 1981 à Avril 2005. Extrait de Barnes et al. (2006).

Enfin, la température en Antarctique montre aussi des différences en fonction de la profondeur. En effet, le littoral Antarctique présente un gradient salin et thermique marqué de 0 à 150 m de profondeur (Dierssen et al., 2002). En surface l'eau de mer est impactée par les échanges thermiques avec l'atmosphère, la formation de glace concentrant les sels dans l'eau liquide et le mélange avec l'eau douce apportée par les précipitations, ainsi l'eau de surface présente typiquement des températures oscillant entre -1,8 et 1°C et une salinité entre 33 et 33,7 ‰. Lors de la formation de glace en surface, les sels sont exclus et l'eau enrichie en sel et froide, plus dense, coule et induit un apport constant d'eau froide et salée en profondeur. Ceci conduit à une relative stabilité des paramètres de l'eau à 100m de profondeur avec une température beaucoup plus stable aux environs de -1,5°C et une forte salinité entre 33,8 et 34,0‰. A faible profondeur, on peut aussi observer des variations à plus haute fréquence. Près de la base de Dumont d'Urville, un capteur de température à haute fréquence placé à 7 m de profondeur permet notamment de mettre en évidence non seulement une variation saisonnière sur l'année pouvant atteindre 2°C mais aussi des variations journalières à fréquence calée sur les marées et pouvant atteindre 1°C en période estivale avec débâcle et environ 0,05°C le reste de l'année (Figure IV.4).



Figure IV-4: Variations de température depuis 2006, enregistrées dans le cadre du programme Nivmer à 7 m de profondeur près de la Station de Dumont d'Uvrille (Terre Adélie). Le graphique en dessous offre une vue plus précise sur 7 mois couvrant l'été austral 2016-2017, avec une débâcle le 7 Décembre. D'après Hourdez et al., in prep.

D - Choix du modèle Polynoidae Harmothoe fuligineum

La famille des Polynoidae est bien représentée en Antarctique, avec pas moins de 28 lignées COI sur la côte du continent (Cowart et al., in prep). Harmothoe fuligineum est, de loin, la plus abondante espèce et présente en différents points de l'Antarctique, de part et d'autre de la péninsule ainsi que dans les provinces de l'Est et de l'Ouest Antarctique (Barnich et al., 2006 ; Pabis & Sicinski, 2010; Brasier et al., 2016; Neal et al., 2017; Cowart et al., in prep). H. fuligineum comme d'autres espèces du même genre se retrouve à différentes profondeurs, depuis le milieu sublittoral (à moins de 50m de profondeur) jusqu'à plus de 400m de profondeur (Pabis & Sicinski, 2010). On dispose de peu d'informations spécifiques quant aux traits d'histoire de vie de H. fuligineum ou sur la biologie des Polynoidae en général (Gambi et al., 2002). Parmi les annélides polychètes, la famille des Polynoidae est l'une des plus fécondes (McHugh & Fong, 2002). De plus, des observations plus précises de plusieurs espèces de Polynoidae Antarctiques suggèrent une relative homogénéité comportementale inter-spécifique avec parfois une conservation brève d'œufs sous les élytres des vers femelles durant une courte période jusqu'à les libérer en grand nombre et avec un développement passant par une phase larvaire planctonique pélagique durant l'été austral (Gambi et al., 2002). Ainsi, Harmothoe fuligineum se retrouve dans des localités soumises à des variations thermiques différentes et colonise des profondeurs présentant elles aussi des profils thermo-salins contrastés. De plus elle présente des traits morphologiques suggérant une grande capacité de dispersion lui permettant éventuellement une connectivité entre localités à grande échelle. Ceci fait donc d'H. fuligineum une espèce idéale pour étudier la structuration des espèces Antarctiques, la connectivité entre populations et éventuellement se faire une idée sur ses capacités à s'adapter au réchauffement climatique déjà en cours. Pour appréhender ce sujet nous nous attacherons à répondre aux questions suivantes :

L'espèce est-elle une espèce pan-Antarctique avec des échanges continus entre la partie Ouest (Péninsule en cours de réchauffement) et la partie Est (Terre Adélie) ?

Existe-t-il des échanges possibles avec les îles Kerguelen situées au-delà du front polaire ? Existe-t-il une structuration génétique en lien avec l'étagement superficiel (environ 20m de profondeur) et profond (150 m) des populations et quelles en sont les causes ?

Dans la mesure où ces populations seraient génétiquement différenciées, présentent-elles des diversités génétiques différentes pouvant traduire des histoires démographiques et/ou écologiques différentes ?

2 - Matériel et Méthodes

Tout comme pour le chapitre précédent, l'ensemble des techniques de laboratoire et de traitement des données RAD pour l'analyse de la structure génétique des populations de *H*. *fuligineum* sont présentées dans le chapitre 2 de cette thèse.

A - Echantillonnage

Pour cette étude, les populations de *H. fuligineum* ont été échantillonnées dans deux régions opposées de l'Antarctique (Péninsule et Terre Adélie) et les îles Kerguelen situées plus au nord dans l'océan austral. L'échantillonnage de l'espèce a été réalisé à différentes profondeurs. En effet, une partie de la population de Terre Adélie et quelques individus de Péninsule ont été échantillonnés sur les stipes de grandes algues *Himantothallus grandifolius* à faible profondeur (moins de 20m) en plongée. Pour les autres échantillons de Terre Adélie ainsi qu'en Péninsule et à Kerguelen, les individus ont été récoltés par dragage ou par chalutage (chalut à perche) à des profondeurs situées généralement entre 50 et 150 m.

Les échantillons de Terre Adélie ont été obtenus au cours des missions POLARIS, KREVET, et REVOLTA. Les échantillons de Péninsule ont été récoltés en face de la base anglaise de Rothera par M. Clark et L. Peck du British Antarctic Survey ou au cours de la mission Allemande ANTXXIX-3 (collaboration M. Eléaume) et celui de Kerguelen (1 seul individu identifié à l'espèce sur la base du COI) à partir des collections du MNHN obtenues lors des campagnes PROTEKER dans le Golfe du Morbihan (avec la collaboration de M. Eléaume). L'ensemble des régions citées sont représentées sur la carte d'Antarctique en Figure IV-5.

Des études préliminaires nous ont amenés à ne prendre en compte la distinction sur la base de l'étagement des populations que dans le cas des individus de Terre Adélie, ceux de la Péninsule ne montrant pas de différence génétique en fonction de la profondeur.

Chapitre IV - Structure, connectivité et diversité des populations d'Harmothoe fuligineum en Antarctique



Figure IV-5 : Carte de l'Antarctique présentant les trois localités étudiées ainsi que l'étagement pris en compte en Terre Adélie.

B - Assignation populationnelle

GeneClass2 (Piry et al., 2004) est un programme qui, à partir d'un échantillonnage aléatoire des allèles dans chaque population, simule des génotypes multi-locus parentaux (1000 génotypes dans notre cas) et détermine ainsi une valeur seuil d'identité des individus à la population de référence à partir de ces génotypes simulés. Il va ensuite en se basant sur l'identité génotypique des individus à tester, calculer une probabilité d'assignement sur la fonction de vraisemblance qui sera considérée comme significative si elle est supérieure à la valeur seuil obtenue pour cette population.

3 - Résultats

A - Contrôle qualité des banques

Tout comme pour l'étude sur *Alvinella*, après le séquençage par la plateforme de génomique de l'Université McGill au Québec, les données nous ont été envoyées déjà démultiplexées pour les 12 index Illumina utilisés (12 fichiers compressés contenant les données de séquençage appariées R1 et R2) et filtrées sur la base des Phred scores (score de qualité). Le démultiplexage

des données sur la base des barcodes (interne aux lectures) au niveau de chaque index nous a ensuite permis de les attribuer à chaque individu et ainsi d'évaluer l'abondance relative de lectures entre individus (cf. Figure IV-6).

De plus le nombre de lectures (*reads*) ayant passé ou non le filtre pour chaque individu nous donne une première idée de la qualité des banques construites. En moyenne, sur 2 057 774 lectures par individu avant filtration (présence de la séquence des adaptateurs dans les lectures, barecodes erronés, sites de restriction non reconnus), 1 317 583 lectures sont conservées soit 64% des lectures initiales.



Figure IV-6 : Abondance relative du nombre de reads par individu sachant que l'individu KER0 (Kerguelen) est le plus couvert en nombre de reads après démultiplexage avec 8 379 463 lectures et une moyenne individuelle de 1 317 583 lectures. (PEN-Péninsule ; TAD-Terre Adélie profond ; TAS-Terre Adélie superficielle)

B - Construction des catalogues de locus

Comme nous ne disposions que d'un transcriptome de référence pour cette espèce, après cette étape, deux possibilités se sont présentées : (1) construire un catalogue en *de novo* (sans génome de référence) ou (2) en ne considérant que les locus alignés sur le transcriptome de référence et

donc associés à des séquences codantes pouvant être ensuite intéressantes à identifier dans le cas où certains locus pourraient être sous sélection diversifiante. Deux catalogues de locus ont donc été construits et les caractéristiques de ces catalogues sont présentées dans la figure IV-7. Dans le catalogue construit en *de novo*, l'appartenance d'un allèle à un locus a été défini selon le critère d'une différence maximale de 10 bases entre deux séquences et en exigeant une couverture de 3 lectures par allèle afin de s'assurer une certaine confiance quant à la fiabilité des séquences identifiées.



Figure IV-7 : (a) Rapport entre les nombres de snps obtenus entre les locus cartographiés sur transcriptome et en de novo pour chaque individu. (b) Nombre de snps obtenu par individu en de novo (bleu) et présents sur le transcriptome (rouge). (c) ratio du nombre de snps sur le nombre de locus identifiés pour chaque individu à partir des données en de novo (bleu) et des données mappées (rouge).

On constate tout d'abord une diminution de l'écart entre les nombres de locus obtenus en *de novo* et sur données mappées dans les catalogues individuels lorsque le nombre initial de lectures obtenues pour un individu donné est faible (Figures IV-6 IV-7 a.). En revanche, pour les individus relativement bien couverts, le rapport entre données mappées et *de novo* se situe aux alentours de 0,3 (70% de locus dans le non-codant).

Comme cela était déjà le cas avec la quantité de lectures brutes après séquençage, on constate une forte disparité du nombre de snps récupérés entre individus quelle que soit la technique de construction de catalogue choisie (Figure IV-7 b.). Lorsque l'on regarde le nombre moyen de snps par locus et par individu, on constate que les individus les moins couverts présentent des moyennes légèrement inférieures à celles des autres individus dont la moyenne tourne autour de 3,2 snps par locus, quelle que soit la technique considérée (Figure IV-7 c.). A la vue de ces résultats nous avons décidé de ne conserver pour la suite des analyses que les individus pour lesquels on obtenait au moins 10 000 SNPs à partir de la construction de catalogue en *de novo*, soit 50% des individus.

Comme nous l'avons dit précédemment, alors que la construction de catalogue en *de novo* couvre l'ensemble du génome, celle basée sur le mapping sur transcriptome se limite à l'information portée par le génome codant. Pour cette raison nous avons réalisé un mapping du catalogue obtenu en *de novo* sur le transcriptome de référence utilisé et en conservant les mêmes critères d'alignement. Ceci nous permet de se faire une idée quant à la redondance de l'information apportée par ces deux catalogues. Les résultats de cet alignement sont présentés dans le Tableau IV-1.

total	2352240	100%
alignés aucune fois	2294904	97,56%
alignés exactement 1 fois	52785	2,24%
alignés plus d'une fois	4551	0,19%

Table IV-5 : Résultats de l'alignement du catalogue obtenu en de novo sur le transcriptome de référence de *Harmothoe fuligineum* à l'aide du programme Bowtie2 en End-to-end.

Le taux d'alignement n'est que de 2,44%. Il est donc possible que le type d'alignement « Endto-End » avec Bowtie2 et surtout le système d'attribution de score d'alignement induise une conservation ou à l'inverse le rejet de certaines séquences conservées dans le catalogue *de novo* basé sur la comparaison deux à deux au sein des lectures elles-mêmes. Ainsi il semble intéressant d'étudier chacun de ces deux jeux de données séparément afin d'avoir une vue d'ensemble de la connectivité entre populations étant donné que chacun traduit le comportement de deux parties distinctes du génome.

C - Définition des populations analysées

Sur la base de l'échantillonnage dont nous disposions, nous avons dans un premier temps cherché à identifier les locus communs entre paires de populations ou entre nos trois échantillons Antarctiques et ce pour chacun des jeux de données, en acceptant 20% de données manquantes dans chaque population. L'obtention de ces locus a permis de calculer à l'aide du programme Genepop les valeurs de Fst entre les 3 populations et par paires de populations et d'estimer la significativité de la différenciation génétique de ces populations par des tests exacts de Fisher sur les données de fréquences alléliques. L'ensemble de ces valeurs est présenté dans la figure IV-8.



Figure IV-8 : (A) Nombre de locus communs entre populations et (B) valeurs du Fst moyen par locus calculées sur la base de ces mêmes locus, *significativité à 0.1% les autres valeurs n'étant pas significative à 0.5%.

On peut constater de façon surprenante que bien que la quantité de locus par individu soit supérieure avec les catalogues *de novo*, le nombre de locus en commun identifiés entre populations est plus grand lorsque l'on considère les données mappées. Il est possible que la converture réelle soit nettement inférieure à la couverture potentielle des sites possibles (PstI MseI encadrant une centaine de paires de bases) ce qui engendre une faible couverture populationnelle en termes d'individus. On peut penser que ce phénomène soit moins marqué dans le cas des données mappées sur transcriptome en raison d'une accessibilité physique supérieure du génome codant limitant ce phénomène (Bell et al., 2011). Ensuite si l'on classe

les ensembles de populations de celui qui partage le plus de locus à celui qui en partage le moins, l'ordre reste le même quel que soit le jeu de données. En effet les populations profondes de Terre Adélie et de Péninsule (sans distinction de profondeur) sont celles partageant le plus de locus alors qu'elles représentent les populations les plus éloignées géographiquement.

Les valeurs de différenciation entre paires de populations (Fst), quant à elles, ne sont pas corrélées au nombre de locus partagés entre populations. Seules les comparaisons incluant les trois populations ou la paire Péninsule/Terre Adélie superficielle montrent des valeurs de Fst significatives avec respectivement 0,0405 et 0,1522.

Etant donné le faible nombre de locus communs entre populations et la différenciation génétique peu marquée entre populations dans la plupart des cas, il a été décidé de réaliser une première analyse sur la base de la présence / absence des locus. Cette analyse a deux objectifs majeurs : dans un premier temps exploiter l'information apportée par l'ensemble des données communes dans le jeu de données, et dans un deuxième temps de valider ou rejeter les *a priori* de regroupements des individus en populations que nous avons effectués sur la base de leur co-localisation spatiale.

Cette analyse a été réalisée sur les deux jeux de données et une distance de Jaccard a été calculée entre paires d'individus et retranscrite à travers un positionnement multidimensionnel (MDS pour « multidimensional scaling » en anglais) présenté dans la figure IV-9.

La différence observée sur le nombre de locus communs partagés ou en termes de Fst entre la population superficielle de Terre Adélie (TAS) et les autres populations est retrouvée dans des graphiques pour les deux jeux de données. Dans les deux cas, on observe deux groupes d'individus distincts, l'un constitué des individus de TAS et l'autre composé des deux autres populations antarctiques ainsi que de l'individu de Kerguelen. La distinction est tout de même plus nette dans le cas de l'analyse basée sur les données obtenues en *de novo*. En effet l'analyse basées sur les données mappées montre l'exitence d'individus en positions intermédiaires pouvant suggérer l'existence d'un léger flux de gènes observable uniquement dans le génome codant et donc potentiellement adaptatif.

Lorsque les individus du groupe TAS sont supprimés de l'analyse pour améliorer la discrimination au sein de l'autre groupe, une légère structuration apparaît avec un groupe compact d'individus et quelques individus plus dispersés (figure IV-7 b). Ces individus plus isolés correspondent aux individus les moins couverts en termes de lectures et constituent vraisemblablement des positionnements artéfactuels liés au nombre important de données manquantes. Au sein du groupe le plus dense, on remarque tout de même que l'individu de Kerguelen est retrouvé à l'intérieur du groupe formé par les individus de Péninsule, alors que

les individus de la population profonde de Terre Adélie (TAD) forment un groupe légèrement décalé au sein de cet ensemble déjà cohérent.



Figure IV-9 : Positionnements multidimensionnels obtenus à partir de matrices de distances de Jaccard sur les données obtenues en *de novo* (2 100 354 locus) ou après cartographie sur le transcriptome (1 024 077 locus) incluant l'ensemble des populations y compris l'individu de Kerguelen et dans un second temps pour chaque jeu de données sans les individus de la population superficielle de Terre Adélie.

Après cette première analyse visant à identifier des grandes tendances, d'autres analyses centrées sur les locus communs entre populations ont été réalisées et cette fois-ci en prenant en compte l'information apportée par les fréquences alléliques. Ainsi quatre lots d'analyses ont été menés, tout d'abord en prenant en compte les trois populations, puis entre paires de populations.

A partir de ces informations nous avons donc réalisé des analyses d'assignation Structure ainsi que des AFC sur les fréquences alléliques afin de confirmer la structure des individus en 2 groupes génétiquement différenciés. Les résultats utilisant l'ensemble des individus des trois populations sont présentés dans la figure IV-10.



Figure IV-10 : Résultats des AFC et analyses Structure obtenus sur la base des 40 et 543 locus communs entre les trois populations (TAS-Terre Adélie superficiel ; TAD-Terre Adélie profond ; PEN-Péninsule) respectivement obtenus dans les jeux de données de novo (a.) et sur données mappées (b.).

Dans les deux cas, les analyses Structure ne permettent pas de distinguer de groupes génétiquement différenciés, cependant des traces d'un fond génétique distinct sont observables chez quelques individus de Terre Adélie tous étagements confondus mais elles n'excèdent pas 10% du génome. Cette apparente homogénéité se retrouve aussi dans les deux AFCs. Les trois premiers axes de l'AFC générée à partir des données *de novo* expliquent respectivement

16,13%, 12,57% et 12,31% de la variance des fréquences alléliques observée dans le jeu de données. En ce qui concerne les données mappées, les trois premiers axes expliquent 12,49%, 8,62% et 8,17% de cette même variance. Dans les deux cas, tous les individus sont regroupés dans une seule et unique entité génétique de laquelle s'éloignent quelques individus de Terre Adélie profond et superficiel constituant l'essentiel de la variance entre individus.

D - Assignation de l'individu de Kerguelen à une population continentale

Une fois les populations identifiées, caractérisées et la différenciation entre celles-ci estimée, il est intéressant de chercher à assigner l'individu échantillonné à Kerguelen aux populations antarctiques. Ces analyses ont uniquement été réalisées sur les jeux de données mappées qui présentent plus de locus et donnent plus de poids statistique aux assignations ; et en utilisant l'information apportée par les locus communs des différents ensembles de populations définis précédemment. Les probabilités d'appartenance obtenues pour chaque jeu de locus sont présentées dans le tableau IV-2. Quelles que soient les probabilités d'appartenance, on remarque que dans tous les cas les probabilités dépassent de façon significative les valeurs seuil (non présentées par le programme) même si elles ne dépassent jamais la valeur de 0,95 qui nous assurerait l'appartenance de l'individu de Kerguelen à l'une de ces populations.

Quel que soit le jeu de données utilisé on peut voir que la population TAS est systématiquement la moins probable des populations proposées. La population PEN est à l'inverse la plus probable comme population d'origine mais la différence avec la population TAD n'est pas très importante. Des tests d'assignation des individus appartenant à ces trois localités donnaient des probalités d'assignation supérieures à 0,95 quelle que soit l'origine de l'individu et la population testée, ce qui montre bien la divergence existant entre Péninsule et continent.

Table IV-6 : Probabilités d'assignation de l'individu de Kerguelen estimées par GeneClass2 à chaque population Antarctique pour chaque jeu de données avec en code couleur le vert correspondant aux probabilités les plus proches de 1 et en rouge celles proches de 0. La dernière colonne indique le nombre de locus ayant permis le calcul des probabilités pour chaque lot d'analyses.

	TAS	TAD	PEN	Nombre de locus	
3 рор	0,291	0,606	0,725	463	
PEN / TAD		0,516	0,717	1413	
TAD / TAS	0,321	0,585		456	
PEN / TAS	0,171		0,664	634	

E - Diversité génétique des trois échantillons de H. fuligineum

Nous avons voulu également savoir si ces populations présentaient des différences du point de vue de leur diversité génétique pour savoir si la faible différenciation génétique observée entre TAD et PEN pouvait être uniquement due à une réduction du polymorphisme dans l'une des deux régions. En effet, le réchauffement climatique se fait déjà ressentir en Péninsule alors que la Terre Adélie n'est pas encore affectée de façon notable. Pour cela nous avons cartographié et quantifié la diversité génétique non-synonyme en fonction de la diversité synonyme pour ces trois populations et le résultat de cette analyse est présenté dans la figure IV-11.

La seule différence significative observée l'est au niveau du polymorphisme non-synonyme entre les populations PEN et TAD. On observe tout de même une tendance montrant un polymorphisme globalement plus important de la population PEN en comparaison avec les deux autres de Terre Adélie, avec une moyenne de polymorphisme non-synonyme plus de deux fois supérieure aux deux autres populations. De plus, on observe une valeur en Terre Adélie légèrement supérieure pour la population superficielle. Lorsque l'on s'intéresse au polymorphisme synonyme, les trois populations ne présentent pas de différences mais on observe le même ordre de classement des populations de la plus diversifiée à la moins différenciée que celui trouvé avec le polymorphisme non-synonyme avec un positionnement intermédiaire de TAS entre TAD et PEN. En plus de la variabilité de polymorphisme qu'il peut y avoir entre les locus utilisés pour chaque population, le nombre de locus utilisés pour l'analyse peut aussi impacter l'estimation de la diversité. Pour cette raison une estimation de la diversité intra- et inter-populations a été réalisée et sur la base des 543 locus mappés communs aux trois populations grâce au programme Genepop (Tableau IV-3).

Qintra et Qinter traduisent respectivement la probabilité de retrouver deux fois le même allèle chez deux individus d'une même population ou de deux populations différentes. Ainsi le calcul 1-Q traduit quant à lui la probabilité inverse et constitue donc un proxy de la diversité existant au sein d'une population ou entre deux. Le calcul réalisé sur les mêmes locus et pour les trois populations montre donc une moins grande disparité que l'analyse précédente. Les trois populations présentent des valeurs faibles, avec par ordre croissant la population Péninsule, suivie des populations profonde et superficielle de Terre Adélie. Ces valeurs sont encore plus faibles lorsque l'on compare les populations à l'ensemble, ce qui laisse penser que le polymorphisme intra-populationnel n'est pas privé pour ces 543 locus.



Figure IV-11 : Polymorphisme synonyme et non synonyme pour les trois populations PEN (Péninsule Antarctique), TAD (Terre Adélie profond) et TAS (Terre Adélie superficiel) à partir de données mappées sur transcriptome et ayant utilisé respectivement des locus RAD associés à 130, 24 et 48 transcrits de ce transcriptome.

Table IV-3 : Diversités intra- et inter-populationnelles calculées pour les 543 locus communsaux trois populations avec le programme Genepop.

	1 - Qintrapopulationnelle	1 - Qinterpopulationnelle
TAS	0,2467	0,1687
TAD	0,2309	0,1577
PEN	0,2293	0,1447

F - Etude de la différenciation génétique des populations locus par locus et rôle de l'étagement/éloignement géographique dans cette différenciation.

-Recherche de locus 'outliers' associés à de la différenciation par l'habitat.

Etant données les valeurs de différenciation moyennes observées entre les deux populations profondes (PEN et TAD) et la population superficielle (TAS), la question s'est posée de savoir si les mêmes locus servaient de support à la différenciation entre chaque paire de populations. Pour ce faire, les locus communs aux trois populations ont été utilisés pour, dans un premier temps, calculer les valeurs de Fst de chaque locus entre paires de populations et, dans un second

temps, les représenter en fonction de la valeur de Fst qu'ils prenaient avec une autre paire de populations. Cette analyse a été réalisée sur les deux jeux de données (*de novo* et mappé) et les résultats de celle-ci sont présentés dans la figure IV-12.



Figure IV-12 : Représentation des valeurs de Fst pour chaque locus commun aux trois populations (TAD, TAS et PEN) croisées entre chaque paire de populations. Les points oranges et les points bleus correspondent respectivement aux valeurs calculées pour les locus des catalogues *de novo* et données cartographiées.

Comme l'on pouvait déjà le voir avec les valeurs de Fst moyennes calculées entre populations, la tendance générale révélée par les valeurs des Fst pour tous les locus montre une proportion importante de locus présentant des valeurs de Fst élevées (> 25 locus « outliers ») lorsque l'on compare l'une des deux populations profondes (TAD ou PEN) avec la population de Terre Adélie superficielle (TAS). Ainsi, alors que les locus présentant le plus de différenciation pour la paire TAD-PEN n'excèdent pas 0,2, une vingtaine de locus présentent des valeurs de Fst supérieures à 0,25 lorsque l'on compare l'une de ces deux populations à TAS, et ces locus

incriminés sont les mêmes entre paires de comparaisons entre sites profonds et celui proche de la surface.

De plus, si l'on s'intéresse à la corrélation des valeurs de Fst pour les deux paires TAS-TAD et TAS-PEN, on constate une forte corrélation positive (coefficient directeur 0,9937) très bien corrélé (R²=0,9802) entre ces valeurs pour le jeu de données obtenu en *de novo*. Et bien que son coefficient de corrélation soit moins grand (R²=0,5869) mais significatif à hauteur de 5% (test de Student pour 510 mesures), la corrélation observée entre la paire TAS-TAD et la paire TAS-PEN a un coefficient directeur de 0,7903. De plus, bien que la différence observée soit faible, il semble que le jeu de données mappées sur le transcriptome permette de mettre en évidence une plus forte différenciation de la paire TAS-PEN que de TAS-TAD. Il semble aussi que pour les locus avec Fst inférieurs à 0,5 dans ces deux paires, la majorité présente des valeurs de Fst supérieures dans la paire TAS-PEN et se retrouve donc au-dessus de la droite de corrélation ce qui pourrait expliquer que seule la différenciation TAS-PEN présente une valeur de Fst moyen significative.

-Distribution du polymorphisme entre les populations

Nous nous sommes intéressés à la distribution des locus communs entre populations selon leurs valeurs de Fis moyen au sein des deux populations et leurs valeurs de Fst entre populations. De plus le comportement de ces locus a été catégorisé par le programme Lositan sur la base de leurs valeurs de Fst et d'hétérozygotie afin d'identifier ceux qui ont un comportement « outlier » ou s'écartant significativement d'une distribution attendue par un modèle en îles avec un taux de migration moyen entre populations. Les locus ont ainsi été classés en trois catégories, ceux au comportement neutre, ceux identifiés comme sous sélection positive et enfin ceux sous sélection balancée. Dans la figure IV-13 sont représentées les valeurs de Fst des locus communs aux deux populations de Terre Adélie en fonction de leur déséquilibre en hétérozygotes (Fis) au sein des deux populations.



Figure IV-13 : Bi-plot des valeurs de Fis et de Fst pour tous les 543 locus communs aux populations de Terre Adélie (profonde et superficielle ; TAD et TAS). Les locus sont identifiés selon trois catégories (neutre - bleu, sous sélection positive - rouge et sous sélection balancée - vert) obtenues à l'aide du programme Lositan simulant des valeurs de Fst dans un modèle de populations en îles. Les diagrammes de densité (en %) présentent la distribution des locus selon leurs valeurs de Fst (gauche) et de Fis pour tous les locus (haut rouge) ou seulement ceux identifiés comme sous sélection positive (haut vert).

Le long de l'axe des ordonnées, le diagramme de densité des valeurs de Fst montre qu'une grande partie des locus est majoritairement centrée sur 0 avec tout de même une queue de distribution assez étendue de locus avec des valeurs supérieures à 0,2. Le même type de diagramme de densité du Fis nous montre que l'essentiel des valeurs de Fis est centré sur 0 avec cependant une nette dissymétrie dans le nombre de locus présentant des valeurs négatives s'étalant sur toute la gamme de -1 à 0. Et si l'on concentre notre attention sur la distribution des locus participant potentiellement à la différenciation des deux habitats (locus sous sélection positive), on remarque qu'à l'exception de quelques rares locus, l'ensemble des locus 'outliers' présente des valeurs de Fis négatives correspondant à un excès en hétérozygotes (dans au moins l'un des deux habitats). La figure IV-14 correspond à la relation entre valeurs de Fis (moyenné sur les deux habitats) et Fst entre les deux populations de Terre Adélie des locus identifiés sous sélection positive avec une valeur de Fis négative. Ce zoom sur ces locus particuliers nous

permet d'identifier une corrélation négative (coefficient directeur de -0.4236 et coefficient de corrélation R^2 de 0,3874) entre Fis moyen et Fst ce qui renforce l'idée que cette distribution des Fis moyens concentrée sur des valeurs négatives correspond à une forte sélection positive locale.



Figure IV-14 : Bi-plot entre Fis et Fst pour les locus communs aux deux populations de Terre Adélie profonde et superficielle identifiés comme sous sélection positive par le programme Lositan et présentant des valeurs de Fis négatives.

Pour comparaison, la même analyse a été réalisée entre les populations Terre Adélie profond (TAD) et Péninsule (PEN) sachant que ces populations ne présentent pas de différenciation génétique malgré l'éloignement géographique. Les résultats sont présentés dans la figure IV-15. La première observation est que la distribution des valeurs de Fst est toujours centrée sur 0 mais les valeurs les plus élevées sont inférieures à 0,4 et s'écartent peu de la moyenne interlocus. Pour les valeurs de Fis, l'ensemble des locus présente une distribution extrêmement resserrée autour de 0 sans aucune déviation apparente. En revanche lorsque l'on regarde la distribution des valeurs de Fis pour les quelques locus outliers identifiés, ces valeurs sont plus ou moins distribuées sans tendance évidente dans des valeurs positives ou négatives.



Figure IV-15 : Bi-plot des valeurs de Fis et Fst pour tous les 1678 locus communs aux deux populations profondes de Terre Adélie et de Péninsule. Les locus sont assignés à trois catégories (neutre, sous sélection positive et sous sélection balancée) obtenues à l'aide du programme Lositan simulant des valeurs attendues sous l'hypothèse d'un modèle en îles. Les diagrammes de densité présentent la distribution des locus selon leurs valeurs de Fst (gauche) et de Fis pour tous les locus (haut rouge) ou seulement ceux identifiés comme sous sélection positive (haut vert).

Dans le cas des deux populations de Terre Adélie se caractérisant par un étagement différent et montrant la plus forte différenciation génétique, on peut se poser la question de savoir si l'un des habitats favorise l'excès en hétérozygotes. Ainsi la figure IV-16 (a.) place chaque locus commun à ces deux populations selon leurs valeurs de Fis au sein de la population profonde et superficielle respectivement en abscisse et en ordonnée.

Les locus y sont encore une fois distingués selon qu'ils sont identifiés comme neutres, sous sélection positive ou balancée. Afin de ne prendre en compte que les locus présentant des valeurs de Fis significativement différentes de 0, nous avons représenté l'ensemble des P-value (valeur de probabilité de ne pas être significativement différent de 0) en fonction des valeurs de Fis elles-mêmes. Ainsi on peut voir dans la figure IV-16 (b.) qu'au-delà de 0,75 ou en deçà de



-0,75, la majorité des valeurs de Fis sont significatives. Le tableau des densités en Figure IV-16 (c.) présente donc le pourcentage de l'ensemble des locus compris dans chaque intervalle.

Figure IV-16 : (a) Représentation des valeurs de Fis associées aux 543 locus communs aux deux populations de Terre Adélie (profonde en abscisses et superficielle en ordonnées). (b) Représentation des P-values associées à la significativité des déséquilibres en hétérozygotes calculées pour chaque locus (Fis de Weir & Cockerham) permettant de définir la valeur de Fis=0,75 comme seuil pour lequel ces estimations présentent des p-value inférieures à 0,15. (c) Tableau traduisant la répartition des locus selon leurs valeurs de Fis dans chaque population dans les intervalles définissant les valeurs de Fis significativement différentes de 0 pour chaque population (<-0,75 et >0,75).

On peut ainsi voir que 27,4% des locus communs aux deux populations présentent des valeurs de Fis significativement négatives dans au moins l'une des deux populations, et 14,9% dans les deux populations à la fois. Aucun locus ne présente de valeur de Fis significativement supérieure à 0 dans les deux populations à la fois, même si 0,4 et 2,2% des locus montrent un déficit significatif en hétérozygotes (Fis>0) respectivement dans les populations superficielle et profonde de Terre Adélie. Et enfin les 70,0% de locus restants ne présentent pas d'écart significatif à 0 dans les deux populations à la fois.

Enfin si l'on regarde de plus près les locus identifiés comme sous sélection positive, la majorité de ceux-ci présentent de faibles valeurs de Fis ne s'écartant pas significativement de zéro. Seulement trois de ces locus présentent des valeurs significativement négatives dont une fixée à -1 dans les deux populations ; ce qui laisse penser à un cas de gènes dupliqués. L'analyse en détail des fichiers ayant permis cette analyse nous permet d'accéder aux fréquences génotypiques populationnelles pour chaque locus et la figure IV-17 présente ces fréquences pour les 4 locus identifiés comme sous sélection positive avec des valeurs de Fis significativement différentes de 0 dans au moins l'un des deux habitats. On peut ainsi confirmer le fait que le locus présentant un Fis de -1 dans les deux populations correspond sûrement à un cas de paralogie puisque l'ensemble des individus tous habitats confondus sont identifiés comme hétérozygote. Les deux autres locus présentant des valeurs de Fis significativement négatives uniquement dans l'habitat profond. Pour le locus identifié en bleu dans la figure IV-18, on ne retrouve que des individus homozygotes dans l'habitat superficiel alors que dans l'habitat profond tous sont hétérozygotes. Pour le locus identifié en vert la situation est plus nuancée avec un type d'homozygotes propre à chaque population mais malgré tout une forte proportion d'hétérozygotes dans les deux populations suggérant une sélection différentielle sur l'un des 2 allèles selon l'habitat considéré. Enfin le locus présentant un excès d'homozygotes dans la population profonde, ne se retrouve qu'à l'état homozygote dans chaque population mais avec des allèles différents. Seule la population profonde présente les 2 allèles à l'état homozygote, cette distribution du polymorphisme laisse penser qu'il existe une forte sélection contre les hétérozygotes tout habitat confondu. Ces quatre locus montrent donc des traces de sélection par l'habitat, mais ceci ne semble affecter qu'une très faible proportion des locus étudiés.



Figure IV-17 : Fréquences génotypiques relatives pour les 5 locus identifiés comme outliers sous sélection positive par Lositan et présentant des valeurs de Fis s'écartant significativement de 0 dans au moins l'un des deux habitats.

-Annotation des locus sous sélection positive selon l'étagement et l'aire géographique

Afin d'identifier les processus biologiques concernés dans la différenciation au sein des deux paires de localités (TAD et PEN) et d'étagement (TAD et TAS), une recherche a été faite pour récupérer l'annotation associée aux transcrits correspondant aux locus identifiés sous sélection positive. Une fois les séquences récupérées, un blast a été réalisé par le biais du site NCBI et l'ensemble des résultats est présenté dans le Tableau IV-4.

Aucun des gènes ou des domaines identifiés n'est commun aux deux comparaisons. En effet dans le cas de la comparaison entre profondeurs en Terre Adélie, on identifie des fonctions associées au stress thermique comme l'HSP 70 (Bukau & Horwich, 1998) et à la réparation de l'ADN avec l'ARN hélicase (Gorbalenya *et al.*, 1989) laissant penser que la différenciation se fait en partie en réponse aux différences de variabilité thermique entre ces habitats. Il est aussi intéressant que l'ARN 18S (sous-unité 40S du ribosome) soit impliquée dans la différenciation entre les localités, en effet cette séquence est connue pour son taux d'évolution relativement lent faisant de lui une cible privilégiée pour les études de phylogénie ancienne (Field *et al.*, 1988), ce qui nous permet de supposer que cette différenciation selon l'étagement découle d'un évènement de divergence ancien potentiellement en lien avec les évènements glaciaires passés.

Dans le cas de la différenciation entre localités, les gènes et fonctions identifiés ne semblent pas permettre d'identifier une tendance générale puisque l'on retrouve à la fois du métabolisme (adénosinetriphosphatase), de la communication cellulaire (domaine couplé à protéine G), un domaine fréquemment retrouvé dans des protéines et jouant un rôle clé dans les interactions protéine-protéine (tétratricopeptide, Blatch & Lässle, 1999). En revanche la cytochrome-b a été identifiée chez le copépode *Microarthridion*, un lien entre 3 lignées haplotypiques de cette protéine et la tolérance au stress salin (Staton *et al.*, 2002). Il est donc possible que malgré la forte homogénéité des fonds génétiques de ces deux localités, il y ait de l'adaptation locale en lien avec la salinité entre autres.

Table IV-7 : Annotation des protéines ou fonctions associées aux locus sous sélection positive différenciant les paires de populations TAD – PEN et TAD – TAS.

	Paire de populations	
Gène ou fonction identifiée	TAD - PEN	TAD - TAS
cytochrome-b5 réductase	1	
domaine couplé à protéine G	X 2	
adénosinetriphosphatase	1	
tétratricopeptide	1	
domaine de liaison ATP à localisation membranaire	1	
sous-unité II RNA polymérase		1
carboxypeptidase		1
glucuronosyltransférase		1
40s ribosome 18s		1
ARN hélicase ATP dépendante		1
HSP70		1

4 - Discussion

A - Connectivité autour de l'Antarctique et avec les îles sub-antarctiques

L'Antarctique de par le courant circum-polaire qui le ceinture se trouve dans une situation d'isolement thermique et biologique. Mais ce courant suit un mouvement bien particulier (figure IV-19). Ainsi la délimitation entre les eaux Antarctiques et les eaux sub-Antarctiques est marquée par le front polaire Antarctique (marqué en rouge dans la figure) et place ainsi les Kerguelen à la limite des eaux Antarctiques. La convergence sub-tropicale marque quant à elle la séparation entre les eaux sub-Antarctiques et les trois grands océans (Atlantique, Indien et Pacifique). Le courant circum-polaire Antarctique (ACC dans la figure IV-19) à proprement

parler tourne dans le sens horaire alors que le courant côtier polaire (ACoC dans la figure IV-18) tourne lui dans le sens anti-horaire.



Figure IV-18 : Carte de l'Antarctique indiquant la localisation du front polaire Antarctique (Antarctic polar front), de la convergence sub-tropicale (Subtropical Convergence), les limites de la calotte glacière hivernale actuelle (Modern winter sea ice) et lors du dernier maximum glaciaire (LGM winter sea ice), le sens du courant circumpolaire (ACC) et côtier (ACoC) ainsi que l'emplacement des îles sub-Antarctiques dont les Kerguelen (KG). Extrait de Fraser et al., 2012.

L'étude des courants et de la biologie des larves permet de prédire au mieux la connectivité entre les populations marines. Les populations d'Antarctique sont isolées par la courant circumpolaire des autres océans à la fois thermiquement mais aussi du point de vue génétique (barrière potentielle au flux de larves). Mais il faut prendre en compte des paramètres comportementaux tels que la rétention larvaire et des paramètres physiques comme des micro-variations environnementales pouvant induire une mortalité larvaire afin de ne pas surestimer les capacités de dispersion des organismes (Cowen et al., 2000). Le développement larvaire d'Harmothoe fuligineum n'est pas bien connu, bien que des observations suggèrent un développement passant par une phase larvaire planctonique pélagique au moins durant l'été austral (Gambi et al., 2002). Ainsi, en théorie, Harmothoe fuligineum peut potentiellement disperser sur de grandes distances étant donné qu'elle produit une grande quantité de petits œufs (produisant des larves planctonotrophes) et que le courant circum-polaire peut constituer un vecteur privilégié pour le transport de ces larves. Il serait alors possible que des larves en provenance de la Péninsule (et soumises à des variations de températures plus importantes que le reste de l'Antarctique), colonisent des territoires pour le moment plus froids et encore épargnés par le réchauffement global comme la Terre Adélie par exemple. De plus les îles sub-antarctiques comme les Iles Kerguelen constituent elles aussi un territoire source d'une potentielle diversité génétique pour répondre au réchauffement climatique. Leur position intermédiaire entre eaux antarctiques et sub-antarctiques rend la connectivité avec le continent encore plus complexe et difficile à prédire étant donné l'isolement que peut aussi constituer le courant circumpolaire entre ces zones. Cependant, des données de modélisation de dispersion en Antarctique suggèrent que des échanges sont possibles depuis le continent vers certaines îles subantarctiques (Brasier et al., 2017). Il faut tout de même noter le fait que n'ayant trouvé qu'un seul individu de cette espèce lors de la campagne d'échantillonnage, soit la méthode d'échantillonnage n'était pas adaptée à cette population, soit il est possible que cette espèce y soit rare ce qui pourrait s'expliquer par une présence résultant de l'arrivée occasionnelle de larves continentales.

Pour se faire une idée de cette connectivité, une étude de la structuration des populations entre localités et de la diversité les caractérisant le cas échéant nous permettra de savoir si les localités pourront par le biais d'introgression adaptative de fonds génétiques plus adaptés ou plus diversifiés répondre au changement climatique.

Les jeux de données dont nous disposons ne montrent pas une forte structuration entre la Terre Adélie et la Péninsule Antarctique. Effectivement les analyses basées sur la composition en locus des individus montrent un fort regroupement des individus de Péninsule et de Kerguelen avec une distribution plus étalée pour les individus de Terre Adélie mais tout de même centrée sur ce même groupe. Mais la résolution de ce genre d'analyse est fortement impactée par les données manquantes, or la population de Péninsule et l'individu de Kerguelen sont justement les mieux couverts. Malgré cela, la composition en locus traduit une forte homogénéité des populations. Et cette homogénéité se retrouve aussi lorsque que l'on compare les fréquences alléliques des populations sur la base des 543 (cartographiés sur le transcriptome) et 40 (obtenus en de novo) locus communs. En effet les AFCs génèrent le même genre de patron que celui obtenu précédemment, c'est-à-dire un regroupement plus resserré des individus pour la population de Péninsule que pour ceux de Terre Adélie. En revanche l'analyse Structure permet de mettre en évidence des traces d'un second fond génétique au niveau de Terre Adélie n'excédant pas 10%. Nos échantillonnages ne nous permettent pas de nous prononcer quant à l'origine de ce fond génétique, et donc de dire si celui-ci constitue une évolution spécifique de certains locus en Terre Adélie ou les traces d'une introgression en provenance d'une autre localité. Ainsi Péninsule et Terre Adélie présentent un fond génétique similaire ne permettant pas de mettre en évidence une quelconque différenciation entre populations pour l'ensemble des locus étudiés. Ces résultats suggèrent donc soit l'existence d'échanges contemporains entre ces localités, soit que les tailles efficaces des populations sont suffisamment grandes pour limiter les effets de dérive. Mais pour savoir si l'échange de migrants entre populations différentiellement soumis au réchauffement global peut être source d'adaptation il faut se concentrer sur les locus au cas par cas.

L'analyse locus par locus de la différenciation entre populations nous révèle deux choses. D'une part la distribution des valeurs de Fst montre une faible différenciation de l'ensemble des locus lorsque l'on compare la population de Péninsule à celle de Terre Adélie profonde. En effet, dans l'ensemble, les valeurs de Fst sont inférieures à 0,2. D'autre part, et à l'inverse, les patrons de distribution de Fst pour les comparaisons PEN/TAS et TAD/TAS présentent une queue de distribution beaucoup plus étalée allant jusqu'à 0,8. De plus, qu'il s'agisse des données obtenues par mapping ou en *de novo*, il existe une forte corrélation entre les valeurs de Fst dans les deux comparaisons (PEN/TAS), ce qui montre que les locus impliqués dans la différenciation avec TAS sont les mêmes lorsque cette population est comparée à la population de Péninsule et à celle de Terre Adélie profonde. Il semble donc que l'étagement soit une source plus grande de différenciation que l'éloignement géographique. Cette différence sera étudiée plus précisément par la suite, et nous nous concentrerons donc désormais sur la comparaison opposant Péninsule et Terre Adélie profonde pour estimer les apports que peuvent constituer des échanges entre ces localités aux profils environnementaux et à l'histoire évolutive récente contrastés.

L'étude plus précise des 1678 locus communs à la Péninsule et la Terre Adélie profonde renforce l'idée d'une différenciation très peu marquée entre ces localités. En effet la distribution

des valeurs de Fis moyen et des Fst entre ces deux populations montre une distribution doublement centrée et resserrée sur zéro. L'augmentation du nombre de locus permet tout de même de montrer que la queue de distribution des valeurs de Fst s'étalant jusqu'à 0,4 pour quelques rares locus. L'essentiel de ces locus est constitué des locus déjà identifiés comme « outlier » sous sélection positive par Lositan. Les valeurs de Fis moyen associées à ces locus en particulier ne montrent pas de tendance particulière et semblent se distribuer aléatoirement entre -1 et 1 ce qui laisse penser que la différenciation très légère entre ces localités se fait de façon relativement neutre.

La plupart des gènes ou fonctions incriminés interviennent dans diverses fonctions métaboliques ou structurelles laissant penser que l'essentiel de cette faible divergence se fait de façon neutre, mais la différence entre ces deux localités est en grande partie marquée par le fait que la Péninsule montre déjà des marques du changement climatique, avec une augmentation de la température moyenne de l'eau de mer de surface de 1°C (Meredith & King, 2005). En plus du réchauffement que cela entraine, des mesures ont permis de montrer une hausse de la salinité depuis 1955, notamment dans les eaux de surface (Meredith & King, 2005). Or l'un des gènes identifiés, la cytochrome-b réductase est associée à la capacité de survie d'un copépode en réponse à un stress salin (Staton et al., 2002). Dans le cas d'Harmothoe fuligineum, la différenciation moyenne observée entre la Péninsule et la Terre Adélie est faible mais les valeurs de Fst pour certains locus sont élevées. La différenciation sur ces locus pourrait être due à l'effet de la température et de sa variabilité : pas de différenciation par la profondeur en Péninsule et variabilité saisonnière marquée (3°C) alors que le site profond en Terre Adélie ne présente pas de variation saisonnière. Il est aussi possible que la hausse de salinité observée depuis ces trente dernières années implique une réponse adaptative à la hausse de salinité en Péninsule. Ceci pourrait représenter une source de polymorphisme adaptatif pour le changement futur en Terre Adélie.

L'étude visant à affilier l'individu échantillonné aux Kerguelen à l'une des populations continentales ne traduit pas d'appartenance directe puisque les probabilités d'appartenance n'excèdent pas la valeur seuil de 0,90. Bien que cet individu ait une plus grande probabilité d'appartenir à la population de Péninsule, ou de Terre Adélie profonde, il pourrait aussi être un représentant d'une population de petite taille limitée aux îles subantarctiques. Cet individu correspond en effet au seul *H. fuligineum* échantillonné parmi la centaine d'individus génotypés de Kerguelen alors que c'est de loin l'espèce la plus abondante autour du continent (Cowart et al., in prep). Un échantillonnage plus important sur ces localités pourrait permettre de déterminer si les populations subantarctiques sont réellement isolées. En revanche on peut

affirmer que cet individu présente une signature génétique différente de celles présentes en Péninsule et en Terre Adélie tout étagement confondu. Ainsi on peut raisonnablement affirmer que dans l'éventualité où un flux de gènes existerait entre ces populations, un apport de diversité permettant au moins la survie au niveau des Kerguelen actuellement plus chaudes que les autres provinces continentales, pourrait permettre l'adaptation de ces populations au réchauffement déjà observable en Péninsule.

B - Impact des glaciations passées sur l'histoire évolutive de l'espèce

La biodiversité actuelle de l'océan Austral porte la signature des climats passés ainsi que des évolutions océanographiques et tectoniques (Allcock & Strugnell, 2012). En effet, au cours des dernières glaciations, la calotte glaciaire s'est étendue au point de recouvrir l'intégralité de plateau continental et contraignant ainsi les espèces benthiques à coloniser des refuges glaciaires dans les profondeurs de l'océan (Thatje *et al.*, 2005). Une étude du polymorphisme du marqueur COI d'*Harmothoe fuligineum* (Cowart *et al.*, in prep) a permis de mettre en évidence la coexistence de 3 haplotypes majoritaires ne montrant pas d'endémisme en Terre Adélie, Péninsule ou en mer de Ross (Figure IV-19). On retrouve aussi un nombre important d'haplotypes dérivés de ces trois majoritaires, et l'individu de Kerguelen dérive d'un haplotype uniquement trouvé en Péninsule. Ce réseau montre également que la population de mer de Ross pourrait être génétiquement différenciée des populations Péninsule/Terre Adélie, et constituer un fond génétique légèrement différent de celui trouvé dans les localités que nous avons échantillonnées pour le ddRAD. Ce fond génétique pourrait expliquer les 10% de fond génétique différent retrouvés en Terre Adélie.



Figure IV-19 : Réseau d'haplotypes basé sur le polymorphisme de COI d'*Harmothoe fuligineum*. Terre Adélie – Bleu ; Péninsule – Vert ; Kerguelen – Rouge ; Mer de Ross – Jaune. (Cowart et al., in prep).

En 2012, dans leur revue de différentes études portant sur la phylogéographie de nombreuses espèces Antarctiques selon leur capacité de dispersion, Allcock et Strugnell affirment que ce genre de réseaux correspond au cas type d'espèces ayant colonisé plusieurs refuges glaciaires. Etant donné le niveau de divergence entre ces haplotypes il reste possible que ce type de réseau corresponde à une coalescence neutre qui se maintiendrait dans des populations qui n'a pas subit de goulot d'étranglement. Quoi qu'il en soit l'absence de signature géographique marquée suggère qu'*H. fuligineum* possède des capacités de dispersion importantes.

Etant données les capacités de dispersion supposées d'*H. fuligineum*, et l'homogénéité des fonds génétiques entre Péninsule et Terre Adélie, comment les évènements glaciaires peuventils avoir marqué l'organisation de la diversité entre et au sein des localités ?

Comme nous l'avons vu, notre étude du polymorphisme nucléaire, tout comme l'étude du polymorphisme du COI révèle une homogénéité des fonds génétiques de Terre Adélie et de Péninsule. L'étude du polymorphisme synonyme et non synonyme dans ces deux localités montre des valeurs relativement faibles. De plus la diversité inter-individuelle soutient l'homogénéité des fonds génétiques et traduit aussi une faible diversité génétique avec des

probabilités d'identité n'excédant pas 0,25 et 0,17 au sein et entre les localités. Cette faible diversité est plutôt surprenante si l'on considère que le flux génétique important entre localités implique un brassage génétique à grande échelle au moins entre nos deux localités d'intérêt et suppose donc que cette population continentale présente une taille efficace importante qui est l'un des facteurs clés dans le maintien de polymorphisme au sein des populations. Ce même cas de figure est observable chez d'autres invertébrés Antarctiques et s'explique par le fait que les glaciations passées en contraignant les espèces à coloniser des refuges induit une réduction des effectifs de la population et la mise en place d'un goulot d'étranglement (Gonzalez-Wevar *et al.*, 2011).

Bien que les valeurs de diversité synonyme et non-synonyme ne puissent pas être prises en compte directement en raison du biais induit par le nombre de contigs ayant permis ce calcul, le ratio piN/piS permet de minimiser ce biais et de mettre en évidence une différence entre localités avec un ratio de 0,31 en Péninsule contre 0,15 et 0,14 en Terre Adélie profonde et superficielle. Il est peu probable que cette différence soit imputable à des phénomènes démographiques propres à l'une ou l'autre des régions Antarctiques puisque celles-ci présentent un flux de gènes entre elles qui devrait gommer ces différences. Cette différence de polymorphisme non-synonyme pourrait suggérer des processus sélectifs différents entre les localités, puisque que contrairement aux deux populations de Terre Adélie, celle de Péninsule couvre les deux étagements considérés, ce qui implique potentiellement une diversification des processus adaptatifs et donc une augmentation de la diversité non-synonyme. Clarke (1998) présente l'idée que la cellule constitue un environnement dans lequel les protéines et autres molécules constitue une communauté. Lorsque des variations de température ou autres facteurs modifient cet environnement cellulaire, la réponse adaptative doit se faire à l'échelle de la communauté pour maintenir un fonctionnement global optimum et à une plus grande échelle la valeur sélective de l'organisme. Or bien que l'environnement marin froid présente l'avantage d'être entre autres caractérisé par une concentration élevée en oxygène dissout, il induit aussi un ralentissement du métabolisme et donc une disponibilité énergétique limitée pour cette communauté de protéines. Il présente ainsi les températures limites de compensation comme les températures limites d'activité de la communauté qui ne sont pas nécessairement les mêmes que les température limites d'activité des protéines prises au cas par cas. Ainsi chez la plupart des espèces Antarctiques cet intervalle entre limite haute et basse de température de compensation est plus resserré que pour des espèces vivant dans d'autres environnements, et des mesures montrent que le milieu benthique Antarctique représente la limite basse de compensation pour certaines espèces (Peck, 2002). Dans le cas d'H. fuligineum, il est probable que le réchauffement climatique en Péninsule ait éloigné l'espèce de sa limite basse de compensation et permette ainsi un relâchement de la sélection purifiante visant à maintenir ce ballet énergétique et donc une augmentation de la diversité non synonyme. Il est aussi possible que la plus grande variabilité de température en Péninsule permette le maintien d'un polymorphisme qui est purgé en Terre Adélie profonde (pas de variabilité annuelle) ou superficielle (amplitude saisonnière plus petite qu'en Péninsule).

C - Structure des populations en réponse à l'adaptation aux contraintes liées à la profondeur

Les échanges thermiques entre atmosphère et océan, la formation de glace, l'évaporation ainsi que les précipitations sont des phénomènes induisant des variations de la température et de la salinité des eaux de surface (Dierssen *et al.*, 2002). Comme nous l'avons déjà vu, la salinité en surface varie entre 33 et 33,7 alors qu'en profondeur elle est plus forte et comprise entre 33,8 et 34‰. Quant à la température, le différentiel se fait surtout sur la variabilité de la température elle-même puisqu'elle est stable en profondeur à $-1,5^{\circ}$ C alors qu'en surface on observe des variations de température entre -1,8 et 1°C pour la Péninsule et -1,8 et 0°C en Terre Adélie. Or *H. fuligineum* présente une distribution bathymétrique étalée, allant des eaux proches de la surface jusqu'à des profondeurs dépassant 100m. Et notre étude a permis de montrer que la population de Péninsule porte déjà des marques imputables au réchauffement climatique et potentiellement en lien avec une augmentation locale de la salinité, menant à des questions sur la différenciation des populations d'*H. fuligineum* en lien avec la profondeur et sur le potentiel adaptatif que pourrait constituer le polymorphisme dans ces populations ?

Les jeux de données dont nous disposons nous permettent de distinguer la population superficielle de Terre Adélie des deux populations profondes (Terre Adélie et Péninsule) sur la base de la composition génétique des individus (présence/absence de locus). Cette structuration n'est plus perceptible avec les analyses AFC et Structure exploitant l'information apportée par les fréquences des allèles des 543 locus communs aux deux populations. Mais les calculs de Fst moyen permettent tout de même de mettre en évidence une différenciation significative avec un Fst moyen de 0,0405. Cette différenciation est d'autant plus marquée lorsque l'on compare PEN et TAS (840 locus commun) avec un Fst significatif de 0,1522.

Ainsi l'étude locus par locus de la différenciation entre populations permet de mettre en évidence deux phénomènes. D'une part, bien que le calcul de Fst moyen ne soit pas significatif entre habitats, on peut voir que la distribution des valeurs de Fst calculées entre profondeurs en

Terre Adélie s'étale jusqu'à des valeurs élevées de Fst (un maximum de 0,8). Et d'autre part, que ce sont les mêmes locus qui tirent la différenciation entre la population de Péninsule et celle de Terre Adélie superficielle. Ainsi, bien que la population de Péninsule soit composée d'individus échantillonnés à la fois en profondeur et en surface, elle présente le même schéma de différenciation avec TAS que la population TAD. Deux scénarios pourraient expliquer la distribution actuelle de ces fonds génétiques (Figure IV-20). On peut en effet envisager qu'il y ait initialement eu deux fonds génétiques, l'un spécifique de l'habitat superficiel et l'autre de l'habitat profond (Figure IV-20 b.), puis les changements environnementaux liés au réchauffement climatique auraient induit par mésadaptation de la population de surface ou grâce au potentiel adaptatif de la population profonde une remontée de celle-ci et une homogénéisation locale des fonds génétiques. L'autre possibilité consisterait dans la présence initiale d'un seul et unique fond génétique dans ces deux localités puis à l'arrivée d'un fond génétique étranger par migration (et non compris dans notre échantillonnage) qui aurait remplacé le premier dans l'habitat superficiel de Terre Adélie. Nos résultats ne nous permettent pas de trancher assurément pour l'un ou l'autre de ces scénarios, cependant le fait que la population TAS partage plus de locus en commun avec PEN (840 en données mappées, 67 en de novo) qu'avec TAD (543 en données mappées, 57 en de novo) pourrait traduire une absorption du fond génétique superficiel par le profond lors de sa remontée en Péninsule (Figure IV-20 b.). Il est aussi possible qu'il n'existe qu'un seul fond génétique soumis à de l'adaptation locale impactant un nombre limité de locus.


Figure IV-20 : Scénarios envisageables pour expliquer la distribution actuelle des fonds génétiques d'*H. fuligineum*. a. Présence initiale d'un seul fond génétique suivi de l'arrivée d'un fond génétique étranger. b. Deux fonds génétiques inféodés respectivement à l'habitat superficiel et profond, et remonté du fond génétique profond en Péninsule suite au réchauffement climatique.

L'étude plus précise des 543 locus communs aux deux habitats de la Terre Adélie permet de se rendre compte du nombre de locus impactés par la sélection par l'habitat. La distribution des valeurs de Fst calculées entre ces deux populations montre une distribution étalée avec une longue queue de distribution avec un maximum à 0,8. La distribution des valeurs de Fis moyen pour l'ensemble des locus est centrée sur zéro ce qui traduit une distribution neutre du polymorphisme. En revanche la distribution des valeurs de Fis moyen pour les locus identifiés comme sous sélection positive ('outliers') montre une distribution contenue entre -1 et 0. Et on peut observer une corrélation négative (0,4234) entre ces valeurs de Fis moyen et les valeurs de Fst prises par ces locus entre ces deux populations. Ceci signifie que les locus tirant le plus la différenciation sont aussi ceux présentant les plus forts excès en hétérozygotes et suppose que cette différenciation pourrait être due à un excès d'individus « hybrides » dans l'un ou l'autre des 2 habitats.

L'essentiel des locus s'écartant de l'équilibre de Hardy-Weinberg (27,4%) présente un excès dans au moins l'une des deux populations. Plus précisément 14,9% présentent un excès d'hétérozygotes dans les deux populations à la fois, ce qui est surement la marque de la paralogie pour les locus fixés à -1 dans les deux populations simultanément. En revanche pour

les autres locus ainsi que pour les 5,7 et 6,8% des locus ne présentant un excès significatif d'hétérozygotes que dans l'habitat superficiel ou le profond, cet excès d'hétérozygotie montre le rôle structurant d'une sélection diversifiante sur le polymorphisme et/ou un apport d'allèles « étrangers » sous la forme hétérozygote. Comme nous l'avons vu les eaux de surface principalement ainsi que les eaux plus profondes dans une moindre mesure sont soumises à des variations de température et de salinité qui peuvent générer momentanément des changements de pression de sélection. Mais, plus encore, la variation saisonnière générant une forte variation de la production primaire peut aussi expliquer cette forte tendance à l'excès d'hétérozygotie.

Lorsque l'on regarde concrètement les fréquences génotypiques relatives dans les deux populations pour les locus outlier présentant des écarts significatifs à la neutralité, à l'exception d'un cas évident de paralogie, trois cas de figure sont observables. L'un des locus montre clairement un excès d'hétérozygotes dans les deux populations suggérant l'action de la sélection diversifiante, avec des génotypes homozygotes spécifiques de chaque habitat. Pour un deuxième locus la situation est plus contrastée puisque l'on ne retrouve que des hétérozygotes dans la population profonde alors que la population superficielle ne présente que des homozygotes. Ceci montre que la sélection n'oppère que sur quelques gènes avec apparemment une forte sélection dirigée dans l'habitat superficiel et à l'inverse une sélection balancée tout aussi forte dans l'habitat profond. Enfin un troisième locus montre une organisation complètement différente puisque l'on ne retrouve aucun hétérozygote dans notre échantillonnage suggérant sont action de gène verrou, avec un mélange des deux types d'allèles dans la population profonde, et un allèle fixé dans la population superficielle. Ceci pourrait résulter du fait que la population apte à se maintenir dans un habitat variable à le potentiel génétique pour se maintenir dans un habitat moins variable mais aux caractéristiques thermiques proches.

Parmi les gènes outliers identifiés, on retrouve l'HSP70 et l'ARN hélicase qui peuvent toutes deux correspondre à une réponse aux conditions thermiques différenciant des deux habitats. De plus, il est intéressant de noter le fait que les HSP 70 ont permis de mettre en évidence chez deux espèces sœurs de krill une adaptation à l'environnement hauturier et côtier avec un niveau contrasté de diversité nucléotidique entre les deux habitats. Ces différences s'expliqueraient par le couplage d'histoires démographiques différentes et de l'action de la sélection directionnelle pour l'habitat le plus stable et le plus froid (Papot *et al.*, 2016). Ces deux espèces de krill colonisant des habitats différents, elles n'ont pas été impactées de la même manière par l'histoire climatique de l'Antarctique. Dans le cas d'*H. fuligineum*, le polymorphisme mitochondrial ne semble pas révéler la coexistence de deux lignées mitochondriales distinctes

et aux histoires évolutives différentes. Comme nous l'avons vu avec l'exemple précédent, l'exploitation d'habitats différents implique une exposition différente à la glaciation et donc des histoires évolutives différentes. Il est très probable qu'H. fuligineum ait été moins impactée que d'autres espèces aux glaciations passées, ce qui lui a permis de mainenir un polymorphisme impotant et commun entre localités aussi grâce à des capacités de dispersion importantes. De plus il semble que du flux de gènes en réponse à la variation de la température est envisageable et dépendant de l'effet que celui-ci aura. Comme on peut le voir en Péninsule, le réchauffement climatique se caractérise dans cette région Antarctique par une hausse de la température et de la salinité. Or si ce sont les populations superficielles qui sont exposées aux variations de température plus importantes, ce sont les populations profondes qui sont exposées aux plus hauts niveaux de salinité. De plus si la sensibilité des populations à la hausse de température est corrélée non pas à l'éventail de températures auxquelles elles sont exposées mais plutôt au minimum expérimenté comme le suggère Clarke (1998), on peut penser que les populations superficielles exposées à des températures atteignant -1,8°C seront plus sensibles à une hausse de la température que les populations plus profondes. Mais quelle que soit la population la plus vulnérable, le fait qu'elles se différencient, entre autres, par une HSP 70 (impliquée, entre autres, dans la réponse au stress thermique) constitue un bon indice suggérant une possible adaptation dans l'un ou l'autre des habitats et par flux de gènes ou remplacement des fonds génétiques un maintien au moins à court terme des populations malgré une hausse de la température.

5 - Conclusion

Harmothoe fuligineum est une espèce de Polynoidae Antarctique qui présente à la fois une distribution circumcontinentale ainsi que sur les Iles Kerguelen, mais aussi le long de l'environnement benthique depuis des profondeurs de quelques mètres jusqu'à plus de 100m. Cette distribution induit une exposition à des conditions environnementales diverses depuis les Kerguelen qui, avec les autres îles sub-antarctiques, constituent l'environnement le plus variable et éloigné des pôles en Antarctique, jusqu'au milieu profond de Terre Adélie qui constitue l'un des environnements les plus stables au monde avec une température stable de - 1,5°C et une salinité comprise entre 33,8 et 34‰.

Comme pour de nombreuses espèces endémiques de l'Antarctique, la question se pose de leur résilience vis-à-vis du changement climatique et du réchauffement global qu'il implique. En effet la tectonique et la mise en place du courant circumpolaire ont généré depuis 33,9 millions

d'années un isolement thermique et biologique du continent. Ainsi contrairement à d'autres espèces colonisant d'autres environnements, il est possible que le polymorphisme soit réduit (résultant d'une grande stabilité thermique et de sélection purifiante associée) et la possibilité de migration vers des latitudes plus froides. Les capacités de résilience au réchauffement climatique ne sont donc pas évidentes. La Péninsule Antarctique où l'on retrouve *H. fuligineum*, est déjà à l'échelle de l'Antarctique ainsi qu'à celle du globe, l'une des régions les plus affectées par le changement climatique et montre depuis une soixantaine d'années une hausse de la salinité et de la température de ses eaux.

Notre étude met en évidence l'homogénéité à l'échelle du génome entier des patrimoines génétiques des différentes populations de Péninsule et de Terre Adélie qui résulte surement de capacité de dispersion permettant à l'espèce d'exploiter la circulation circumpolaire pour propager ces larves produites en grand nombre. Malgré cette homogénéité, il est possible de distinguer une différenciation en Terre Adélie entre les habitats profonds et superficiels, potentiellement en réponse à des conditions thermiques différentes. Cette différence ne se retrouve pas en Péninsule où seul le fond génétique profond de Terre Adélie est retrouvé.

Enfin, la présence d'un individu aux Iles Kerguelen non affilié à l'une de ces trois populations laisse supposer qu'il existe un fond génétique capable de se maintenir au niveau des îles subantarctiques, plus chaudes et exposées à une amplitude saisonnière plus importante. Bien qu'on ne puisse affirmer avec certitude la possibilité qu'un flux de gènes s'opère entre cette population et le continent, on peut supposer que cette population constitue un pool de gènes qui permettrait aux populations continentales de répondre aux variations environnementales à venir.

1 - Introduction

Comme nous l'avons vu précédemment pour les Alvinella en milieu hydrothermal, ainsi que pour Harmothoe fuligineum en Antarctique, la diversité génétique des espèces est à la fois impactée par l'histoire démographique des populations, le degré de connectivité entre elles et la variabilité environnementale à échelle locale qui peut conduire tantôt à des balayages sélectifs tantôt à un maintien des polymorphismes. Ainsi le fait d'être exposé ou non à un environnement hypervariable dans l'espace et dans le temps, ainsi que l'amplitude et la fréquence de ces variations pourraient constituer une force suffisante pour contraindre et partitionner la diversité génétique d'une espèce, au moins au niveau de sa composante nonsynonyme. Or la diversité génétique constitue, avec la taille efficace et l'hétérozygotie, un proxy particulièrement bien corrélé à la valeur sélective des espèces et leur potentiel adaptatif vis-à-vis des modifications de l'environnement (Reed & Frankham, 2003). Par exemple, il a ainsi été mis en évidence, en comparant deux espèces soeurs du krill Antarctique ayant des tailles efficaces comparables mais vivant dans des environnements thermiques différents, d'une part, que la diversité non-synonyme était nettement supérieure chez l'espèce exposée aux conditions thermiques les plus variables et, d'autre part, que l'espèce soumise aux conditions les plus froides (et stables) avait subi une baisse drastique de sa diversité nucléotidique en réponse à de la sélection purifiante (et/ou balayage sélectif) ou un goulot d'étranglement (Papot et al., 2016).

A - Les traits d'histoire de vie

De nombreuses études ont cependant montré que la diversité génétique neutre d'une espèce était directement liée à la taille efficace de ses populations à travers l'équation θ =4.Ne.µ, où Ne est la taille efficace de l'espèce et µ, le taux de mutation (Kimura & Crow, 1964). Dans ce cadre, les principaux facteurs influençant cette taille efficace sont le mode de reproduction et les traits d'histoire de vie des espèces, et notamment leur capacité à produire de nombreux descendants (Hamrick *et al.*, 1979). Les récents travaux de Romiguier et al. (2012, 2014) ont d'ailleurs confirmé à l'aide d'une approche RNAseq sur un échantillon représentatif d'espèces terrestres et marines de vertébrés et d'invertébrés que les polymorphismes synonyme et nonsynonyme des espèces étaient régulés par le mode de reproduction et la capacité à produire un grand nombre de descendants. Néanmoins, les espèces caractérisées par de très grandes tailles efficaces, de fortes fécondités et un grand nombre de descendants telles que trouvées chez un grand nombre d'invertébrés marins (Poulsen *et al.*, 2006) sont également celles ou les chances de voir apparaître des mutations bénéfiques sont les plus grandes (Kimura, 1983), laissant supposer que chez ces espèces, la sélection positive peut également jouer un rôle important dans la mise en place de polymorphismes adaptatifs à l'échelle du génome.

B - Variabilité environnementale et sélection purifiante

Mais pour des traits d'histoire de vie similaires, on observe une forte corrélation entre la variabilité de l'environnement et la diversité génétique (Vellend, 2005). Ainsi en milieu hétérogène, en plus des traits d'histoire de vie des espèces, les variations temporelles et spatiales de l'environnement peuvent en effet affecter la diversité génétique des espèces si les différents génotypes présentent momentanément ou localement des valeurs sélectives plus ou moins élevées pour un environnement donné (Nevo, 2001). Des modèles ont tenté d'expliquer de quelle façon des environnements compartimentés en différentes niches pouvaient induire le maintien d'un polymorphisme stable dans les populations (Levene, 1953 ; Gillespie, 1976), mais ce genre de modèles présente des limites, telles que le fait de ne pas prendre en compte le recrutement d'œufs/larves/juvéniles qui s'opère au sein de chaque niche ou celui de surestimer la valeur sélective des hétérozygotes et ainsi biaiser les simulations des modèles (Smith & Hoekstra, 1980).

Dans les environnements hypervariables comme l'environnement hydrothermal, les espèces sont plutôt caractérisées par des diversités relativement faibles, possiblement dues à la fois aux phénomènes d'extinction/recolonisation récurrents dans les populations (Gilpin 1991) et aux fortes pressions de sélection purifiante s'exerçant sur les espèces. Vrijenhoek (1997, 2003) a notamment montré que les espèces associées aux habitats les plus fragmentés présentaient une diversité génétique plus faible et que l'hétérozygotie était négativement corrélée au taux d'extinction des sites hydrothermaux malgré la variabilité extrême de l'environnement.

On s'attend également à retrouver des espèces présentant des niveaux faibles de diversité génétique dans les environnements stables en raison de la sélection purifiante exercée par ces

environnements sur les espèces. Il a d'ailleurs été montré que bien que la diversité génétique soit corrélée avec l'intensité du stress induit par l'environnement, une fois passé un certain seuil la diversité génétique chute drastiquement en raison de la forte sélection purifiante que ce stress implique (Kis-Papo *et al.*, 2003). Ainsi l'Antarctique qui est parmi les environnements les plus froids et stables thermiquement est aussi celui qui possède les espèces les plus sténothermes au monde (Aronson *et al.*, 2007). Mais cette particularité environnementale imposant des températures très faibles voire négatives et donc potentiellement stressantes laisse supposer que l'impact de la sélection purifiante est grand dans cet environnement stable à l'extrême. Il existe cependant des cas ne suivant pas cet attendu. Il a en effet été montré que les populations d'*Euphausia superba* en mer de Scotia présentent une diversité génétique mitochondriale supérieure à ce que l'on retrouve habituellement pour les espèces antarctiques avec une absence de structuration des populations. Goodall-Copestake et collaborateurs (2010) expliquent cette situation par une probable expansion démographique ancienne à la suite du refroidissement progressif du continent durant le Pleistocène.

C - Les phénomènes démographiques : extinction/recolonisation, goulots d'étranglement et expansion démographique

Un autre facteur impactant la diversité actuelle des populations dépend aussi de leur histoire passée. Ainsi, dans certains cas, les épisodes de glaciation ont réduit drastiquement les aires de distribution de certaines espèces, et amené celles-ci à émigrer dans des refuges. Ces réductions de la taille efficace des populations passées laissent des traces dans le génome, encore observables aujourd'hui (Hewitt, 2004), bien que des remises en contact entre zones refuges aient pu localement conduire à des niveaux élevés de diversité au sein des populations et à de la différenciation génétique entre celles-ci (Hewitt, 1999 ; Maggs *et al.*, 2008). L'isolement du continent Antarctique rend les espèces particulièrement sensibles à ces glaciations puisque qu'une retraite à des latitudes plus hautes impliquerait de s'éloigner du plateau continental (Brey *et al.*, 1996). L'histoire évolutive de ces populations et les isolements passés constituent aussi une source importante de variabilité génétique.

D - Les contacts secondaires entre populations génétiquement différenciées

Les remises en contact secondaires entre populations précédemment isolées géographiquement constituent également un moteur pour générer de la diversité génétique localement. Les flux de gènes même limités opérant entre ces entités génétiques aux histoires évolutives et adaptatives différentes génèrent de nouveaux allèles par recombinaison à la suite d'introgression entre génomes et donc une augmentation locale du polymorphisme dans la zone de contact (Hilbish, 1996). Dans le cas de l'Antarctique et des sources hydrothermales (à l'échelle d'une dorsale) qui sont pourtant des environnements dans lesquels les différentes localités sont globalement similaires du point de vue de leurs conditions thermo-chimiques, nous avons pu noter que de légères différences entre localités sont suffisantes pour générer des adaptations locales et, par la même occasion, un apport de diversité au travers des flux de gènes entre ces populations et le recrutement de certains allèles dans certains habitats. Mais cette variabilité environnementale peut jouer d'une autre façon sur la diversité génétique. En effet, le complexe d'espèces *Mytilus edulis / M. galloprovencialis* qui présente une histoire évolutive complexe se retrouve aujourd'hui en sympatrie de façon répétée sur les côtes Européennes, et cette répétition de contacts secondaires permet de mettre en évidence aujourd'hui des introgressions d'allèles suivant des modalités différentes selon les zones de contact en réponse à des contraintes environnementales (Fraïsse et al., 2014). Ceci laisse supposer que des environnements présentant plus de contraste ou tout simplement des espèces couvrant plusieurs environnements comme celles des côtes Européennes disposent d'un potentiel génétique d'autant plus diversifié.

E - Implications fonctionnelles

D'autre part, les diversités génétiques observées chez les espèces ne prennent pas en compte les implications physiologiques des mutations qui les caractérisent. Il a été par exemple montré le rôle capital de l'environnement cellulaire et notamment des protéines chaperonnes dans la compensation des mutations à effet délétère (Tokuriki & Tawfik, 2009). En effet, la séquence primaire d'une protéine et particulièrement celle d'une enzyme définit sa structure et sa fonction et, par la même occasion, sa température optimale d'action. Ainsi, l'augmentation de la température conduit dans un premier temps à une augmentation de l'efficacité catalytique de l'enzyme puis à la dénaturation de la protéine en perturbant la structure secondaire et tertiaire de celle-ci. A l'inverse une baisse de température, en induisant une diminution de la flexibilité de l'enzyme, causera une diminution de son efficacité catalytique (Fields et al., 2015). Ainsi la question du maintien d'un métabolisme basal en fonction de la température se pose à la fois pour des espèces exposées à des variations hautes fréquences et de grande amplitude de l'environnement telle que celles trouvées au niveau des sources hydrothermales et, celles exposées à une variabilité plus saisonnière de l'environnement. Pour les organismes ectothermes, on peut donc s'attendre à ce que la sensibilité thermique reflète la température de vie des espèces considérées. Les systèmes enzymatiques des organismes ectothermes doivent

en effet maintenir leur efficacité à la température imposée par l'environnement (Somero & Hochachka, 1971). Ceci a notamment été montré pour la lactate deshydrogénase des muscles des poissons (pour une revue, voir Somero, 2004) et, pour les invertébrés, la malate déshydrogénase cytosolique (cMDH) est plus particulièrement utilisée (Fields et al., 2006, Dong et Somero, 2009). Ainsi, chez les espèces du genre *Mytilus* sur la côte californienne, la distribution latitudinale des espèces est corrélée à la thermostabilité de la cMDH (Fields *et al.,* 2006). Ceci est aussi vrai pour les patelles du genre *Lottia*, dont l'aire de distribution des espèces s'est déplacée de 75 km vers le pôle en 30 ans, en relation avec le réchauffement climatique (Dong et Somero, 2009). Dans cette même étude, les auteurs ont montré que les espèces intertidales, exposées à de fortes variations journalières de température, possédaient des cMDH moins sensibles aux variations de température que la même enzyme de l'espèce subtidale (seulement soumise à des variations saisonnières de température). Ces études ne se sont par contre pas intéressées aux différentes formes alléliques de ces enzymes au sein de chaque espèce.

Le fait d'observer une forte proportion de polymorphisme, même non-synonyme n'implique pas nécessairement de retrouver cette diversité à l'échelle structurale ou fonctionnelle des protéines. En effet la substitution d'un acide aminé par un autre n'aura pas le même impact selon leur nature et la position dans la protéine (Jiang *et al.*, 2018). Ainsi la majorité des mutations synonymes comme non-synonymes n'ont un effet que limité sur la structure et la fonction des protéines et adoptent un comportement neutre, seules les recombinaisons entre allèles permettent une exploration à grande échelle du paysage adaptatif (Xia & Levitt, 2002). Il apparait donc capital d'étudier aussi directement l'impact de l'environnement sur des cas concrets.

Dans ce chapitre, nous étudierons le polymorphisme nucléotidique d'espèces d'annélides polychètes appartenant aux sous-ordres des Terebellomorpha (familles Alvinellidae et Terebellidae) et des Aphroditiformia (famille Polynoidae) qui occupent une large gamme d'environnements thermiques contrastés allant du milieu hydrothermal aux eaux froides de l'Antarctique. En étudiant l'information apportée par leur polymorphisme nucléotidique et les performances catalytiques de leurs enzymes nous voulons déterminer si l'environnement et sa variabilité thermique ont un effet spécifique sur la diversité génétique de ces espèces. Nous chercherons à répondre aux questions suivantes :

La variabilité thermique de l'environnement favorise-t-elle un mode de sélection en particulier ? et, si oui, lequel ?

Les variations thermiques hautes fréquences marquent-t-elles préférentiellement le polymorphisme non-synonyme des espèces et existe-t-il une signature environnementale sur la diversité génétique des espèces vivant dans les environnements extrêmes stables et instables ? Les substitutions des acides aminés entre allèles sont-elles dépendantes d'un environnement donné dès lors qu'elles affectent la structure 3D de la protéine ?

Dans le cas précis de la cMDH, observe-t-on une corrélation entre variabilité de l'environnement et diversité des enzymes du point de vue de leur thermostabilité entre individus d'une même espèce ?

2 - Matériel et Méthodes

Tout comme pour le chapitre précédent, l'ensemble des techniques de laboratoire et de traitement bioinformatique des données ddRAD obtenues sur les 12 espèces d'annélides polychètes pour estimer leurs diversités nucléotidiques synonyme et non synonyme sont présentés dans le chapitre 2 de cette thèse. Dans ce cas précis néanmoins, les locus ddRAD obtenus proviennent de la cartographie des lectures du séquençage Illumina de chaque espèce sur leur propre transcriptome de référence, à l'exception du terebellidae côtier *A. edwardsi* pour lequel nous n'avons pu obtenir d'ARNm de qualité suffisante pour en séquencer le transcriptome. Tous les transcriptomes ont été assemblés à l'aide du package Trinity (Grabherr et al. 2011) selon la procédure décrite dans le chapitre 2 et la cartographie des lectures a été réalisée sur la partie codante des transcrits en recherchant l'ORF de chaque transcrit avec le programme GetORF (https://github.com/peterjc/pico_galaxy/tree/master/tools) dans l'environnement Galaxy.

A - Echantillonnage

Pour étudier l'effet de la stabilité/instabilité thermique des environnements marins sur la diversité génétique des espèces, nous avons choisi de nous restreindre aux annélides polychètes associés à la plus large gamme d'habitats thermiques possible, du milieu Antarctique côtier aux sources hydrothermales profondes. Notre choix s'est porté sur 2 groupes de polychètes, la famille des Polynoidae et les terebellomorphes (familles proches Alvinellidae et Terebellidae). Notre stratégie d'échantillonnage a été d'appréhender le polymorphisme des espèces à partir de 16 individus par espèce, en prenant 2 espèces par famille et par environnement (3 environnements X 2 familles X 2 espèces par famille), en essayant de prendre en compte des espèces ayant à peu près les mêmes traits d'histoire de vie (grandes tailles efficaces des

populations, des fécondités modérées à fortes et un mode de développement avec larves pélagiques). A noter, cependant que le mode de reproduction apparaît légèrement différent entre les espèces hydrothermales profondes et les espèces des 2 autres environnements (côtier tempéré et polaire) avec un mode de fécondation interne des ovocytes et un développement lécithotrophe des larves chez les espèces hydrothermales et une fécondation des ovocytes dans la colonne d'eau avec un développement planctonotrophe des larves chez les espèces des deux autres environnements. Un récapitulatif des différentes espèces étudiées avec leurs caractéristiques biologiques et écologiques est présenté dans le tableau V-1. Les animaux ont été récoltés (1) sur la grève de Roscoff lors des marées de vive eaux sous les pierres pour A. gelatinosa et T. lapidaria, et dans le sable pour P. furcosetosa et A. edwardsi, (2) lors de 2 expéditions d'été sur la base Dumont d'Urville à partir de plongées en scaphandre autonome sous la glace entre 5 et 20 m de profondeur, et (3) lors des campagnes océanographiques Mescal 2010 et Mescal 2012 (chefs de mission : N. Le Bris et F.H. Lallier) à 9°50 N sur la dorsale est Pacifique avec le bathyscaphe Nautile et son navire support N/O L'Atalante et la campagne américaine 2009 TM-235 avec le ROV Jason II et son navire support le R/V Thomas G. Thompson.

Sous- ordre	Famille	Environnement	Espèce	Mode de reproduction	Fécondité (nombre d'oeufs moyen)	Type de développement larvaire		
		Côtier tempéré - Roscoff intertidal	Terebella lapidaria					
	ellidae	Côtier tempéré - Roscoff subtidal	Amphitrite edwardsi	épitoquie	146021	planctonique sur réserves (McHugh, 1993)		
orpha	Terebe	Antarctique -	Thelepus sp.	2004)	Fong, 2002)			
Terebellom		d'Urville	Amphitritides sp.					
	ellidae	Hydrothermal -	Alvinella pompejana	fertilisation interne (Jollivet	jusqu'à 137500	lécitotrophique ou benthique (Desbruyères <i>et al.,</i> 1998)		
	Alvine	Pacifique, 2500 m	Alvinella caudata	Jouin-Toulmond et al., 2002)	(Pradillon <i>et</i> <i>al.,</i> 2005)			
		Côtier tempéré - Roscoff subtidal	Pettibonesia furcosetosa					
		Côtier tempéré - Roscoff intertidal	Alentia gelatinosa	épitoquie				
Aphroditiformia	Antarctique - Péninsule		Harmothoe fuligineum	(Ruppert <i>et al.,</i> 2004)				
	Polynoida	Antarctique - Station Dumont d'Urville	Harmothoe crosetensis		526000 (McHugh & Fong, 2002)	planctonique (Rouse, 2000)		
		Hydrothermal - dorsale Est Pacifique, 2500 m	Lepidonotopodium fimbriatum	lecithotrophique ou				
		Hydrothermal - Bassin de Lau 1700m	Branchinotogluma Branchinotoglumasegonzaci	direct (Dreyer et al., 2005)				

Table V-8 : Récapitulatif des environnements, modes de reproduction, fécondité et types de développements larvaires pour les espèces étudiées

B - Estimation des diversités nucléotidiques globales, synonymes et nonsynonymes

Les diversités nucléotidiques des différentes espèces ont été estimées à partir des fichiers de sortie du programme 'population' du logiciel STACKs, notamment les fichiers de sortie Genepop et les fichiers de séquences des locus en format fasta. Ces fichiers de séquences ont été ensuite utilisés sous DNAsp v.5 (Librado et Rozas, 2009) en mode 'batch' pour estimer la diversité nucléotidique globale et effectuer le test D de Tajima afin de déterminer la part respective des effets démographiques et sélectifs sur cette diversité.

Le calcul des diversités synonyme et non-synonyme a été réalisé en utilisant le pipeline

Reads2snps développé par N. Galtier et collaborateurs (Gayral et al. 2013). Dans cette étude, seules les polymorphismes synonyme et non-synonyme ont été estimés en l'absence d'une espèce outgroup suffisamment proche pour être utilisée pour calculer les divergences synonyme et non-synonyme entre espèces d'environnements différents. Cette restriction ne nous a donc pas permis d'effectuer de tests plus puissants comme le test de McDonald-Kreitman pour détecter les effets d'une sélection diversifiante entre habitats.

3 - Distribution des mutations rares ou en fréquences intermédiaires dans le polymorphisme des espèces (spectres de fréquences alléliques)

Après concaténation des alignements des différents allèles obtenus par Reads2snps pour une espèce donnée (alignements de 12 à 20 allèles selon l'espèce), pour chaque position polymorphe la fréquence de l'allèle alternatif a été calculée en utilisant un minimum de 16 allèles (8 pour *P. furcosetosa,* espèce pour laquelle nous n'avons que 6 individus séquencés) identifiés. Puis des comptes par classe de fréquence sont réalisés pour identifier le nombre de mutations par classes de fréquence à l'aide d'un script python développé par V. Mataigne (M2 Bioinfo, ABIMS).

A - Distribution des mutations polymorphes non-synonymes selon la nature des acides aminés

Comme précédemment, après concaténation des alignements des différents allèles obtenus par Reads2snps pour une espèce donnée (alignements de 12 à 20 allèles selon l'espèce), les remplacements d'acides aminés entre allèles ainsi que leur directionalité ont été comptés entre paires de séquences à l'aide du programme MutCount de la suite AdaptSearch implémentée dans Galaxy3 (AdaptSearch v. beta, V. Mataigne ; Fontanillas et al. 2016). Les fréquences de ces remplacements selon les 12 espèces analysées ont ensuite fait l'objet d'une analyse en composantes principales avec le module FactomineR (Hê et al. 2008) sous R.

B - Mesures d'activité de la cMDH entre espèces et détermination de la température de dénaturation thermique (Tm) des différentes isoformes

Afin d'étudier un cas concret, et de se faire une idée de l'effet de l'environnement et de l'action de la sélection nous avons cherché à caractériser le polymorphisme d'un système enzymatique central. La malate déshydrogénase cytoplasmique (cMDH) joue un rôle important dans l'alternance des métabolismes aérobie et anaérobie (navette malate-aspartate). Elle est présente dans tous les tissus, en quantité suffisamment importante pour permettre la mesure de son activité sans nécessité de purification. Elle catalyse l'inter-conversion de l'oxaloacétate (OAA) en L-malate en utilisant comme co-facteur le NADH. Le suivi de la réaction dans le sens de la formation de L-malate est possible grâce au suivi de la consommation simultanée de NADH qui présente un pic d'absorbance à 340 nm. Pour étudier la cinétique de cet enzyme entre les différentes espèces et leur habitat thermique, les tissus de chaque individu/espèce sont placés dans 500 µL de tampon d'extraction anti-protéase (NaCl 150 mM ; PMSF 1 mM ; NaH₂PO₄ 10 mM ; pH 7,2) pour y être broyés sur glace. Le broyat est ensuite centrifugé à 4°C, 12000 rpm pendant 15 min afin de récupérer le surnageant contenant les protéines cytosolubles et conservé sur glace jusqu'à son injection dans un milieu réactionnel afin de faire ensuite les mesures d'activité dans un spectrophotomètre thermostaté à 20°C.

C - Mesure d'activité et détermination du Tm

Le mélange réactionnel (tampon Imidazole 50 mM ; MgCl₂ 1,5 mM ; NADH 0,1 mM ; OAA 0,1 mM ; pH 7,4) est placé directement dans les cuves de spectrophotométrie et 10 μ L d'extrait protéique brut est ajouté. Des tests préliminaires de dilution des extraits protéiques sont réalisés afin que la concentration en cMDH permette de suivre le déroulement de la réaction enzymatique sur 5 min et de déterminer la vitesse initiale de la réaction. La réaction s'effectue à 20°C avec des extraits protéiques ayant été incubés auparavant durant 20 min à 0, 15, 25, et une gamme de température entre 30 et 45°C (10 échantillons dans une machine PCR à gradient), au besoin d'autres mesures ont été faites dans une gamme de températures plus étroite autour de la température ou une baisse d'activité est constatée. Les vitesses initiales ainsi obtenues pour les différentes températures d'incubation permettent de définir la température (Tm) pour laquelle la cMDH perd la moitié de son activité initiale (référence mesurée avec l'extrait incubé à 0°C). La valeur de Tm est déterminée par ajustement d'une courbe de type 'survie' avec le logiciel JPM11. L'algorithme explore différentes valeurs permettant d'ajuster la valeur du

plateau, la valeur de Tm et une valeur correspondant à la pente autour du Tm. Chaque valeur permet de tester la corrélation avec les valeurs mesurées de façon à minimiser les écarts entre valeur théorique et mesurée. Les valeurs de Tm sont ensuite comparées entre individus au sein de chaque espèce et entre espèces à l'aide de boîtes à moustaches obtenues par ggplot sour R pour estimer le degré d'hétérogénéité de la sensibilité des différentes isoformes selon l'habitat thermique de l'espèce analysée.

D - Test de Tukey

Le test de Tukey ou test DSH (différences significatives honnêtes) permet une comparaison simultanée deux à deux de plusieurs moyennes en se basant sur des distributions de Student. Il permet ainsi de mettre en évidence des groupes statistiques au sein d'un ensemble de distributions. Ce test est réalisable sous R et permet une plus grande confiance dans les différences mises en évidence puisqu'il est plus conservateur que la plupart des tests habituels réalisés pour ce genre de comparaison (Faraway 2002). Ce test nous a également permis tester la significativité des différences entre les distributions du D de Tajima et des indices de diversité associés obtenus par espèce.

4 - Résultats

A - Contrôle qualité des banques

Après un séquençage Illumina sur 2 lignes d'un séquenceur Hiseq 2500 par le service de génomique dédié de l'Université McGill (plateforme Génome-Québec, Canada), les données nous ont été envoyées sous la forme de fichiers après démultiplexage des lectures selon les 12 index choisis et nettoyés sur la base des phred scores (12 fichiers compressés contenant les données de séquençage appariées (R1 & R2)).

Dans un premier temps une vérification du nombre de lectures obtenues pour chaque combinaison barcode/index (données individu) a été réalisée pour identifier la proportion relative de chaque individu dans la composition de la banque pour le séquençage. La Figure V-1 montre une comparaison des individus les plus couverts de chaque espèce. Cette comparaison permet de nous assurer que le nombre de lectures obtenues par séquençage Illumina est comparable entre individus au sein d'une même espèce et entre espèces.



Figure V-18 : Abondances relatives des lectures de chaque individu dans les banques spécifiques pour nos espèces de Terebellomorpha (Alvinellidae et Terebellidae)



Figure V-19 : Abondances relatives des lectures de chaque individu dans les banques spécifiques pour nos espèces de Polynoidae



Figure V-20 : Couvertures (en nombre de lectures) comparées entre les 12 espèces sur les 10 individus les plus couverts (6 pour *P. furcosetosa*).

B - Alignement des séquences sur les transcriptomes de référence

Pour chacune des banques associées à une espèce, les lectures des 10 individus les plus couverts ont été conservées et alignées sur un transcriptome de référence. Le tableau V-2 récapitule les conditions d'alignement des lectures sur les transcrits ainsi que le pourcentage de lectures associé à une région codante pour chaque espèce. Pour la majorité des cas, l'alignement réalisé en End-to-End a permis d'obtenir un taux d'alignement des lectures compris entre 0,23 et 5,14%. Dans les autres cas, différents tests ont été réalisés afin de définir les conditions d'alignement en mode Local (plus permissif en fonction de la longueur de la graine permettant l'alignement et le nombre de mésappariements sur cette graine). La taille optimale de la séquence d'ancrage (graine) pour obtenir un pourcentage de locus maximal est comprise entre 18 et 20 bp.

Table V-9 : Bilan des conditions et résultats des alignements réalisés pour chaque espèce. L'espèce de référence correspond au transcriptome utilisé comme référence. Le taux d'alignement (en % de lectures alignées sur le total) est indiqué.

Espèce	nb moyen de reads/ind	type d'alignement	Espèce de référence	nb moyen de reads mappés/ind	taux d'allignement	
Harmothoe crosetensis	5603429,9	End-to-end	Harmothoe crosetensis	60865,7	1,09	
Harmothoe fuligineum	2858774,1	End-to-end	Harmothoe fuligineum	76607,6	2,68	
Branchinotogluma segonzaci	3569269	End-to-end	Branchinotogluma segonzaci	109623,8	3,07	
Lepidonotopodium fimbriatum	5125025,1	End-to-end	Lepidonotopodium fimbriatum	263568,1	5,14	
Alentia gelatinosa	5185736,5	End-to-end	Alentia gelatinosa	12583,2	0,24	
Pettibonesia furcosetosa	5282525,4	End-to-end	Pettibonesia furcosetosa	169805,2	3,21	
Amphitritides	4153254,1	Local 18	Amphitritides	132790,7	3,20	
Thelepus	6421459,8	Local 20	Thelepus	120364,6	1,87	
Alvinella caudata	2742490,5	End-to-end	Alvinella caudata	31789,3	1,16	
Alvinella pompejana	3772521	End-to-end	Alvinella pompejana	125219	3,32	
Terebella lapidaria	4925976,7	End-to-end	Terebella lapidaria	146323,3	2,97	
Amphitrite edwardsi	6081180,3	Local 20	Thelepus	123561,6	2,03	

Ces alignements étant réalisés sur des transcriptomes limités aux cadres de lecture ouverte, il a été possible d'identifier des sites polymorphes (SNP) pour un petit nombre de contigs et de caractériser les polymorphismes synonyme et non-synonyme à l'aide du script dNdSpiNpiS (développé par Nicolas Galtier et collaborateurs). Les figures V-4 et V-5 représentent respectivement la moyenne et l'écart type des valeurs de π_N et π_S mesurées pour l'ensemble de ces contigs pour les Terebellomorpha et Aphroditiformia avec le nombre de contigs utilisés pour chaque espèce. On constate dans les deux groupes d'annélides (Terebellomorpha et Aphroditiformia) une grande diversité dans la distribution des valeurs de π_N et π_S selon l'espèce. Bien que très variable d'une espèce à l'autre, aucune corrélation n'a été trouvée entre la variabilité inter-locus (différence entre les valeurs maximale et minimale) et le nombre de contigs utilisés pour chaque espèce.

Dans le cas des Aphroditiformia, une corrélation linéaire positive est observée entre π_N et π_S avec une distinction marquée des deux espèces hydrothermales (*L. fimbriatum* et *B. segonzaci*) présentant des diversités synonyme et non-synonyme nettement inférieures à celles des autres espèces. En effet, ces deux espèces ont une diversité trois fois plus petite que celle des espèces originaires d'Antarctique et de Bretagne. De plus, bien qu'aucune différence significative n'ait été trouvée entre les espèces antarctiques et bretonnes, les espèces de l'environnement polaire semblent présenter des valeurs de diversité légèrement plus élevées que celle du milieu côtier breton.



Figure V-21 : Diversités synonyme (piS) et non-synonyme (piN) des espèces de Polynoidae hydrothermales (rouge), antarctiques (bleu) et bretonnes (vert). Les moyennes et écarts types présentés sont obtenus sur la base des mesures réalisées un nombre de contigs relativement variable et indiqué pour chaque espèce.

Dans le cas des Terebellomorpha, on ne retrouve pas de corrélation linéaire positive entre les valeurs de π_N et π_S . La gamme de valeurs de π_N moyen est bien plus petite, n'excédant pas

0,0022 alors qu'elle atteint 0,0036 pour le Polynoidae *H. crosetensis*. Les valeurs maximales de π_S sont par contre très similaires. Au sein des Terebellomorpha, alors qu'aucune différence significative n'est observable sur la base des valeurs de π_N , les espèces du milieu côtier breton présentent des valeurs de π_S significativement supérieures à celles observée dans le milieu hydrothermal avec des valeurs intermédiaires dans l'environnement polaire. Bien que la diversité synonyme soit plus élevée chez les espèces côtières par rapport à leurs homologues hydrothermaux, les niveaux de contrainte sur le polymorphisme des espèces sont à peu près semblables avec des ratios non-synonyme/synonyme compris entre 0,1 et 0,3. Ces résultats ne permettent donc pas de soutenir l'hypothèse que les pressions de la sélection purifiante soient plus fortes chez les espèces hydrothermales profondes par rapport aux 2 autres environnements, et même d'envisager un effet inverse avec un polymorphisme non-synonyme qui n'augmente pas avec la diversité synonyme chez les espèces polaires et côtières les mieux échantillonnées.



Chapitre V – Effet de la variabilité environnementale sur la diversité génétique des invertébrés marins et sur la thermotolérance de leurs enzymes

Figure V-22 : Diversités synonyme (piS) et non-synonyme (piN) des espèces de Terebellomorpha hydrothermales (rouge), antarctiques (bleu) et bretonne (vert). Les moyennes et écarts types présentés sont obtenus sur la base des mesures réalisées sur un nombre de contigs variable selon l'espèce considérée (indiqué pour chaque espèce).

Dans l'ensemble, chez les Aphroditiformia, on retrouve une hiérarchisation des environnements que l'on observe à la fois au niveau du synonyme et du non-synonyme (figure V-6). Les amplitudes observées entre les valeurs monolocus extrêmes de π_N et π_S sont plus importantes dans l'environnement côtier que dans l'environnement Antarctique présentant lui-même des valeurs supérieures à celles de l'environnement hydrothermal. Cette tendance se retrouve mais en moindre mesure chez les Terebellomorpha pour lesquels certaines espèces sont mal échantillonnées (*A. edwardsi* = 30 contigs).



Figure V-23 : Amplitude des diversités synonymes (piS) et non-synonymes (piN) obtenues pour chaque contig chez les espèces hydrothermales (rouge), antarctiques (bleu) et bretonne (vert) des sous ordres Aphroditiformia et Terebellomorpha.

C - Distribution des valeurs de diversité nucléotidique et des valeurs du test de Tajima monolocus entre les différentes espèces

Les comportements du polymorphisme synonyme et non-synonyme ne suivant pas les mêmes modalités chez les deux sous-ordres étudiés, il a été décidé de regarder plus en détail la diversité génétique globale et les écarts à l'accumulation neutre des mutations dans le polymorphisme afin de savoir si certaines espèces subissent ou ont subi des goulots d'étranglements dans leurs populations. Ainsi pour chaque espèce, les régions alignées sur les transcriptomes et présentant du polymorphisme ont été utilisées pour calculer le nombre de sites ségrégeant (nombre de sites polymorphes, Eta) et le nombre moyen de différences observées entre deux séquences alléliques prises au hasard dans la population. Le D de Tajima a aussi été estimé à chacun des locus afin de déterminer si le transcriptome de ces espèces est impacté par des changements démographiques (effet global) ou si certains locus sont affectés par des processus sélectifs (Figures V-7 et V-8). Pour les espèces d'Aphroditiformia, aucun des trois estimateurs utilisés ne permet d'identifier de composantes environnementales. On remarque tout de même que toutes les espèces étudiées présentent des valeurs de D de Tajima globalement négatives ce qui peut s'expliquer par un phénomène d'expansion faisant suite à un goulot d'étranglement. On

retrouve aussi pour toutes les espèces (exceptée *B. segonzaci*) une queue de distribution des valeurs de D de Tajima atteignant des valeurs positives et élevées indiquant qu'une petite partie des séquences étudiées pourrait être marquée par de la sélection balancée, notamment pour les espèces polaires et *L. fimbriatium*, qui présentent de nombreux locus 'outliers'. La distribution du nombre de sites ségrégeant est particulièrement étroite vers des valeurs très faibles chez les 2 espèces hydrothermales, laissant penser que ces espèces ont été fortement impactées par des réductions d'effectifs au cours du temps. Cette constatation peut également être faite pour l'espèce *H. fuligineum* qui présente une distribution géographique circonscrite à la côte antarctique et pourrait être plus sensible aux effets des glaciations.



Figure V-24 : Boîtes à moustaches présentant la distribution des valeurs du D de Tajima, le nombre moyen de différences entre séquences et le nombre de sites ségrégeant pour les espèces du milieu côtier (vert), Antarctique (bleu) et hydrothermal (rouge) d'Aphroditiformia. Les boîtes à moustaches indiquent du centre vers l'extérieur, la valeur médiane, les quartiles ainsi que le premier et le dernier décile. Les croix indiquent les valeurs moyennes et les points les valeurs non comprises dans les boîtes à moustaches (outliers). Pour chaque espèce, le nombre de séquences utilisées est indiqué et pour chaque mesure les groupes statistiques identifiés grâce au test de Tukey (seuil de significativité 0,01).

Le cas des espèces de Terebellomorpha est différent. Les distributions du D de Tajima permettent de mettre en évidence une différence significative entre les espèces de l'environnement hydrothermal et les autres. En effet *A. pompejana* et *A. caudata* ainsi que l'espèce antarctique *Thelepus* sp. présentent des D de Tajima globalement plus faibles et plus

négatives que les autres espèces (celles-ci ayant une valeur médiane très proche de zéro). On retrouve aussi tout comme avec les Aphroditiformia des valeurs de D de Tajima globalement négatives à l'exception du cas de *T. lapidaria* pour laquelle la distribution est centrée sur zéro et significativement différente de celles des autres espèces. En revanche, les mesures de diversité ne diffèrent pas selon les environnements. Globalement, même si des différences significatives sont observées entre espèces, il ne semble y avoir aucune tendance liée à l'environnement en terme de diversité génétique.



Figure V-25 : Boîtes à moustaches présentant la distribution des valeurs du D de Tajima, le nombre moyen de différences entre séquences et le nombre de sites ségrégeant pour les espèces du milieu côtier (vert), Antarctique (bleu) et hydrothermal (rouge) de Terebellomorpha. Les boîtes à moustaches indiquent du centre vers l'extérieur, la valeur médiane, les quartiles ainsi que le premier et le dernier décile. Les croix indiquent les valeurs moyennes et les points les valeurs non comprises dans les boîtes à moustaches. Pour chaque espèce, le nombre de séquences utilisées est indiqué et pour chaque estimateur les groupes statistiques identifiés grâce au test de Tukey (seuil de significativité 0,01).

Les résultats suggèrent que certaines espèces ont subi une expansion démographique à la suite d'un goulot d'étranglement et présentent une diversité génétique plus réduite qui pourrait être également liée au mode de reproduction des espèces, notamment celles vivant dans l'environnement hydrothermal. Pour s'affranchir des effets démographiques/traits d'histoire de vie des espèces nous nous sommes alors intéressés à l'évolution des remplacements des acides aminés dans le polymorphisme des espèces afin de savoir si certains types de remplacements sont plus fréquents dans l'un des 3 environnements. Pour cela, nous avons identifié chaque position polymorphe dans les alignements protéiques de nos espèces et estimé la fréquence de l'allèle alternatif (SNP bi-allélique) en tolérant 20% de données manquantes (soit 4 séquences sur les 12 possibles pour *P. furcosetosa* et 20 pour les autres espèces). Les figures V-9 et V-10 représentent pour les Terebellomorpha et les Aphroditiformia les diagrammes de classe de fréquence pour la mutation alternative de chaque SNP non-synonyme de chaque espèce.

On constate ainsi une tendance généralisée chez les Aphroditiformia (excepté pour *P. furcosetosa*) montrant un excès de singletons non-synonymes pouvant traduire soit une expansion démographique pour la plupart des espèces, soit les effets d'une très forte sélection purifiante. La distribution apparemment bimodale observée pour *P. furcosetosa* est possiblement un artefact dû au fait que pour cette espèce, seuls 6 individus ont été utilisés, limitant ainsi l'interprétation de ce genre d'analyse.

Chez les Terebellomorpha, des différences entre les spectres de fréquences alléliques sont observées entre espèces avec un excès de fréquences intermédiaires par rapport aux singletons attendus par une sélection purifiante chez les espèces hydrothermales et *Thelepus* sp.. Ce cas de figure pourrait aussi bien être le résultat d'une introgression au sein des populations échantillonnées ou des effets de sélection balancée plus fréquents.



Figure V-26 : Spectres de fréquences de l'allèle alternatif aux positions polymorphes nonsynonymes dans les séquences protéiques pour les espèces de Terebellomorpha.



Figure V-27 : Spectre de fréquences de l'allèle alternatif aux positions polymorphes non synonymes dans les séquences protéiques pour les espèces d'Aphroditiformia.

D - Catégorisation du polymorphisme protéique et identification des remplacements préférentiels selon les espèces

Pour s'émanciper d'un effet lié à l'histoire démographique ou aux traits d'histoire de vie des espèces, une seconde étude a été réalisée sur les remplacements d'acides aminés dans le polymorphisme protéique des espèces, en tablant sur l'hypothèse qu'un polymorphisme neutre ne devrait pas entraîner de différences dans la fréquence des remplacements entre les différentes classes biochimiques de résidus protéiques d'une espèce à l'autre. Plus précisément, les positions des SNPs non-synonymes ayant été identifiées, une liste des remplacements a été

réalisée en identifiant les acides aminés concernés et en les classant en quatre catégories (aromatique ; chargé ; polaire ; apolaire). Ainsi le tableau V-3 présente pour l'ensemble des espèces, la fréquence de chaque type de remplacement en indiquant par un code couleur les transitions les plus fréquentes. Le tableau V-4 quant à lui est recentré sur les remplacements entre les différentes classes d'acides aminés (hétérotypiques) afin d'augmenter le contraste entre remplacements les plus fréquents et ceux les plus rares. Ce second tableau présente aussi la fréquence cumulée des remplacements d'une classe d'acides aminés vers une autre.

Table V-10 : Fréquences des différents types de remplacements entre acides aminés de même classe biochimique chez les espèces de Terebellomorpha et d'Aphroditiformia.

			Aron	atique L7 Hor	hatique chat	obite ^{L7} Polai	are Apole And Apole	one Constinue Long	atte	natione ^{1,7} AP	hate 27 Polai	hatter P	bile 27 house
a	Miliou côtion	T. lapidaria	0,07508	0,25460	0,30893	0,35812	0,00011	0,00022	0,00025	0,00102	0,00067	0,00100	
morph	Willed Cotler	A. edwardsi	0,09628	0,26191	0,28487	0,35514	0,00010	0,00014	0,00014	0,00062	0,00027	0,00051	
	A	Thelepus sp.	0,08480	0,21008	0,34623	0,35189	0,00032	0,00108	0,00064	0,00165	0,00093	0,00240	
ello	Antarctique	Amphitritides sp.	0,06157	0,18834	0,42130	0,32188	0,00008	0,00106	0,00058	0,00167	0,00147	0,00206	
ereb	Miliou bydrothormal	A. pompejana	0,07190	0,21962	0,35780	0,34556	0,00019	0,00048	0,00032	0,00142	0,00083	0,00187	
Ţ	willeu nyurothermai	A. caudata	0,07113	0,22134	0,35800	0,34203	0,00017	0,00082	0,00051	0,00209	0,00136	0,00256	
iformia	Miliou sôtion	P. furcosetosa	0,07181	0,21279	0,37547	0,33166	0,00026	0,00081	0,00048	0,00227	0,00146	0,00299	
	willeu cotier	A. gelatinosa	0,06248	0,22535	0,41629	0,28791	0,00025	0,00075	0,00030	0,00183	0,00131	0,00353	
	Antonetinun	H. fuligineum	0,07762	0,24360	0,32844	0,34187	0,00049	0,00088	0,00056	0,00236	0,00156	0,00263	
odi	Antarctique	H. crosetensis	0,07832	0,20874	0,36379	0,33267	0,00041	0,00213	0,00077	0,00343	0,00311	0,00662	i i
phr	Miliou budeeth erreal	L. fimbriatum	0,06339	0,24910	0,34645	0,33903	0,00015	0,00018	0,00012	0,00060	0,00035	0,00062	i i
4	willeu nydrothermal	B. seaonzaci	0.07904	0.24746	0.33535	0.33539	0.00022	0.00027	0.00014	0.00081	0.00043	0.00089	1

Table V-11 : Fréquences des remplacements entre classes d'acides aminés différentes chez les espèces de Terebellomorpha et d'Aphroditiformia.

	mattere							
	T lanidaria	A10	0140	A101	0.00102	0.00067	Q 00100	0.00227
C Milieu côtier	A. edwardsi	0.00011	0.00014	0.00014	0.00062	0.00027	0.00051	0.00179
nor	Thelepus sp.	0.00032	0.00108	0.00064	0.00165	0.00093	0.00240	0.00701
· Antarctique	Amphitritides sp.	0,00008	0,00106	0,00058	0,00167	0,00147	0,00206	0,00692
de la	A. pompejana	0,00019	0,00048	0,00032	0,00142	0,00083	0,00187	0,00512
Milleu nydrothermal	A. caudata	0,00017	0,00082	0,00051	0,00209	0,00136	0,00256	0,00750
	P. furcosetosa	0,00026	0,00081	0,00048	0,00227	0,00146	0,00299	0,00827
E Milleu cotier	A. gelatinosa	0,00025	0,00075	0,00030	0,00183	0,00131	0,00353	0,00797
li l	H. fuligineum	0,00049	0,00088	0,00056	0,00236	0,00156	0,00263	0,00847
o Antarctique	H. crosetensis	0,00041	0,00213	0,00077	0,00343	0,00311	0,00662	0,01648
	L. fimbriatum	0,00015	0,00018	0,00012	0,00060	0,00035	0,00062	0,00203
Milleu nydrothermal	B. segonzaci	0,00022	0,00027	0,00014	0,00081	0,00043	0,00089	0,00277

Quel que soit le sous-ordre des espèces et l'environnement considérés, la grande majorité des remplacements a lieu entre acides aminés de même classe et représente 98,3% des remplacements répertoriés, ce qui constitue un attendu neutre du polymorphisme non-

synonyme. Lorsque l'on se concentre sur les remplacements entre classes d'acides aminés, un différentiel apparaît entre espèces. Dans le cas des Aphroditiformia, les espèces de l'environnement hydrothermal ont une fréquence plus faible de remplacements hétérotypiques ($\approx 0,2\%$) par rapport à celles des deux autres environnements ($\approx 0,8\%$ au minimum). Pour les espèces de Terebellomorpha, on observe une plus grande variabilité inter-espèces de ces remplacements. Malgré cela, les deux espèces du milieu côtier présentent une fréquence de remplacements hétérotypiques inférieure aux autres espèces avec, respectivement, 0,32 et 0,18% pour *T. lapidaria* et *A. edwardsi* contre au moins 0,51% pour les espèces des autres environnements. Ces différences observables dans le milieu hydrothermal pour les Aphroditiformia et dans le milieu côtier pour les Terebellomorpha ne semblent pas imputables à un remplacement hétérotypique d'une classe particulière vers une autre.

Les fréquences centrées et réduites de ces remplacements ont permis de réaliser une ACP (figure V-11). Le premier axe de cette ACP explique 81% de la variance observée et permet de distinguer les espèces antarctiques de celles de l'environnement hydrothermal pour les deux sous-ordres mais de façon plus marquée chez les Aphroditiformia. Le deuxième axe explique quant à lui 12% de la variabilité observée. Cet axe permet de distinguer pour les Aphroditiformia, les espèces antarctiques de celles de l'environnement côtier breton, et pour les Terebellomorpha, celles de l'environnement côtier breton des espèces hydrothermales. Ainsi les deux axes cumulés permettent pour les Aphroditiformia de distinguer clairement les trois environnements. Pour les Terebellomorpha, la distinction entre environnements est moins évidente mais ne semble pas spécifiquement conditionné par l'axe 2. Seuls les espèces hydrothermales se regroupent entre elles sur la base de leurs remplacements hétérotypiques entre les différentes classes d'acides aminés.



Figure V-28 : ACP représentant les différentes espèces de Terebellomorpha (triangles) et d'Aphroditiformia (ronds) selon leurs fréquences centrées et réduites des remplacements entre classes d'acides aminés. L'environnement d'origine de chacune est indiqué par la couleur : vert - environnement côtier breton ; bleu – Antarctique ; orange – milieu hydrothermal. L'encadré indique le pourcentage de variance expliquée pour chaque axe.

Nous avons dans un second temps réalisé la même approche en considérant chaque changement d'acide aminé en particulier (soit 380 changements possibles ; figure V-12). Les deux premières dimensions de cette ACP expliquent respectivement 57 et 19% de la variance observée. Cette analyse ne permet pas de distinguer un quelconque effet environnemental et est fortement conditionnée par le fort pourcentage des remplacements s'effectuant au sein d'une même classe d'acides aminés, ce type de remplacement étant présumé neutre. Même si la distribution des espèces semble aléatoire, on peut tout de même noter un regroupement pour les Aphroditiformia des espèces hydrothermales et pour les Terebellomorpha des espèces côtières bretonnes et

Antarctiques. Il est fort probable que ces relations de proximité soient dues à la proximité phylogénétique de ces différentes paires d'espèces.



Figure V-29 : ACP représentant les différentes espèces de Terebellomorpha (triangles) et d'Aphroditiformia (ronds) selon leurs fréquences centrées et réduites des remplacements d'acides aminés. L'environnement d'origine de chacune est indiqué par la couleur : vert - environnement côtier breton ; bleu – Antarctique ; orange – milieu hydrothermal. L'encadré indique le pourcentage de variation expliquée pour chaque dimension.

E - Etude de la variabilité du Tm de la cMDH entre les différentes espèces selon le type d'environnement

Afin d'étudier un cas concret de l'effet de l'environnement sur la thermotolérance enzymatique des espèces, nous avons réalisé des mesures de Tm de la cMDH pour une partie des espèces dont nous avons étudié le polymorphisme. Ces mesures nous ont permis de calculer pour chaque cas, la moyenne de Tm et l'indice de dispersion des mesures (figure V-13). Les valeurs de Tm ne permettent pas de distinguer les environnements côtier et antarctique quel que soit le sousordre d'espèces considéré. Dans le cas des Aphroditiformia pour lesquels on dispose de mesures pour une espèce hydrothermale, cette moyenne ne semble pas non plus s'écarter particulièrement des mesures obtenues pour les autres environnements. En effet *L. fimbriatum* présente un Tm moyen de 41,3°C contre plus de 39°C pour *P. furcosetosa* et *H. crosetensis* provenant du milieu côtier breton et de l'Antarctique, respectivement.



Figure V-30 : Mesures de Tm (°C) et représentation de la dispersion par densité de classes (bandes latérales). Pour chaque espèce, le nombre d'observations, la moyenne de Tm et l'indice de dispersion (D) des points sont présentés sous le graphe.

Comme indiqué par la couleur sur la figure V-13, les valeurs de Tm tendent à s'agréger en plusieurs groupes, une possible réflexion de la thermostabilité des isoformes. Cette tendance apparaît évidente pour *H. fuligineum* pour laquelle les points se regroupent en trois groupes distincts. En ce qui concerne l'amplitude de la gamme de valeurs, représenté par l'indice de dispersion, des différences sont observables entre espèces. Pour rappel, des mesures de D = 1 correspondent à une distribution Poissonnienne centrée autour d'une valeur donnée. Des valeurs de D supérieures ou inférieures à 1 correspondent respectivement à une distribution dispersée ou à une distribution agrégée. Ainsi, cet indice ne nous permet pas ici de ressortir une tendance environnementale, mais plutôt de distinguer les deux sous-ordres. En effet pour l'ensemble des espèces d'Aphroditiformia les calculs d'indice de dispersion traduisent une surdispersion plus ou moins marquée des valeurs et donc potentiellement une diversification des isoformes dans une espèce donnée (i.e. une seule isoforme spécialisée ou ubiquitaire doit être

caractérisée par une valeur spécifique de Tm). Au contraire pour les espèces de Terebellomorpha, l'indice de dispersion révèle pour *A. edwardsi* et *Thelepus sp.* une agrégation des valeurs de Tm laissant supposer que chez ces espèces l'action de la sélection est plus contraignante que dans les autres cas avec cependant un distribution couvrant une plus large gamme de température pour l'espèce Antarctique. Pour *T. lapidaria* on observe une légère dispersion des valeurs de Tm mais cette dispersion est inférieure à ce qui est observé pour les Aphroditiformia. Ce qui laisse penser que pour les Terebellomorpha la sélection purifiante est plus intense que chez les Aphroditiformia.

Si l'on ne considère que l'amplitude de la distribution des valeurs de Tm, on peut voir que tout sous-ordre confondus, cette amplitude n'excède pas 15°C pour les environnements tempérés et hydrothermaux alors que cette amplitude peut atteindre les 30°C pour une espèce Antarctique, et couvrent des températures dépassant largement celles expérimentées par les espèces dans leur milieu. Il est donc probable que la tolérance thermique ancestrale de la cMDH étant largement supérieure aux besoins des espèces Antarctique, un relâchement de la sélection sur cette caractéristique enzymatique permet une plus grande liberté du polymorphisme.

5 - Discussion

A - Milieu hydrothermal, hyper variabilité thermique et température élevées

Comme nous l'avons vu précédemment, en raison des phénomènes répétés d'extinction recolonisation (Gilpin, 1991) et de la forte sélection purifiante induite par les conditions extrêmes (Vrijenhoek, 1997 ; Vrijenhoek, 2003), les espèces de l'environnement hydrothermal sont généralement caractérisées par de faibles diversités génétiques. De plus le mode de reproduction des *Alvinella* et certaines espèces de Polynoidae (Van Dover et al. 2000, Jollivet et al. 2000) avec choix du partenaire et appariement au sein des populations équivaut à une diminution des tailles efficaces des populations et potentiellement une diminution de la fécondité qui sont deux paramètres dont on sait qu'ils sont corrélés positivement avec la diversité génétique (Romiguier *et al.*, 2014-a ; Romiguier *et al.*, 2014-b). Cette réduction du polymorphisme chez les espèces hydrothermales est largement confirmée par notre étude chez les espèces de la famille des Polynoidae,

De plus les particularités physiques de cet environnement où la pression hydrostatique est élevée (Marteinsson *et al.*, 1999) et des températures dépassant parfois 50°C (Chevaldonné et

al., 1992; Le Bris et al., 2006) ont probablement amené à des adaptations pour répondre à la fois aux fortes amplitudes thermiques et aux variations 'haute fréquence' de la température (Somero & Hochachka, 1971; Somero, 1992). On peut donc s'attendre d'une part à observer une évolution moléculaire dirigée vers des isoformes beaucoup plus plastiques (e.g. efficacité catalytique d'une enzyme maintenue sur une large amplitude thermique) avec pour corollaire une baisse de la diversité génétique, et d'autre part, à une diversification et spécialisation des isoformes pour travailler sur une partie de la gamme des conditions rencontrées (duplications « adaptatives », polymorphisme balancé) (Somero & Hochachka, 1971), avec pour corollaire une augmentation de la diversité génétique. Ce maintien de la diversité n'est observé que chez les Alvinella même si les quatre espèces hydrothermales partagent les mêmes tendances quant aux types de remplacements d'acides aminés observés dans leur polymorphisme. Cette sélection sur les types et la fréquence des remplacements hétérotypiques chez les espèces hydrothermales pourrait soutenir l'hypothèse d'une sélection purifiante beaucoup plus marquée sur les protéines des espèces hydrothermales et plus particulièrement marquée pour les Aphroditiformia hydrothermaux qui présentent un taux extrêmement faible de transitions entre acides aminés de types différents (< 0,3%). La quasi-absence de remplacements hétérotypiques permettrait alors de maintenir d'une part la structure tridimensionnelle des protéines par la conservation de résidus hydrophiles sur les surfaces exposées de la protéine et les résidus hydrophobes au cœur de celle-ci en conservant un site actif ni trop flexible ni trop rigide pour assurer à la fois une spécificité de l'activité et une flexibilité permettant la réalisation de la fonction (Jiang et al., 2018). Ces données pourraient donc soutenir l'hypothèse d'une évolution des protéines hydrothermales vers des isoformes moins diversifiées mais plus plastiques (i.e. capables de fonctionner sur une large gamme de température), même si l'évolution vers une forme 'optimale' constitue un processus long pour trouver la bonne combinaison de mutations dans le polymorphisme existant. A ce titre, les mesures de Tm réalisées sur L. fimbriatum (Aphroditiformia hydrothermal) mettent en évidence une valeur moyenne nettement supérieure à celle des autres espèces en réponse à l'exposition à des températures environnementales plus élevées avec une gamme de valeurs de Tm relativement étroite en comparaison avec les autres espèces, ce qui laisse penser que cette distribution, bien que multimodale pourrait correspondre à une seule isoforme relativement plastique. Il convient néanmoins de souligner que contrairement aux Alvinellidae, les Polynoidae sont des espèces vagiles capables d'éviter les variations environnementales de trop grande amplitude et, donc susceptibles de vivre dans un environnement « extrême » mais beaucoup moins variable temporellement.

Il semble donc évident que l'extrême variabilité thermique spatiale et temporelle des sources hydrothermales dont les températures peuvent excéder 50°C constitue un facteur clé marquant le polymorphisme des espèces à la fois quantitativement et qualitativement et à l'échelle du génome comme à celle des protéines. Cependant cet environnement se caractérise aussi par un fort taux d'extinction/recolonisation des populations et pour ces espèces une sélection vers un mode de reproduction favorisant plutôt une plus faible fécondité avec des œufs riches en réserves vitellines. L'effet d'une forte sélection purifiante liée à la nature 'extrême' de l'environnement hydrothermal s'additionnant aux effets d'une fécondité limitée et du taux d'extinction des sites, on peut penser alors que cet environnement pourrait contraindre la diversité des populations hydrothermales et qu'il faut donc prendre en compte lorsque l'on s'intéresse à ces espèces.

Néanmoins, il convient de souligner le comportement atypique des polychètes Alvinella en termes de polymorphisme. Ces polychètes présentent en effet des diversités assez semblables à celles des espèces des autres environnements. On retrouve d'ailleurs un excès de mutants en fréquence intermédiaire chez les deux Alvinella ce qui a tendance à renforcer l'idée que la dynamique des habitats favorise le maintien de lignées alléliques différentes selon que les cheminées soient plus ou moins chaudes ou l'arrivée et le maintien d'allèles introgressés dans la population échantillonnée. Cette différence pourrait s'expliquer par le fait que les deux espèces d'Alvinella constituent des espèces pionnières avec de grandes capacités de recolonisation de l'habitat hydrothermal (Piccino et al. 2004). Jang et al. (2016) soulignent également cette particularité pour l'espèce A. pompejana lorsqu'elle est comparée à la plupart des autres espèces vivant en milieu hydrothermal. Différentes hypothèses peuvent être proposées pour expliquer cette diversité 'anormalement' élevée chez les espèces d'Alvinella en considérant d'une part, des capacités de dispersion conférant aux populations de plus grandes tailles efficaces (Vrijenhoek, 1997), le mélange de plusieurs fonds génétiques entre l'EPR sud et nord et la dorsale Pacifique-Antarctique générant de l'introgression d'allèles à certaines localités (c.f. chapitre 3), et enfin l'action d'une sélection balancée maintenant du polymorphisme en fonction de la dynamique des habitats « chauds » et « froids » comme ce qui a été observé dans le cas de la PGM chez A. pompejana (Piccino et al. 2004, Plouviez et al,. 2010, Bioy et al. soumis).
B - L'Antarctique, une stabilité thermique dans l'extrême

Le fait que l'Antarctique soit isolé thermiquement et géographiquement par le courant circumpolaire implique que les espèces qui y vivent sont à la fois soumises à des conditions thermiques extrêmes permanentes (températures gélives) et particulièrement vulnérables aux successions de périodes de glaciations entrainant une faible diversité génétique chez la plupart des espèces (Kis-Papo et al., 2003 ; Hewitt, 2004). Mais on observe aussi dans certains cas chez des espèces dont le mode de vie leur permet d'être moins vulnérables aux glaciations des valeurs de diversité dépassant l'attendu observé chez des espèces plus circonscrites à l'environnement côtier (Papot et al., 2016). Ces diversités élevées telles que celles observées chez les crinoïdes du genre Promachocrinus (Hemery et al. 2012) ou le krill Antarctique (Goodall-Copestake et al., 2010) sont sans doute à mettre en lien avec les abondances extrêmement élevées de ces espèces le long de la côte antarctique et la forte connectivité circumpolaire de leurs populations. Bien que les espèces de polychètes choisies soient très abondantes sur le plateau continental antarctique, nos données génomiques semblent plutôt confirmer les signes d'une expansion des espèces suite aux derniers épisodes de glaciation. Toutes présentent en effet une distribution des valeurs de D de Tajima centrées sur des valeurs négatives, et un excès de singletons typique d'une expansion démographique consécutive à un goulot d'étranglement. Cependant on ne retrouve pas la faible diversité génétique attendue. En effet ces espèces montrent un niveau de polymorphisme semblable dans le cas des Terebellomorpha ou même supérieur dans le cas des Aphroditiformia à celui retrouvé chez les autres espèces. Ceci est probablement dû au fait que lors des glaciations, on assiste en effet au sein des refuges glaciaires à une diminution effective de la taille efficace et de la diversité génétique des populations, mais la multiplicité de ces refuges permet lors des retraites glaciaires une remise en contact des différentes populations dans les zones recolonisées avec une augmentation locale de la diversité génétique (Petit et al., 2003). De plus, l'estimateur considéré est important pour caractériser la diversité génétique de ces populations. On constate que dans les zones refuges après une glaciation, alors que l'on observe effectivement une diminution de la diversité génétique de Nei (1973) due à l'effet de fondation, la richesse allélique est en partie maintenue (Widmer & Lexer, 2001). Et c'est donc ce maintien, même partiel, de la richesse allélique qui pourrait expliquer en grande partie le fait que les π_N et π_S ne sont pas particulièrement plus faibles que ceux des espèces des autres environnements.

Dans le cas du métabolisme et plus précisément de la cMDH, le compromis qui existe entre stabilité et efficacité de la molécule supposerait que l'enzyme soit fonctionnelle dans une

gamme de température inférieure à 5°C et donc fortement contrainte du point de vue de sa structure primaire la rendant de ce fait extrêmement sensible à une hausse de la température (Arnold et al., 2001). On retrouve cependant des valeurs moyennes de thermostabilité pour les espèces antarctiques similaires à celle mesurées pour les espèces tempérées mais avec des valeurs maximale et minimale beaucoup plus dispersées, suggérant l'existence de plusieurs isoformes pouvant travailler à des températures différentes. Il faut prendre en considération le fait que l'Antarctique avant la fragmentation du Gondwana (185-60 Ma) (Hawkesworth et al., 1999) était en contact avec des habitats à plus hautes latitudes. Jusqu'à environ 34 Ma, le climat autour du continent était aussi plus tempéré et les populations n'étaient pas séparées de celles à plus basse latitude. Les ancêtres des espèces antarctiques actuelles étaient donc potentiellement distribuées ou en contact avec des populations soumises à des températures moins extrêmes comme ce qui s'observe chez certaines espèces de krill (Goodall-Copestake et al., 2010). On retrouve ainsi que des protéines psychrophiles ne présentent que peu de différences avec leurs homologues mésophiles en raison d'une évolution plutôt ciblée sur la flexibilité de la molécule permettant une rétention plus grande du polymorphisme non-synonyme neutre (Arnold et al., 2001). Dans ce cas particulier, l'augmentation de la flexibilité de la molécule pour fonctionner à des températures plus basses ne se traduit pas forcément par une diminution de sa température de dénaturation et l'enzyme est donc à même de pouvoir travailler sur une gamme plus étalée de températures que celle rencontrée dans son environnement. Ceci pourrait expliquer le fait que les mesures de Tm pour les espèces antarctiques étudiées se distribuent sur une gamme de températures identiques à celles des espèces des autres environnements. On retrouverait ainsi des valeurs moyennes de Tm élevées héritées des formes ancestrales mais avec de nombreuses isoformes dues à un relâchement de la sélection sur ce trait phénotypique depuis l'isolement du continent. Ce résultat fait écho à plusieurs travaux menés sur les réponses aux chocs thermiques via l'expression d'HSP70 chez plusieurs espèces Antarctiques et ayant montré pour certaines espèces une perte de cette capacité (par mutation des sites promoteurs) et pour d'autres le maintien du caractère ancestral avec une température d'induction supérieure à ce que l'on peut rencontrer en Antarctique (pour revue Clark & Peck, 2009). Les auteurs suggèrent comme dans notre cas que ce maintien de trait ancestral peut résulter du fait que la forte stabilité thermique de cet environnement n'induit pas de pression de sélection sur ce trait.

C - Le milieu côtier tempéré, un milieu hypervariable mais cyclique

Si l'on compare l'environnement côtier tempéré au milieu hydrothermal ou antarctique, d'un point de vue thermique celui-ci est clairement intermédiaire. Cependant, contrairement aux deux autres environnements, cette variabilité thermique est étroitement associée au cycle des saisons, des marées et au rythme circadien. De plus cette variabilité et le stress qu'elle peut occasionner dépendent grandement du régime local des marées plus ou moins marqué selon les localités (Helmuth *et al.*, 2006), créant ainsi une véritable mosaïque environnementale selon les conditions locales d'émersion. Cette hétérogénéité géographique et temporelle de la température le rapproche de l'environnement hydrothermal mais introduit ici une possibilité d'acclimatation pour les espèces qui y vivent. D'un autre côté, comme les espèces antarctiques, les espèces tempérées boréales ont subi les successions glaciaires et on s'attend à observer une diminution de la diversité génétique de Nei (1973) due à la contraction des aires de distribution des espèces au cours des derniers millénaires avec cependant un maintien de la richesse allélique (Widmer & Lexer, 2001).

Les données génomiques que nous avons obtenues sur les espèces du milieu tempéré montrent des niveaux de diversités synonyme et non synonyme semblables à ceux trouvées pour les espèces antarctiques avec pour les aphroditiformes des distributions du D de Tajima centrées autour des mêmes valeurs négatives. En revanche, les espèces de Terebellomorpha sont les seuls à présenter une distribution centrée sur des valeurs proches de zéro. Ce décalage de distribution peut s'expliquer par de la rétention de polymorphisme ancestral ou à des mélanges de populations issues de refuges différents et sélectionnées différentiellement dans le temps ou l'espace (Bierne *et al.*, 2013).

Les distributions des valeurs de Tm pour la cMDH sont particulièrement intéressantes. On retrouve des valeurs moyennes comprises entre 32,4 et 39,4°C ne permettant pas de les distinguer des espèces antarctiques mais les amplitudes thermiques couvertes par ces Tm n'excèdent pas 15°C. Comme pour les espèces hydrothermales pour lesquelles les valeurs mesurées sont plutôt resserrées, on peut supposer que les faibles différences de Tm observées pour les espèces tempérées reflètent également une stratégie d'adaptation plutôt orientée vers la mise en place d'isoformes relativement plastiques capables de travailler sur une large gamme de températures.

D - Relations entre traits d'histoire de vie et diversité synonyme

En 2014, Romiguier et al. ont cherché à corréler la diversité génétique des espèces avec leurs stratégies écologiques telles que l'investissement parental, la quantité de descendants produit, la taille des adultes, des propagules, etc. Pour ce faire, ils ont utilisé les données de transcriptomes de 76 espèces distinctes de métazoaires, en caractérisant les diversités synonyme et non-synonyme en séquençant de 2 à 10 transcriptomes d'individus par espèce. Leurs analyses ont permis de montrer que parmi les espèces choisies, la diversité synonyme varie d'un facteur 100 entre les espèces avec un π_s moyen de 0,001 pour les espèces les moins diversifiées (notamment les vertébrés) contre 0,1 chez la plupart des invertébrés marins n'ayant pas d'investissement parental (larves pélagiques). Ceci place donc nos espèces d'intérêt dans la catégorie des espèces présentant une faible diversité synonyme (valeurs moyennes comprises entre 0,002 et 0,02) malgré des caractéristiques plutôt en faveur de taille efficace élevée, forte fécondité et larves pélagiques. Leurs travaux mettent aussi en avant une forte corrélation négative entre diversité synonyme et taille des propagules, cette taille étant elle-même négativement corrélée à la quantité de propagules produits. Ceci devrait impliquer une forte diversité pour nos espèces au vu de leur fécondité et taille de propagules (Tableau V-1), ce qui n'est apparemment pas le cas. Or les auteurs réussissent aussi à identifier une corrélation négative entre taille de l'individu adulte et diversité génétique, mais cette corrélation met surtout en avant le rôle de l'investissement parental sur la diversité génétique de l'espèce. En effet les espèces qui présentent des niveaux de diversité synonyme inférieurs à l'attendu de leur prédiction selon la taille moyenne des adultes sont aussi celles qui s'investissent le plus dans leur descendance. Ainsi il est probable que la fécondation interne explique en partie les faibles valeurs de diversités observées chez les Alvinellidae. Dans le cas des autres espèces, il est très probable que l'on observe comme chez d'autres Polynoidae une rétention des œufs sous les élytres de certains (Daly, 1972), ou pour certains Terebellidae une rétention des œufs à l'intérieur des tubes construits par la femelle ou au sein de structures gélatineuses regroupant les œufs (McHugh, 1993). Ces comportements et peut-être d'autres pourraient donc en partie expliquer les niveaux de diversité synonyme observés chez nos Annélides marines. Une autre observation particulièrement intéréssante est que l'on retrouve chez les Aphroditiformia une corrélation positive entre polymorphisme synonyme et non-synonyme alors que dans le cas des Terebellomorpha, la diversité non-synonyme semble contrainte et n'excède pas 0,0025 pour l'ensemble des espèces tous environnements confondus. Ceci pourrait traduire une action plus forte de la sélection purifiante sur les Terebellomorpha, potentiellement en lien avec le fait que ces animaux sont relativement peu mobiles. On peut en effet penser que ceci les rend plus vulnérables aux différentes contraintes environnementales que les Aphroditiformia qui ont la possibilité de se déplacer et possiblement de pratiquer l'évitement. Une autre possibilité qu'il faut tout de même prendre en compte est que l'écart observé entre nos mesures de diversité et celles de Romiguier et al. (2014) peut résulter du fait qu'en calculant ces indices de diversité sur des lectures de 120 paires de bases, notre analyse pourrait légèrement sous-estimer cette diversité (par opposition au séquençage de transcriptomes de Romiguier et al., 2014).

6 - Conclusion

La température a donc un effet puissant sur la diversité génétique et phénotypique des annélides en milieu marin, notamment dans les environnements extrêmes. En effet la température contraint intrinsèquement la structure des protéines qui doivent pour une valeur donnée présenter une structure moléculaire suffisamment flexible de sorte que sa fonction soit assurée tout en maintenant suffisamment de rigidité pour éviter la dénaturation de la structure tridimensionnelle. Cette contrainte est particulièrement marquée dans le milieu hydrothermal où la forte pression hydrostatique et les températures élevées imposent une forte pression de sélection différente selon la température de l'habitat.

De plus, le fort taux d'extinction/recolonisation des sites s'additionnant à la sélection purifiante de l'environnement hydrothermal implique par son effet sur la taille efficace des populations des niveaux de diversité bien plus faibles que ceux que l'on retrouve dans d'autres environnements. Cependant il reste difficile d'estimer le pouvoir respectif de chacun de ces deux facteurs sur les patrons observés. Cet environnement présente tout de même une signature qui lui est propre quant à la fréquence des remplacements hétérotypiques d'acides aminés que l'on peut raisonnablement imputer aux spécificités thermiques de cet environnement.

En revanche la comparaison entre les espèces d'Antarctique et celles du milieu côtier tempéré met en évidence le rôle prédominant des épisodes de glaciations sur la diversité génétique actuelle des populations. Ainsi de ce point de vue, la distinction entre ces deux environnements thermiquement contrastés n'est pas possible avec les outils génétiques utilisés dans cette étude. Cependant la variabilité des valeurs de Tm de la cMDH permet quant à elle de mettre en évidence des pressions de sélection purifiante plus forte sur cet enzyme chez les espèces de l'environnement tempéré par rapport à celles de l'Antarctique.

Ainsi bien que le choix des espèces ait été fait afin d'éviter tout biais lié aux traits d'histoire de vie, on ne peut s'émanciper des spécificités indirectes liées à l'environnement et n'impliquant

pas directement la température. Tout comme Romiguier et al. (2014), notre étude montre que les événements démographiques passés ont un effet puissant sur la diversité actuelle des populations et ces effets ont la capacité de masquer l'effet plus ciblé de la sélection dans le génome. Malgré cela, il est tout de même possible d'identifier l'effet de la variabilité thermique de l'environnement en multipliant les approches et en exploitant le fait que ces approches sont plus ou moins sensibles aux traits d'histoire de vie, aux histoires démographiques ou à la sélection.

VI – Conclusion, Discussion, Perspectives

1 - Conclusion / Discussion

A - Sélection et Evolution

Les espèces poïkilothermes comme les ectothermes ont une température corporelle qui ne peutêtre régulée par le métabolisme de l'individus, en revanche les ectothermes contrairement à ces premières vont pouvoir ajuster leur comportement pour se maintenir dans une gamme de température qu'ils tolèrent. Ainsi des études ont cherché un lien entre la grande diversité allozymique que l'on retrouve chez certaines espèces poïkilothermes et la variabilité environnementale (Somero & Soulé, 1974). Dans ce genre de recherche l'important est de prendre en compte un système pour lequel on est certain qu'il subit l'influence de la composante variable considérée, tels que les systèmes enzymatiques des espèces ectothermes ou poïkilothermes et leur sensibilité à la température (Somero, 2004). Ces premières études n'ont pas permis de corréler cette diversité allozymique avec la variabilité environnementale, les auteurs émettent l'hypothèse que la variabilité observée dans une population reflète en fin de compte l'ancienneté de l'espèce dans son environnement, la stabilité de celui-ci ainsi que la complexité des interactions trophiques au sein de cette environnement (Somero & Soulé, 1974). Cette théorie fait tout de même de fortes hypothèses, en effet elle considère que l'hétérozygotie constitue le plus généralement un avantage et que l'on assiste au cours des temps géologiques à un cumul de sélection directionnelle (purgeant la diversité) et de sélection stabilisante permettant l'accumulation d'allèles. Plus récemment une étude de Romiguier et al. (2014) a cherché à expliquer les différences de diversité génétique observées entre métazoaires. Ils ont ainsi pu mettre en évidence une absence de relation entre la distribution des espèces et leur diversité génétique, mais à l'inverse leur étude a mis en avant le lien étroit qui existe entre traits d'histoire de vie, et plus particulièrement la stratégie de reproduction et l'investissement parental avec cette diversité intra-espèce.

Ces différences de variabilité thermique peuvent aussi jouer un rôle majeur dans la structuration des populations. Nous avons pu montrer grâce à nos travaux que la température en tant que telle constitue un moteur de la diversification des espèces chez les invertébrés marins. En effet, les populations exposées à des gradients de température ou à une forte variabilité de celle-ci développent des mécanismes en réponse à ces variations pour mieux coloniser l'environnement,

soit à travers l'acquisition d'une forte plasticité physiologique et des capacités de réponse à très court terme, soit à travers l'acquisition d'une plus grande diversité génétique leur permettant une meilleure adaptation locale. Ainsi, chez Alvinella, nous avons pu mettre en évidence de l'introgression préférentielle de fond génétique en fonction de la température de l'habitat malgré l'existence d'une barrière au flux de gènes, ainsi qu'une proportion importante de locus présentant un excès d'hétérozygotie potentiellement due à un maintien adaptatif de polymorphisme dans des localités étudiées. Cette différence de diversité génétique a aussi pu être mise en évidence chez l'espèce antarctique H. fuligineum pour laquelle la population de Péninsule à la fois intrinsèquement et anthropologiquement exposée à des températures plus hautes et variables que celles de Terre Adélie présente un niveau de diversité légèrement supérieur. En revanche si l'on se place à l'échelle des espèces, nos résultats sont plus contrastés que ce que montre Romiguier et al. (2014). En se basant sur des espèces aux traits d'histoire relativement moins contrastés que ce qui peut être observé dans leur étude entre un papillon et une tortue, nos données laissent penser que si l'empreinte d'une sélection diversifiante existe à l'échelle du génome, elle est tout de même limitée ou tout du moins rapidement masquée par les processus démographiques, les isolements géographiques et les reprises de flux migratoires entre les populations d'une même espèce. En effet contrairement à notre hypothèse qui supposait que la diversité génétique étaitdirectement corrélée à la variabilité de l'environnement, il apparaît que les espèces du milieu hydrothermal, pourtant présenté comme l'un des environnements les plus variables de notre planète, soient celles qui ont les niveaux de diversité synonyme et non-synonyme les plus faibles par rapport à ceux retrouvés chez les espèces des milieux tempérés ou polaires, en tout cas chez les annélides polychètes marines. Dans ce cas, cette réduction du polymorphisme peut s'expliquer par un taux d'extinction/recolonisation des sites hydrothermaux très important qui entraîne une forte réduction de la taille efficace des espèces sessiles ou à la mobilité limitée associée à cet environnement.

D'un point de vue qualitatif, quel que soit le niveau de polymorphisme non synonyme des espèces dans un environnement donné, de façon logique la grande majorité des mutations nonsynonymes correspondent à des transitions entre acides aminés ayant les mêmes propriétés structurelles et fonctionnelles (e.g. polaire vs. polaire) et permettent ainsi d'assurer un maintien de structure et de stabilité des protéines (Zeldovich et al., 2007). De plus, la plupart de ces mutations apparaissent sous la forme de singletons et constitue donc pour la plupart des mutations neutres, ou faiblement délétères encore non purgées par la sélection purifiante. Bien que moins fréquentes que les autres, les transitions hétérotypiques montrent des biais mutationnels diagnostiques de chaque type d'environnement, ces biais étant communs aux deux sous-ordres de polychètes analysés ici. Ces mutations diagnostiques pourraient refléter soit un biais dans la composition des génomes pour certains codons (Wang & Lercher, 2010), soit l'action d'une sélection orientée vers certains acides aminés dans une logique d'augmenter ou diminuer la flexibilité de certains variants vis-à-vis des conditions thermiques prévalant dans certains environnements. Cependant, lorsque l'on s'intéresse précisément aux transitions entre acides aminés dans leur ensemble, ces tendances disparaissent même si l'on retrouve des cas de regroupements entre espèces pour certaines paires d'environnement ou sous-ordres. Cette perte de signal environnemental et ce différentiel observé entre les deux sous-ordres montre que l'histoire phylogénique des espèces marque plus profondément le génome et, que les rapprochements observés sont attribuables à un biais mutationnel propre aux génomes d'espèces proches (Berezovsky & Shakhnovich, 2005). Cette perte du signal environnemental confirme alors l'idée qu'une partie des mutations non-synonymes au sein d'une même classe d'acides aminés est neutre. Lorsque l'on s'intéresse aux propriétés enzymatiques de la cMDH, des différences sont observables entre espèces quant à la tolérance thermique de l'enzyme (Tm), avec une tolérance supérieure de la cMDH chez l'espèce hydrothermale et la plus thermophile (L. fimbriatium). En revanche la dispersion des valeurs d'activité pour chaque espèce semble là aussi plus en lien avec le sous-ordre considéré qu'avec l'environnement auquel appartient l'espèce. Ainsi sur la base d'observations qualitatives, les caractéristiques du génome telles que les échanges préférentiels entre acides aminés ou l'intensité de la sélection opérant sur la tolérance thermique des enzymes suivent des tendances plus liées à l'histoire évolutive des espèces sur le long terme qu'à l'histoire environnementale récente des espèces.

B - Sélection et Différenciation

Lorsque l'on parle de différenciation des populations, la question du niveau de différenciation se pose de même que cette question se pose lorsque l'on étudie des cas de populations en voie de spéciation puisqu'il s'agit d'un véritable continuum le long duquel, en partant de deux populations d'une même espèce, on va observer un degré plus ou moins important d'isolement reproductif et de différenciation génétique amenant à deux espèces distinctes (Nosil *et al.*, 2009b). Ainsi on peut considérer que tant que ces deux conditions ne sont pas totalement validées nous sommes peut-être en phase de spéciation, mais les entités concernées sont belles

et bien des populations d'une même espèce. Alors quand deux populations sont exposées à des conditions différentes, la sélection divergente va entrainer une divergence plus ou moins importante des populations selon son intensité et le nombre de locus concernés (Nosil et al., 2009b). Ainsi même une différenciation ne touchant qu'un gène peut avoir de larges répercutions sur le génome en raison du balayage sélectif qu'elle peut entrainer (Smith & Haigh, 1974). En effet des locus neutres liés à un gène sous sélection vont voir leur fréquence allélique augmenter avec celle du premier, et cet effet sera directement corrélé au ratio r/s avec r le taux de recombinaison et s le coefficient de sélection (Barton 2000). Ce genre de mécanisme va donc entrainer une hétérogénéité de la différenciation le long du génome pour ces populations qui sera directement corrélée à la proximité au site sous sélection (Via & West, 2008). Le flux de gènes tient aussi son rôle dans la différenciation des populations, ainsi Egan et al. (2008) ont comparé différentes populations de scarabées définies par leurs localités géographiques et leur plante hôte, ils ont ainsi déterminé quels étaient les locus intervenant dans la préférence en matière de plante hôte, et ils ont mis en évidence le fait que ces locus « outliers » et les locus neutres ne traduisaient pas la même histoire évolutive. En effet, la construction d'arbres populationnels utilisant les premiers locus mettait en évidence la proximité écologique entre les populations et de quelle façon la sélection divergente avait maintenu deux lignées écologiques pour ces locus. A l'inverse la même analyse réalisée avec les locus neutres a révélé quant à elle les relations de proximité géographique et mettait ainsi en évidence les opportunités de flux de gènes plus grandes entre certaines populations. Mais la sélection écologique peut aussi renforcer la différenciation des populations en générant des isolements reproductifs qui vont alors limiter ce flux de gènes (Nosil et al., 2009a). Mais comme nous l'avons vu précédemment, même s'il y a isolement des fonds génétiques, l'un des critères nécessaires pour parler de différenciation des populations ou de spéciation est la distinction des fonds génétiques. Ainsi, même isolées, le facteur temps est capital pour permettre une différenciation des populations suffisante pour générer cette divergence (Coyne & Orr, 1989).

Nos travaux montrent bien l'effet majeur de la limitation du flux de gènes chez *Alvinella* pour laquelle se cumulent l'effet d'une barrière physique et, pour *A. pompejana*, potentiellement celui de barrières pré- et post-zygotiques. Cette barrière datée à 1-2 millions d'années (Plouviez *et al.*, 2009) cumulée à un taux d'accrétion plus fort au sud de la barrière ont par divergence neutre et, par sélection divergente, généré une telle différenciation des populations de part et d'autre que ceci se retrouve dans les différences de locus communs entre populations étant ou non du même côté de cette barrière. Mais cette barrière n'est pas complètement perméable, et cette semi-perméabilité suit des modalités différentes pour nos deux espèces avec une zone de

contact à 7°S pour A. pompejana et à 9°50N pour A. caudata. Cette situation en miroir laisse penser à une possible évolution de ces deux espèces en allopatrie à laquelle aurait succédé une recolonisation de l'ensemble de l'EPR et qui expliquerait leur syntopie actuelle. A l'inverse notre étude d'H. fuligineum en Antarctique révèle la forte homogénéité des fonds génétiques des populations de Péninsule et de Terre Adélie, traduisant un flux de gènes à l'échelle du continent ou un temps d'isolement suffisamment court pour ne pas permettre une différenciation des populations par dérive identifiable grâce à nos analyses. Pour ce qui est de l'étagement en Terre Adélie, nos résultats montrent qu'il existe effectivement une différenciation des fonds génétiques détectable sur la composition en locus, mais un maintien de flux de gènes entre ces étagements et potentiellement de la sélection balancée permettent aussi de limiter la différenciation que l'on observe sur quelques locus identifiés sous sélection par l'habitats. En effet un grand nombre de ces gènes présentent un excès d'hétérozygotes dans les deux habitats et la sélection balancée en contraignant l'évolution de ces gènes pourrait limiter d'une part l'accumulation de divergence entre ces habitats et favoriser les flux de gènes pour maintenir ce polymorphisme adaptatif. Nos travaux montrent aussi l'impact de la sélection sur le flux de gènes en lui-même. On constate en effet chez Alvinella que le flux de gènes déjà limité (discontinuité géographique, barrières pré- et post-zygotiques), est en plus conditionné par l'habitat et la localité. Selon les locus considérés, on constate en effet que le passage de rares allèles du sud vers le nord est selon les cas plus probable au sein d'un type d'habitat que dans l'autre et ces tendances varient selon les localités. Ainsi, alors que l'on ne parvient pas à identifier de différences majeures en comparant les fonds génétiques entre localités ou entre habitats, d'une part une sélection divergente opère sur l'introgression du fond génétique nord par le sud et d'autre part l'accumulation aléatoire dans le temps de divergence entre les localités semble aussi conditionner cette introgression comme ce qui a déjà été montré chez des populations isolées de drosophiles (Coyne & Orr, 1989). Ainsi divergence (par mutation et dérive) et sélection divergente s'additionnent pour conditionner le flux de gènes et générer de la différenciation entre populations.

Dans le cas de figure d'*A. pompejana*, une approche de type « δaδi » (Le Moan *et al.*, 2016) permettrait de définir l'histoire évolutive de ces populations et de savoir notamment si la situation actuelle correspond à un isolement strict avec un début de spéciation, un isolement avec migration ou à une remise en contact secondaire des populations après le relâchement de la barrière. Des tests préliminaires que nous avons réalisés laissent penser que l'on se trouve dans une situation ressemblant au modèle d'isolement strict avec donc un début de spéciation (Figure VI-1). Dans un cas comme celui-ci, le choix de la méthode de détection des locus

« outliers » est importante puisque les méthodes conventionnelles se basent sur la comparaison entre locus ou sur des corrélations entre fréquences alléliques et variable environnementale. Or ce genre de méthode génère une quantité de faux positifs (faux « outliers ») non négligeable qui peut être diminuée en prenant en compte des différences de taux d'expansion, de taux de mutation ou l'effet d'autostop (Villemereuil & Gaggiotti, 2015). Un autre moyen de s'émanciper de ce problème est l'analyse des locus au cas par cas par le biais d'approches bayesiennes limitant ainsi encore l'obtention de faux positifs (Foll & Gaggiotti, (2008). Enfin l'utilisation de programmes dédiés tels que LFMM (Modèle Mixte à Facteurs Latents ; Frichot, 2012) nous permettrait de pouvoir valider l'hypothèse de la sélection locale, LFMM étant parmi les programmes existants celui qui présente le meilleur compromis dans des situations telles que la nôtre où plusieurs locus sont différemment impactés et où la structure des populations suit des modalités complexes habitat et localité dépendantes (Villemereuil *et al.*, 2014).



Figure VI-31 : Test d'analyse δaδi réalisée avec le modèle d'isolement strict. Haut-gauche : matrice de densité présentant le nombre d'allèles alternatifs en fonction de leur abondance dans chaque population (Nord en abscisse, Sud en ordonnées). Haut-droite : résultat du modèle pour une situation d'isolement strict pour deux populations avec le même nombre d'individus et le même nombre de locus. Bas-gauche : matrice des résidus de la comparaison du modèle avec les données réelles. Bas-droite : Diagramme de densité du nombre de résidus en fonction de leur valeur.

2 - Perspectives

La température et ses variations spatio-temporelles bien qu'elles jouent un rôle central dans la différenciation des populations et la spéciation n'ont donc qu'un effet limité et ne laissent que des traces éphémères à l'échelle du temps de l'évolution. Un moyen d'identifier une signature directe de l'environnement sur cet équilibre qui existe au sein du milieu cellulaire serait d'étudier les propriétés et les patrons d'expression de certaines protéines chaperonnes telles que celle de réponse aux chocs thermiques puisque ce sont elles qui assurent l'intégrité du fonctionnement cellulaire lors de stress environnementaux (Hofmann, 1999) et de lier cette diversité fonctionnelle à leur diversité génétique. Il a ainsi pu être mis en évidence chez des espèces sœurs de Krill une perte adaptative de diversité d'HSP70 chez l'espèce expérimentant les températures les plus basses et stables (Papot et al., 2016). Un autre moyen serait de chercher une réponse physiologique globale en étudiant les rendements métaboliques des espèces en fonction de la température ou d'étudier l'ensemble des protéines impliquées dans une voie métabolique en particulier comme le cycle de Krebs ou la voie de synthèse des acides aminés. On peut en effet penser que si cet équilibre cellulaire est contraint par la sélection de façon globale, une réponse adaptative serait perceptible à l'échelle de ces voies métaboliques à défaut de l'être à celle des enzymes elles-mêmes (Pörtner, 2002).

On a pu voir dans nos analyses que la duplication de gènes peut poser problème lorsque que l'on souhaite utiliser l'information apportée par les fréquences alléliques. En effet ces duplications peuvent être considérées comme de l'allélisme et générer du polymorphisme fictif et non informatif dans les jeux de données. Mais la duplication de gènes et le maintien de copies constitue un moyen évolutif efficace pour s'adapter à un environnement variable en produisant rapidement du polymorphisme fonctionnel (Kondrashov, 2012). En effet, alors que l'une des copies maintient sa conformité et un fonctionnement de base, la dérive des autres versions par relâchement de la pression de sélection permet une véritable exploration des possibilités structurales et fonctionnelles de celles-ci pouvant aboutir à l'obtention d'une adaptation pour l'espèce. Il faut quand même prendre en compte le fait que pour qu'une duplication soit adaptative, elle doit balayer dans la population ce qui a aussi un prix pour l'espèce en termes de perte de diversité à d'autres locus. Mais dans des populations comme celles des sources hydrothermales où les extinctions/recolonisations sont nombreuses, on peut penser qu'un tel mécanisme peut en théorie facilement montrer son efficacité lors de la recolonisation de nouveaux sites. Il serait alors possible de voir si cet environnement dont les variations thermiques sont stochastiques et donc probablement plus stressantes pour les espèces, favorise

chez celles-ci un nombre plus important de gènes multi-copies que chez les espèces vivant dans des environnements moins variables en intensité ou en fréquence.

Un autre moyen de répondre rapidement à une nécessité environnementale est de jouer sur les patrons d'expression de ses gènes, soit par modification des séquences des régions promotrices, soit par modification épigénétique qui peut être transmise à sa descendance, par exemple par les éléments transposables (Mirouze & Paszkowski, 2011). Bien que les mécanismes d'action et l'impact des éléments transposables restent encore méconnus, on sait aujourd'hui qu'ils ont un fort impact sur l'hétérogénéité du taux de recombinaison des gènes dans le génome de nombreuses espèces et peuvent même influencer les processus de spéciation lors de la remise en contact de populations préalablement isolées (Rose & Doolittle, 1983 ; Giraud *et al.*, 1997 ; Hegarty & Hiscock, 2005) ou constituer des facteurs d'isolement entre des individus faisant face à des stress environnementaux différents. Ces stress biotiques ou abiotiques sont à même de modifier le taux de transposition des éléments dans le génome (de Jesus Ferreira *et al.*, 2001) et conduire à des incompatibilités génétiques entre individus différentiellement stressés. La diversité et l'abondance des éléments transposables entre espèces pourraient donc être gouvernées par l'hétérogénéité et la variabilité environnementale notamment chez les annélides et les mollusques (Thomas-Bulle *et al.*, soumis).

En termes de perspectives sur la méthodologie employée, il serait également intéressant d'augmenter la couverture de nos jeux de données en re-séquençant certaines banques individuelles déjà utilisées, notamment pour certaines localités géographiques souséchantillonnées. Ceci permettrait d'éviter les biais liés à l'échantillonnage des gènes entre espèces et une meilleure couverture de certains transcriptomes, tout en augmentant notre couverture en locus communs afin d'affiner nos observations concernant la structuration des populations d'Alvinella et d'H. fuligineum. De façon plus concrète, il serait aussi judicieux d'augmenter le nombre d'individus et de localités échantillonnées. En effet, dans le cas d'A pompejana, ceci nous permettrait d'une part de confirmer l'existence d'une zone de contact secondaire à 7°25S. D'autre part, l'échantillonnage de populations en provenance de la ride Pacifique-Antarctique, de la zone des Galápagos, et de localités plus au nord nous permettrait d'identifier plus clairement l'origine géographique de certains fonds génétiques détectés dans nos analyses. Ceci nous permettrait aussi d'obtenir une vue d'ensemble sur la répartition de cette espèce et potentiellement de définir plus en détail les patrons de connectivité longue distance entre populations. Dans le cas d'H. fuligineum, disposer de plus de localités autour de l'Antarctique permettrait d'une part, de vérifier le fait que l'espèce présente une distribution circum-Antarctique, et d'autre part, de préciser les relations avec d'autres populations ou espèces proches ayant colonisé les îles sub-Antarctiques afin d'identifier d'éventuelles zones refuges ayant permis à l'espèce de se maintenir lors des périodes de glaciation.

Enfin, utiliser des espèces vagiles et moins sujettes aux phénomènes d'extinction/recolonisation au niveau des sources hydrothermales permettrait de tester l'effet d'une migration orientée sur la différenciation locale des populations entre habitats. Ainsi l'infra-ordre des Caridea et l'ordre des décapodes incluent des espèces colonisant les trois environnement pris en compte dans cette étude dont, entre autres, la crevette *Rimicaris exoculata* et crabe *Cyanagraea praedator* qui colonisent tous les deux le pôle chaud de l'environnement hydrothermal profond. En effet les *Rimicaris* sont des crevettes présentant des épibioses nécessitant un contact régulier avec les émissions hydrothermales pour maintenir la chimiosynthèse bactérienne, et *Cyanagraea praedator* qui est le prédateur principal des deux espèces d'*Alvinella*. Ces espèces sont elles aussi soumises aux fortes amplitudes de la température rencontrée sur les cheminées hydrothermales, mais leur mobilité leur permet d'échapper à ces conditions lorsqu'elles sont extrêmes et de se déplacer entre cheminées au sein d'un champ hydrothermal et ainsi d'éviter les extinctions locales de sites. Ces modèles biologiques pourraient constituer une alternative pour atténuer l'impact des phénomènes démographiques sur l'architecture de la diversité génétique et nous permettre d'accéder à la signature laissée par la variabilité environnementale.

Bibliographie

Akashi, H. (1995). Inferring weak selection from patterns of polymorphism and divergence at" silent" sites in Drosophila DNA. *Genetics*, *139*(2), 1067-1076.

Allcock, A. L., & Strugnell, J. M. (2012). Southern Ocean diversity: new paradigms from molecular ecology. *Trends in Ecology & Evolution*, 27(9), 520-528.

Alt, J. C. (1995). Subseafloor Processes in Mid-Ocean Ridge Hydrothennal Systems. *Seafloor Hydrothermal Systems: Physical, Chemical, Biological, and Geological Interactions*, 85-114.

Andrew, R. L., & Rieseberg, L. H. (2013). Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*, 67(9), 2468-2482.

Antao, T., Lopes, A., Lopes, R. J., Beja-Pereira, A., & Luikart, G. (2008). LOSITAN: a workbench to detect molecular adaptation based on a F ST-outlier method. *BMC Bioinformatics*, *9*(1), 323.

Arnaud-Haond, S., Alberto, F., Teixeira, S., Procaccini, G., Serrao, E. A., & Duarte, C. M. (2005). Assessing genetic diversity in clonal organisms: low diversity or low resolution? Combining power and cost efficiency in selecting markers. *Journal of Heredity*, *96*(4), 434-440.

Arnold, F. H., Wintrode, P. L., Miyazaki, K., & Gershenson, A. (2001). How enzymes adapt: lessons from directed evolution. *Trends in Biochemical Sciences*, *26*(2), 100-106.

Aronson, R. B., Thatje, S., Clarke, A., Peck, L. S., Blake, D. B., Wilga, C. D., & Seibel, B. A. (2007). Climate change and invasibility of the Antarctic benthos. *Annual Review of Ecology, Evolution, and Systematics*, *38*, 129-154.

Asins M.J., Carbonell E.A. (1987) Concepts involved in measuring genetic variability and its importance in conservation of plant genetic resources. *Evolutionary Trends in Plants* 1:51-62

Autem, M., Salvidio, S., Pasteur, N., Desbruyères, D., & Laubier, L. (1985). Mise en évidence de l'isolement génétique des deux formes sympatriques d'Alvinella pompejana (Polychaeta: Ampharetidae), annélides inféodées aux sites hydrothermaux actifs de la dorsale du Pacifique oriental. *Comptes rendus de l'Académie des Sciences. Série 3, Sciences de la Vie, 301*(4), 131-135.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS One*, *3*(10), e3376.

Barker, P. F., & Thomas, E. (2004). Origin, signature and palaeoclimatic influence of the Antarctic Circumpolar Current. *Earth-Science Reviews*, *66*(1-2), 143-162.

Barnes, D. K., Fuentes, V., Clarke, A., Schloss, I. R., & Wallace, M. I. (2006). Spatial and temporal variation in shallow seawater temperatures around Antarctica. *Deep Sea Research Part II: Topical Studies in Oceanography*, *53*(8), 853-865.

Barnich, R., Fiege, D., Micaletto, G., & Gambi, M. C. (2006). Redescription of Harmothoe spinosa Kinberg, 1856 (Polychaeta: Polynoidae) and related species from Subantarctic and Antarctic waters, with the erection of a new genus. *Journal of Natural History*, *40*(1-2), 33-75.

Barrett, R. D., & Schluter, D. (2008). Adaptation from standing genetic variation. *Trends in Ecology & Evolution*, 23(1), 38-44.

Barton, N. H. (1979). Gene flow past a cline. Heredity, 43(3), 333.

Barton, N. H. (2000). Genetic hitchhiking. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *355*(1403), 1553-1562.

Bataillon, T. M., David, J. L., & Schoen, D. J. (1996). Neutral genetic markers and conservation genetics: simulated germplasm collections. *Genetics*, *144*(1), 409-417.

Bates, A. E., Tunnicliffe, V., & Lee, R. W. (2005). Role of thermal conditions in habitat selection by hydrothermal vent gastropods. *Marine Ecology Progress Series*, *305*, 1-15.

Bateson, W., & Mendel, G. (2013). Mendel's principles of heredity. Courier Corporation.

Beaumont, M. A., & Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society of London B: Biological*, 263(1377), 1619-1626.

Behrman, E.L., Watson, S.S., O'brien, K. R., Heschel, M. S., & Schmidt, P. S. (2015). Seasonal variation in life history traits in two *Drosophila* species. *Journal of Evolutionary Biology*, 28(9), 1691-1704.

Belkhir K., Borsa P., Chikhi L., Raufaste N. & Bonhomme F. 1996-2004 GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CNRS UMR 5171, Université de Montpellier II, Montpellier (France).

Bell, O., Tiwari, V. K., Thomä, N. H., & Schübeler, D. (2011). Determinants and dynamics of genome accessibility. *Nature Reviews Genetics*, *12*(8), 554.

Benzécri, J. P. (1973). L'analyse des données (Vol. 2, p. 1). Paris: Dunod.

Berezovsky, I. N., & Shakhnovich, E. I. (2005). Physics and evolution of thermophilic adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36), 12742-12747.

Bessey, C.E. (1880). The Supposed Dimorphism of *Lithospermum Longiflorum* (L. Angustifolium Michx. of Gray's Synoptical Flora.). *The American Naturalist*, 14(6), 417-421.

Bierne, N., Borsa, P., Daguin, C., Jollivet, D., Viard, F., Bonhomme, F., & David, P. (2003). Introgression patterns in the mosaic hybrid zone between Mytilus edulis and M. galloprovincialis. *Molecular Ecology*, *12*(2), 447-461.

Bierne, N., Welch, J., Loire, E., Bonhomme, F., & David, P. (2011). The coupling hypothesis: why genome scans may fail to map local adaptation genes. *Molecular Ecology*, *20*(10), 2044-2072.

Bierne, N., Gagnaire, P. A., & David, P. (2013). The geography of introgression in a patchy environment and the thorn in the side of ecological speciation. *Current Zoology*, *59*(1), 72-86.

Blatch, G. L., & Lässle, M. (1999). The tetratricopeptide repeat: a structural motif mediating proteinprotein interactions. *Bioessays*, 21(11), 932-939.

Bleich, V. C., Wehausen, J. D., & Holl, S. A. (1990). Desert-dwelling Mountain Sheep: Conservation Implications of a Naturally Fragmented Distribution. *Conservation Biology*, *4*(4), 383-390.

Bogdanova, V. S., Zaytseva, O. O., Mglinets, A. V., Shatskaya, N. V., Kosterin, O. E., & Vasiliev, G. V. (2015). Nuclear-cytoplasmic conflict in pea (Pisum sativum L.) is associated with nuclear and plastidic candidate genes encoding acetyl-CoA carboxylase subunits. *PloS One*, 10(3), e0119835.

Bradford, D. F., Neale, A. C., Nash, M. S., Sada, D. W., & Jaeger, J. R. (2003). Habitat patch occupancy by toads (Bufo punctatus) in a naturally fragmented desert landscape. *Ecology*, *84*(4), 1012-1023.

Brakefield, P. M. (1985). Differential winter mortality and seasonal selection in the polymorphic ladybird *Adalia bipunctata* (L) in the Netherlands. *Biological Journal of the Linnean Society*, 24(2), 189-206.

Brasier, M. J., Wiklund, H., Neal, L., Jeffreys, R., Linse, K., Ruhl, H., & Glover, A. G. (2016). DNA barcoding uncovers cryptic diversity in 50% of deep-sea Antarctic polychaetes. *Royal Society Open Science*, *3*(11), 160432.

Brasier M. J., Harle J. , Wiklund H., Jeffreys R. M. , Linse K., Henry A. Ruhl H. A. & Glover, A. G. (2017). Distributional patterns of polychaetes across the west Antarctic based on DNA barcoding and particle tracking analyses. *Frontiers in Marine Science*, 4, 356.

Brey, T., Dahm, C., Gorny, M., Klages, M., Stiller, M., & Arntz, W. E. (1996). Do Antarctic benthic invertebrates show an extended level of eurybathy?. *Antarctic Science*, 8(1), 3-6.

Broquet, T., Viard, F., & Yearsley, J. M. (2013). Genetic drift and collective dispersal can result in chaotic genetic patchiness. *Evolution*, *67*(6), 1660-1675.

Brown, A. C. (1960). Desiccation as a factor influencing the vertical distribution of some South African Gastropoda from intertidal Rochy Shores. *Portugal Acta biologica*. (B), Vol. 8, pp.11-23

Bukau, B., & Horwich, A. L. (1998). The Hsp70 and Hsp60 chaperone machines. Cell, 92(3), 351-366.

Caldwell, J. P. (1982). Disruptive selection: a tail color polymorphism in Acris tadpoles in response to differential predation. *Canadian Journal of Zoology*, *60*(11), 2818-2827.

Callaghan, D.T., Blanchfield, P. J., & Cott, P. A. (2016). Lake trout (*Salvelinus namaycush*) spawning habitat in a northern lake: the role of wind and physical characteristics on habitat quality. *Journal of Great Lakes Research*, 42(2), 299-307.

Cary, S. C., Cottrell, M. T., Stein, J. L., Camacho, F., & Desbruyeres, D. (1997). Molecular identification and localization of filamentous symbiotic bacteria associated with the hydrothermal vent annelid *Alvinella pompejana*. *Applied and Environmental Microbiology*, *63*(3), 1124-1130.

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, *22*(11), 3124-3140.

Chain, P. S., Comerci, D. J., Tolmasky, M. E., Larimer, F. W., Malfatti, S. A., Vergez, L. M., ... & Garcia, E. (2005). Whole-genome analyses of speciation events in pathogenic Brucellae. *Infection and Immunity*, 73(12), 8353-8361.

Chevaldonné, P., Desbruyères, D., & Le Haître, M. (1991). Time-series of temperature from three deepsea hydrothermal vent sites. *Deep Sea Research Part A. Oceanographic Research Papers*, *38*(11), 1417-1430.

Chevaldonné, P., Desbruyéres, D., & Childress, J. J. (1992). ... and some even hotter. *Nature*, *359*(6396), 593.

Chevaldonné, P., & Jollivet, D. (1993). Videoscopic study of deep-sea hydrothermal vent alvinellid polychaete populations: biomass estimation and behaviour. *Marine Ecology Progress Series*, 251-262.

Charlesworth, D., Charlesworth, B., & Morgan, M. T. (1995). The pattern of neutral molecular variation under the background selection model. *Genetics*, *141*(4), 1619-1632.

Chevaldonne, P., Jollivet, D., Vangriesheim, A., & Desbruyeres, D. (1997). Hydrothermal-vent alvinellid polychaete dispersal in the eastern Pacific. 1. Influence of vent site distribution, bottom currents, and biological patterns. *Limnology and Oceanography*, 42(1), 67-80.

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

Chevin, L. M., & Hoffmann, A. A. (2017). Evolution of phenotypic plasticity in extreme environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1723), 20160138.

Clark, M. S., & Peck, L. S. (2009). HSP70 heat shock proteins and environmental stress in Antarctic marine organisms: a mini-review. *Marine genomics*, 2(1), 11-18.

Clarke, A. (1998). Temperature and energetics: an introduction to cold ocean physiology. *Cold Ocean Physiology*, *66*, 3-32.

Clarke, A., Meredith, M. P., Wallace, M. I., Brandon, M. A., & Thomas, D. N. (2008). Seasonal and interannual variability in temperature, chlorophyll and macronutrients in northern Marguerite Bay, Antarctica. *Deep Sea Research Part II: Topical Studies in Oceanography*, *55*(18), 1988-2006.

Collinge, S. K. (1996). Ecological consequences of habitat fragmentation: implications for landscape architecture and planning. *Landscape and Urban Planning*, *36*(1), 59-77.

Cosham, J., Beazley, K. F., & McCarthy, C. (2016). Environmental factors influencing local distributions of European green crab (*Carcinus maenas*) for modeling and management applications. *Environmental Reviews*, 24(3), 244-252.

Cowen, R. K., Lwiza, K. M., Sponaugle, S., Paris, C. B., & Olson, D. B. (2000). Connectivity of marine populations: open or closed?. *Science*, *287*(5454), 857-859.

Cox, P. M., Betts, R. A., Jones, C. D., Spall, S. A., & Totterdell, I. J. (2000). Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature*, *408*(6809), 184-187.

Coyne, J. A., & Orr, H. A. (1989). Patterns of speciation in Drosophila. Evolution, 43(2), 362-381.

Crandall, K. A., & Hillis, D. M. (1997). Rhodopsin evolution in the dark. Nature, 387(6634), 667.

Crow, J. F. (1970). Genetic loads and the cost of natural selection. In *Mathematical topics in population genetics* (pp. 128-177). Springer, Berlin, Heidelberg.

Dahlhoff, E., O'Brien, J., Somero, G. N., & Vetter, R. D. (1991). Temperature effects on mitochondria from hydrothermal vent invertebrates: evidence for adaptation to elevated and variable habitat temperatures. *Physiological Zoology*, *64*(6), 1490-1508.

Dai, A. (2013). Increasing drought under global warming in observations and models. *Nature Climate Change*, *3*(1), 52-58.

Dalongeville, A., Andrello, M., Mouillot, D., Albouy, C., & Manel, S. (2016). Ecological traits shape genetic diversity patterns across the Mediterranean Sea: a quantitative review on fishes. *Journal of Biogeography*, 43(4), 845-857.

Daly, J. M. (1972). The maturation and breeding biology of Harmothoe imbricata (Polychaeta: Polynoidae). *Marine Biology*, *12*(1), 53-66.

Darwin, C. (1859) On the Origin of Species by Means of Natural Selection. London: John Murray.

David, P., & Jarne, P. (1997). Context-dependent survival differences among electrophoretic genotypes in natural populations of the marine bivalve Spisula ovalis. *Genetics*, *146*(1), 335-344.

de Jesus Ferreira, M. C., Bao, X., Laizé, V., & Hohmann, S. (2001). Transposon mutagenesis reveals novel loci affecting tolerance to salt stress and growth at low temperature. *Current Genetics*, 40(1), 27-39.

Desbruyeres, D., & Laubier, L. (1980). Alvinella pompejana gen. sp. nov., Ampharetidae aberrant des sources hydrothermales de la ride Est-Pacifique. *Oceanologica Acta*, *3*(3), 267-274.

Desbruyères, D., & Laubier, L. (1986). Les Alvinellidae, une famille nouvelle d'annélides polychètes inféodées aux sources hydrothermales sous-marines: systématique, biologie et écologie. *Canadian Journal of Zoology*, 64(10), 2227-2245.

Desbruyères, D., Chevaldonné, P., Alayse, A. M., Jollivet, D., Lallier, F. H., Jouin-Toulmond, C., ... & Arndt, C. (1998). Biology and ecology of the "Pompeii worm" (Alvinella pompejana Desbruyères and Laubier), a normal dweller of an extreme deep-sea environment: a synthesis of current knowledge and recent developments. *Deep Sea Research Part II: Topical Studies in Oceanography*, 45(1-3), 383-422.

Desbruyères, D., Segonzac, M., & Bright, M. (Eds.). (2006). Handbook of deep-sea hydrothermal vent fauna.

Dierssen, H. M., Smith, R. C., & Vernet, M. (2002). Glacial meltwater dynamics in coastal waters west of the Antarctic peninsula. *Proceedings of the National Academy of Sciences*, *99*(4), 1790-1795.

Dobzhansky, T. (1937). *Genetics and the Origin of Species* (Third Edition 1951). Columbia University Press.

Dobzhansky, T., & Pavan, C. (1950). Local and seasonal variations in relative frequencies of species of Drosophila in Brazil. *The Journal of Animal Ecology*, 1-14.

Dong, Y. and Somero, G. N. (2009). Temperature adaptation of cytosolic malate dehydrogenases of limpets (genus *Lottia*): differences in stability and function due to minor changes in sequence correlate with biogeographic and vertical distributions. *Journal of Experimental Biology*. 212, 169-177.

Dreyer, J. C., Knick, K. E., Flickinger, W. B., & Van Dover, C. L. (2005). Development of macrofaunal community structure in mussel beds on the northern East Pacific Rise. *Marine Ecology Progress Series*, *302*, 121-134.

Eakins, B. W., & Lonsdale, P. F. (2003). Structural patterns and tectonic history of the Bauer microplate, Eastern Tropical Pacific. *Marine Geophysical Researches*, *24*(3-4), 171-205.

Egan, S. P., Nosil, P., & Funk, D. J. (2008). Selection and genomic differentiation during ecological speciation: isolating the contributions of host association via a comparative genome scan of Neochlamisus bebbianae leaf beetles. *Evolution*, 62(5), 1162-1181.

Eldon, B., Riquet, F., Yearsley, J., Jollivet, D., & Broquet, T. (2016). Current hypotheses to explain genetic chaos under the sea. *Current Zoology*, 62(6), 551-566.

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

Ellingson, R. A., & Krug, P. J. (2016). Reduced genetic diversity and increased reproductive isolation follow population-level loss of larval dispersal in a marine gastropod. *Evolution*, *70*(1), 18-37.

Faith, D. P., Minchin, P. R., & Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, *69*(1-3), 57-68.

Faraway, J. J. (2002). Practical regression and ANOVA using R.

Faure, B., Chevaldonné, P., Pradillon, F., Thiébaut, E., & Jollivet, D. (2007). Spatial and temporal dynamics of reproduction and settlement in the Pompeii worm Alvinella pompejana (Polychaeta: Alvinellidae). *Marine Ecology Progress Series*, *348*, 197-211.

Ferronnière, G. (1901). Études biologiques sur les zones supra-littorale de la Loire-Inférieure (Vol. 11). A. Dugas. Field, K. G., Olsen, G. J., Lane, D. J., Giovannoni, S. J., Ghiselin, M. T., Raff, E. C., ... & Raff, R. A. (1988). Molecular phylogeny of the animal kingdom. *Science*, *239*(4841), 748-753.

Fields, P. A., & Somero, G. N. (1998). Hot spots in cold adaptation: localized increases in conformational flexibility in lactate dehydrogenase A4 orthologs of Antarctic notothenioid fishes. *Proceedings of the National Academy of Sciences*, 95(19), 11476-11481.

Fields, P. A., Rudomin, E. L., & Somero, G. N. (2006). Temperature sensitivities of cytosolic malate dehydrogenases from native and invasive species of marine mussels (genus Mytilus): sequence-function linkages and correlations with biogeographic distribution. *Journal of Experimental Biology*, *209*(4), 656-667.

Fields, P. A., Dong, Y., Meng, X., & Somero, G. N. (2015). Adaptations of protein structure and function to temperature: there is more than one way to 'skin a cat'. *Journal of Experimental Biology*, *218*(12), 1801-1811.

Fisher, C. R., & Childress, J. J. (1992). Organic carbon transfer from methanotrophic symbionts to the host hydrocarbon-seep mussel. *Symbiosis*, *12*(3), 221-235.

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, *180*(2), 977-993.

Fontanillas, E., Galzitskaya, O. V., Lecompte, O., Lobanov, M. Y., Tanguy, A., Mary, J., ... & Jollivet, D. (2017). Proteome evolution of deep-sea hydrothermal vent alvinellid polychaetes supports the ancestry of thermophily and subsequent adaptation to cold in some lineages. *Genome Biology and Evolution*, 9(2), 279-296.

Fornari, D.J. & Embley, R.W., 1995. In: Seafloor hydrothermal systems (eds. SE Humphris, RA Zierenberg, LS Mullineaux & RE Thomson). Washington DC: Geophysical Monograph, 91: 1-46.

Fouquet, Y., Auzende, J. M., Ballu, V., Batiza, R., Bideau, D., Cormier, M. H., ... & Spadea, P. (1994). Variabilité des manifestations hydrothermales actuelles le long d'une dorsale ultra rapide. Dorsale Est Pacifique entre 17° et 19° S (campagne NAUDUR). *Comptes Rendus de l'Academie des Sciences Serie II*, *319*(11), 1399-1406.

Fraïsse, C., Roux, C., Welch, J. J., & Bierne, N. (2014). Gene-flow in a mosaic hybrid zone: is local introgression adaptive?. *Genetics*, *197*(3), 939-951.

Francheteau, J., & Ballard, R. D. (1983). The East Pacific Rise near 21 N, 13 N and 20 S: Inferences for along-strike variability of axial processes of the mid-ocean ridge. *Earth and Planetary Science Letters*, *64*(1), 93-116.

Francis, R., Maine, E., & Schedl, T. (1995). Analysis of the multiple roles of gld-1 in germline development: interactions with the sex determination cascade and the glp-1 signaling pathway. *Genetics*, 139(2), 607-630.

Frankham, R. (1995). Effective population size/adult population size ratios in wildlife: a review. *Genetics Research*, 66(2), 95-107.

Fraser, C. I., Nikula, R., Ruzzante, D. E., & Waters, J. M. (2012). Poleward bound: biological impacts of Southern Hemisphere glaciation. *Trends in Ecology & Evolution*, *27*(8), 462-471.

Frichot, E. (2012). A short manual for LFMM.

Fu, Y. X., & Li, W. H. (1993). Statistical tests of neutrality of mutations. Genetics, 133(3), 693-709.

Gaill, F., Desbruyeres, D., & Prieur, D. (1987). Bacterial communities associated with "Pompei worms" from the East Pacific Rise hydrothermal vents: SEM, TEM observations. *Microbial Ecology*, *13*(2), 129-139.

Gaill, F., Mann, K., Wiedemann, H., Engel, J., & Timpl, R. (1995). Structural comparison of cuticle and interstitial collagens from annelids living in shallow sea-water and at deep-sea hydrothermal vents. *Journal of Molecular Biology*, 246(2), 284-294.

Gambi, M. C., Patti, F. P., Micaletto, G., & Giangrande, A. (2002). Diversity of reproductive features in some Antarctic polynoid and sabellid polychaetes, with a description of Demonax polarsterni sp. n.(Polychaeta, Sabellidae). In *Ecological Studies in the Antarctic Sea Ice Zone* (pp. 24-32). Springer, Berlin, Heidelberg.

Gayral, P., Melo-Ferreira, J., Glémin, S., Bierne, N., Carneiro, M., Nabholz, B., ... & Belkhir, K. (2013). Reference-free population genomics from next-generation transcriptome data and the vertebrate– invertebrate gap. *PLoS Genetics*, *9*(4), e1003457.

Gemmell, N. J., & Slate, J. (2006). Heterozygote advantage for fecundity. PLoS One, 1(1), e125.

Gillespie, J. H. (1976). A general model to account for enzyme variation in natural populations. II. Characterization of the fitness functions. *The American Naturalist*, *110*(975), 809-821.

Gillespie, J. H. (1977). A general model to account for enzyme variation in natural populations. III. Multiple alleles. *Evolution*, 31(1), 85-90.

Gillespie, J. H. (1985). The interaction of genetic drift and mutation with selection in a fluctuating environment. *Theoretical Population Biology*, 27(2), 222-237.

Gilpin, M. (1991). The genetic effective size of a metapopulation. In *Metapopulation Dynamics: Empirical and Theoretical Investigations* (pp. 165-175).

Giraud, T., Fortini, D., Levis, C., Leroux, P., & Brygoo, Y. (1997). RFLP markers show genetic recombination in Botryotinia fuckeliana (Botrytis cinerea) and transposable elements reveal two sympatric species. *Molecular Biology and Evolution*, *14*(11), 1177-1185.

Gohli, J., Lifjeld, J. T., & Albrecht, T. (2016). Migration distance is positively associated with sexlinked genetic diversity in passerine birds. *Ethology Ecology & Evolution*, 28(1), 42-52.

Gómez-Palacio, A., & Triana, O. (2014). Molecular evidence of demographic expansion of the Chagas disease vector Triatoma dimidiata (*Hemiptera, Reduviidae, Triatominae*) in Colombia. *PLoS Neglected Tropical Diseases*, 8(3), e2734.

Gonzalez-Wevar, C. A., David, B., & Poulin, E. (2011). Phylogeography and demographic inference in Nacella (Patinigera) concinna (Strebel, 1908) in the western Antarctic Peninsula. *Deep Sea Research Part II: Topical Studies in Oceanography*, 58(1-2), 220-229.

Goodall-Copestake, W. P., Perez-Espona, S., Clark, M. S., Murphy, E. J., Seear, P. J., & Tarling, G. A. (2010). Swarms of diversity at the gene cox1 in Antarctic krill. *Heredity*, *104*(5), 513.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., & Blinov, V. M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Research*, *17*(12), 4713-4730.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Chen, Z. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644.

Graur, D. (1991). Fundamentals of molecular evolution. Massachusetts: Sinauer Associates.

Hague, M. T., & Routman, E. J. (2016). Does population size affect genetic diversity? A test with sympatric lizard species. *Heredity*, *116*(1), 92-98.

Hamilton, W. D., & May, R. M. (1977). Dispersal in stable habitats. Nature, 269(5629), 578.

Hamrick, J. L., Linhart, Y. B., & Mitton, J. B. (1979). Relationships between life history characteristics and electrophoretically detectable genetic variation in plants. *Annual Review of Ecology and Systematics*, *10*(1), 173-200.

Hannington, M. D., Jonasson, I. R., Herzig, P. M., & Petersen, S. (1995). Physical and chemical processes of seafloor mineralization at mid-ocean ridges. *Seafloor hydrothermal systems: Physical, chemical, biological, and geological interactions*, 115-157.

Hansen, J., Sato, M., Ruedy, R., Lacis, A., & Oinas, V. (2000). Global warming in the twenty-first century: An alternative scenario. *Proceedings of the National Academy of Sciences*, 97(18), 9875-9880.

Harnly, M. H. (1936). The temperature-effective periods and the growth curves for length and area of the vestigial wings of *Drosophila melanogaster*. *Genetics*, 21(2), 84.

Harris, H. (1966). C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London B: Biological*, 164(995), 298-310.

Harrison, S., & Hastings, A. (1996). Genetic and evolutionary consequences of metapopulation structure. *Trends in Ecology & Evolution*, *11*(4), 180-183.

Hartfield, M., Bataillon, T., & Glémin, S. (2017). The Evolutionary Interplay between Adaptation and Self-Fertilization. *Trends in Genetics*, 33(6), 420-431.

Hawkesworth, C., Kelley, S., Turner, S., Le Roex, A., & Storey, B. (1999). Mantle processes during Gondwana break-up and dispersal. *Journal of African Earth Sciences*, 28(1), 239-261.

Haymon, R. M., & Macdonald, K. C. (1985). The Geology of Dee-Sea Hot Springs. *American Scientist*, 73, 441-449.

Hê S, Josse J, Husson F. (2008). FactoMineR: an R package for multivariate analysis. Journal of Statistical Software. 25:1–18.

Hedgecock, D. (1994). Does variance in reproductive success limit effective population sizes of marine organisms. *Genetics and Evolution of Aquatic Organisms*, 122.

Hedgecock, D., & Pudovkin, A. I. (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. *Bulletin of Marine Science*, 87(4), 971-1002.

Hedrick, P. W. (1985). Inbreeding and selection in natural populations. In *Population Genetics in Forestry* (pp. 71-91). Springer, Berlin, Heidelberg.

Hedrick, P. W. (1986). Genetic polymorphism in heterogeneous environments: a decade later. *Annual Review of Ecology and Systematics*, 17(1), 535-566.

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics*, *117*(2), 331-341.

Hedrick P. W. (1999). Genetics of populations. Second Edition. Jones and Bartlett Publishers.

Hedrick, P. W. (2012). What is the evidence for heterozygote advantage selection?. *Trends in Ecology* & *Evolution*, 27(12), 698-704.

Hegarty, M. J., & Hiscock, S. J. (2005). Hybrid speciation in plants: new insights from molecular studies. *New Phytologist*, *165*(2), 411-423.

Helmuth, B., Broitman, B. R., Blanchette, C. A., Gilman, S., Halpin, P., Harley, C. D., ... & Strickland,D. (2006). Mosaic patterns of thermal stress in the rocky intertidal zone: implications for climate change.*Ecological Monographs*, 76(4), 461-479.

Hemery, L. G., Eléaume, M., Roussel, V., Améziane, N., Gallut, C., Steinke, D., ... & Wilson, N. G. (2012). Comprehensive sampling reveals circumpolarity and sympatry in seven mitochondrial lineages of the Southern Ocean crinoid species Promachocrinus kerguelensis (Echinodermata). *Molecular Ecology*, *21*(10), 2502-2518.

Hewitt, G. M. (1999). Post-glacial re-colonization of European biota. *Biological Journal of the Linnean Society*, 68(1-2), 87-112.

Hewitt, G. M. (2004). Genetic consequences of climatic oscillations in the Quaternary. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1442), 183-195.

Hijmans, R. J., & Graham, C. H. (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*, *12*(12), 2272-2281.

Hilbish, T. J. (1996). Population genetics of marine species: the interaction of natural selection and historically differentiated populations. *Journal of Experimental Marine Biology and Ecology*, 200(1-2), 67-83.

Hillman, T. L., Keenlance, P., Moore, J. A., Swanson, B. J., Jacquot, J. J., Witt, J. C., & Cornman, A. (2017). Genetic diversity of reintroduced American martens in Michigan's Lower Peninsula. *Journal of Mammalogy*, *98*(5), 1489-1496.

Hofmann, G. E. (1999). Ecologically relevant variation in induction and function of heat shock proteins in marine organisms. *American Zoologist*, *39*(6), 889-900.

Houghton JT, Ding Y, Griggs DJ, Noguer M, Van der Linden PJ, Dai X, Maskell K, Johnson CA, editors. Climate change 2001: the scientific basis. Contribution of Working Group I to the third assessment report of the Intergovernmental Panel of Climate Change.

Hourdez, S., & Weber, R. E. (2005). Molecular and functional adaptations in deep-sea hemoglobins. *Journal of Inorganic Biochemistry*, *99*(1), 130-141.

Hudson, R. R., Kreitman, M., & Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, *116*(1), 153-159.

Hurtado, L. A., Lutz, R. A., & Vrijenhoek, R. C. (2004). Distinct patterns of genetic differentiation among annelids of eastern Pacific hydrothermal vents. *Molecular Ecology*, *13*(9), 2603-2615.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise de Sciences Naturelles*, *37*, 547-579.

Jacobs, R. P. W. M. (1979). Distribution and aspects of the production and biomass of eelgrass, Zostera marina L., at Roscoff, France. *Aquatic Botany*, *7*, 151-172.

Jang, S. J., Park, E., Lee, W. K., Johnson, S. B., Vrijenhoek, R. C., & Won, Y. J. (2016). Population subdivision of hydrothermal vent polychaete Alvinella pompejana across equatorial and Easter Microplate boundaries. *BMC Evolutionary Biology*, *16*(1), 235.

Jangjoo, M., Matter, S. F., Roland, J., & Keyghobadi, N. (2016). Connectivity rescues genetic diversity after a demographic bottleneck in a butterfly population network. *Proceedings of the National Academy of Sciences*, *113*(39), 10914-10919.

Jenkins, A., Dutrieux, P., Jacobs, S., Steig, E. J., Gudmundsson, G. H., Smith, J., & Heywood, K. J. (2016). Decadal ocean forcing and Antarctic ice sheet response: Lessons from the Amundsen Sea. *Oceanography*, 29(4), 106-117.

Jiang, Q., Teufel, A. I., Jackson, E. L., & Wilke, C. O. (2018). Beyond Thermodynamic Constraints: Evolutionary Sampling Generates Realistic Protein Sequence Variation. *Genetics*, 208(4), 1387-1395.

Jirkov IA, Leontovitch MK. (2013) Identification keys for Terebellomorpha (Polychaeta) of the eastern Atlantic and the North Polar basin. Invertebrate Zoology [aka Zoologiya Bespozvonochnykh]. 10(2):217–43.

Johansson, M., Primmer, C. R., & Merilae, J. (2007). Does habitat fragmentation reduce fitness and adaptability? A case study of the common frog (Rana temporaria). *Molecular Ecology*, *16*(13), 2693-2700.

Johnson, K. S., Childress, J. J., & Beehler, C. L. (1988). Short-term temperature variability in the Rose Garden hydrothermal vent field: an unstable deep-sea environment. *Deep Sea Research Part A*. *Oceanographic Research Papers*, *35*(10-11), 1711-1721.

Johnson, S. B., Won, Y. J., Harvey, J. B., & Vrijenhoek, R. C. (2013). A hybrid zone between Bathymodiolus mussel lineages from eastern Pacific hydrothermal vents. *BMC Evolutionary Biology*, *13*(1), 21.

Jokinen, H., Wennhage, H., Ollus, V., Aro, E., & Norkko, A. (2016). Juvenile flatfish in the northern Baltic Sea—long-term decline and potential links to habitat characteristics. *Journal of Sea Research*, *107*, 67-75.

Jollivet, D., Desbruyères, D., Ladrat, C., & Laubier, L. (1995). Evidence for differences in the allozyme thermostability of deep-sea hydrothermal vent polychaetes (Alvinellidae): a possible selection by habitat. *Marine Ecology Progress Series*, 125-136.

Jollivet, D., Dixon, L. R. J., Desbruyères, D., & Dixon, D. R. (1998). Ribosomal (rDNA) variation in a deep sea hydrothermal vent polychaete, Alvinella pompejana, from 13 N on the East Pacific Rise. *Journal of the Marine Biological Association of the United Kingdom*, 78(1), 113-130.

Jollivet, D., Chevaldonne, P., & Planque, B. (1999). Hydrothermal-vent alvinellid polychaete dispersal in the eastern pacific. 2. A metapopulation model based on habitat shifts. *Evolution*, *53*(4), 1128-1142.

Jollivet, D., Empis, A., Baker, M. C., Hourdez, S., Comtet, T., Jouin-Toulmond, C., ... & Tyler, P. A. (2000). Reproductive biology, sexual dimorphism, and population structure of the deep sea hydrothermal vent scale-worm, Branchipolynoe seepensis (Polychaeta: Polynoidae). *Journal of the Marine Biological Association of the United Kingdom*, 80(1), 55-68.

Jouin-Toulmond, C., Mozzo, M., & Hourdez, S. (2002). Ultrastructure of spermatozoa in four species of Alvinellidae (Annelida: Polychaeta). *Cahiers de Biologie Marine*, *43*(3/4), 391-394.

Juan, C., Guzik, M. T., Jaume, D., & Cooper, S. J. (2010). Evolution in caves: Darwin's 'wrecks of ancient life'in the molecular era. *Molecular Ecology*, *19*(18), 3865-3880.

Kimura, M., & Crow, J. F. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, *49*(4), 725-738.

Kimura, M., & Weiss, G. H. (1964). The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, *49*(4), 561.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217(5129), 624-626.

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

Kis-Papo, T., Kirzhner, V., Wasser, S. P., & Nevo, E. (2003). Evolution of genomic diversity and sex at extreme environments: fungal life under hypersaline Dead Sea stress. *Proceedings of the National Academy of Sciences*, *100*(25), 14970-14975.

Kondrashov, F. A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society of London B: Biological*, 279(1749), 5048-5057.

Kremer, A., Ronce, O., Robledo-Arnuncio, J. J., Guillaume, F., Bohrer, G., Nathan, R., ... & Kuparinen, A. (2012). Long-distance gene flow and adaptation of forest trees to rapid climate change. *Ecology Letters*, *15*(4), 378-392.

Kröger, K., & Rowden, A. A. (2008). Polychaete assemblages of the northwestern Ross Sea shelf: worming out the environmental drivers of Antarctic macrobenthic assemblage composition. *Polar Biology*, *31*(8), 971-989.

Lacassagne, A. (1940). Sexual dimorphism of the submaxillary gland in mice. *Compte Rendu des Séances de la Sociéte de Biologie*, 133, 180-181.

Lalou, C. (1991). Deep-sea hydrothermal venting: A recently discovered marine system. *Journal of Marine* Systems 1, 403-440.

Lalou, C., Reyss, J.E, Arnold, M., Thompson, G., Fouquet, Y. & Rona, P.A. (1993). New age dating for MidAtlantic Ridge hydrothermal sites: TAG and Snake Pit chronology revisited. *Journal of Geophisical Research*, 98(B3), 9705-9713.

Lande, R., & Orzack, S. H. (1988). Extinction dynamics of age-structured populations in a fluctuating environment. *Proceedings of the National Academy of Sciences*, 85(19), 7418-7421.

Landi, P., Hui, C., & Dieckmann, U. (2015). Fisheries-induced disruptive selection. *Journal of Theoretical Biology*, *365*, 204-216.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), 357-359.

Latch, E. K., Dharmarajan, G., Glaubitz, J. C., & Rhodes, O. E. (2006). Relative performance of Bayesian clustering software for inferring population substructure and individual assignment at low levels of population differentiation. *Conservation Genetics*, 7(2), 295-302.

Lavie, B., & Nevo, E. (1982). Heavy metal selection of phosphoglucose isomerase allozymes in marine gastropods. *Marine Biology*, *71*(1), 17-22.

Le Bris, N., Sarradin, P. M., & Caprais, J. C. (2003). Contrasted sulphide chemistries in the environment of 13 N EPR vent fauna. *Deep Sea Research Part I: Oceanographic Research Papers*, *50*(6), 737-747.

Le Bris, N., Govenar, B., Le Gall, C., & Fisher, C. R. (2006). Variability of physico-chemical conditions in 9 50' N EPR diffuse flow vent habitats. *Marine Chemistry*, 98(2-4), 167-182.

Le Bris, N., & Gaill, F. (2007). How does the annelid Alvinella pompejana deal with an extreme hydrothermal environment?. *Reviews in Environmental Science and Bio/Technology*, 6(1-3), 197.

Le Moan, A., Gagnaire, P. A., & Bonhomme, F. (2016). Parallel genetic divergence among coastalmarine ecotype pairs of European anchovy explained by differential introgression after secondary contact. *Molecular Ecology*, *25*(13), 3187-3202.

Lefebvre V, Donnadieu Y (2012) Deciphering the role of southern gateways and carbon dioxide on the onset of the Antarctic Circumpolar Current. *Palaeoceanography*, 27, 1–9.

Levene, H. (1953). Genetic equilibrium when more than one ecological niche is available. *The American Naturalist*, 87(836), 331-333.

Levin, S. A., Cohen, D., & Hastings, A. (1984). Dispersal strategies in patchy environments. *Theoretical Population Biology*, 26(2), 165-191.

Lewontin, R. C., & Hubby, J. L. (1966). A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics*, *54*(2), 595-609.

Lewontin, R. C. (1974). *The genetic basis of evolutionary change* (Vol. 560). New York: *Columbia University Press.*

Lewontin, R. C. (1991). Twenty-five years ago in Genetics: electrophoresis in the development of evolutionary genetics: milestone or millstone?. *Genetics*, *128*(4), 657-662.

Librado, P., & Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25(11), 1451-1452.

Lisiecki, L. E., & Raymo, M. E. (2005). A Pliocene-Pleistocene stack of 57 globally distributed benthic δ18O records. *Paleoceanography*, 20(1).

Lonsdale, P. (1977). Clustering of suspension-feeding macrobenthos near abyssal hydrothermal vents at oceanic spreading centers. *Deep Sea Research*, *24*(9), 857-863.

Loyau, T., Bedrani, L., Berri, C., Métayer-Coustard, S., Praud, C., Coustham, V., & Hennequet-Antier, C. (2015). Cyclic variations in incubation conditions induce adaptive responses to later heat exposure in chickens: a review. *Animal*, *9*(1), 76-85.

Luther, G. W., Rozan, T. F., Taillefert, M., Nuzzio, D. B., Di Meo, C., Shank, T. M., et al. (2001). Chemical speciation drives hydrothermal vent ecology. *Nature* 410, 813–816.

Lutz, F. E. (1913). Experiments concerning the sexual difference in the wing length of *Drosophila ampelophila*. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology*, 14(2), 267-273.

Lutz, R. A., Jablonski, D., & Turner, R. D. (1984). Larval development and dispersal at deep-sea hydrothermal vents. *Science*, 226(4681), 1451-1454.

Lutz, R. A. (1988). Dispersal of organisms at deep-sea hydrothermal vents: a review. *Oceanologica Acta, Special Issue*.

Maggs, C. A., Castilho, R., Foltz, D., Henzler, C., Jolly, M. T., Kelly, J., ... & Viard, F. (2008). Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. *Ecology*, 89(sp11).

Malcolm, J. R., Markham, A., Neilson, R. P., & Garaci, M. (2002). Estimated migration rates under scenarios of global climate change. *Journal of Biogeography*, *29*(7), 835-849.

Malfant, M., Coudret, J., Le Merdy, R., & Viard, F. (2017). Effects of temperature and salinity on juveniles of two ascidians, one native and one invasive, and their hybrids. *Journal of Experimental Marine Biology and Ecology*, 497, 180-187.

Mammerickx, J., & Klitgord, K. D. (1982). Northern East Pacific Rise: Evolution from 25 my BP to the present. *Journal of Geophysical Research: Solid Earth*, 87(B8), 6751-6759.

Marchinko, K. B., Matthews, B., Arnegard, M. E., Rogers, S. M., & Schluter, D. (2014). Maintenance of a genetic polymorphism with disruptive natural selection in stickleback. *Current Biology*, 24(11), 1289-1292.

Marteinsson, V. T., Birrien, J. L., Reysenbach, A. L., Vernet, M., Marie, D., Gambacorta, A., ... & Prieur, D. (1999). Thermococcus barophilus sp. nov., a new barophilic and hyperthermophilic archaeon isolated under high hydrostatic pressure from a deep-sea hydrothermal vent. *International Journal of Systematic and Evolutionary Microbiology*, *49*(2), 351-359.

Matabos, M., Thiébaut, E., Le Guen, D., Sadosky, F., Jollivet, D., & Bonhomme, F. (2008). Geographic clines and stepping-stone patterns detected along the East Pacific Rise in the vetigastropod Lepetodrilus elevatus reflect species crypticism. *Marine Biology*, *153*(4), 545-563.

Matabos, M., Plouviez, S., Hourdez, S., Desbruyeres, D., Legendre, P., Warén, A., ... & Thiébaut, E. (2011). Faunal changes and geographic crypticism indicate the occurrence of a biogeographic transition zone along the southern East Pacific Rise. *Journal of Biogeography*, *38*(3), 575-594.

McCauley, D. E. (1991). Genetic consequences of local population extinction and recolonization. *Trends in Ecology & Evolution*, 6(1), 5-8.

McDonald, J. H., & Kreitman, M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature*, *351*(6328), 652.

McHugh, D. (1993). A comparative study of reproduction and development in the polychaete family Terebellidae. *The Biological Bulletin*, *185*(2), 153-167.

McHugh, D., & Fong, P. P. (2002). Do life history traits account for diversity of polychaete annelids?. *Invertebrate Biology*, *121*(4), 325-338.

McPeek, M. A., & Holt, R. D. (1992). The evolution of dispersal in spatially and temporally varying environments. *The American Naturalist*, *140*(6), 1010-1027.

Meredith, M. P., & King, J. C. (2005). Rapid climate change in the ocean west of the Antarctic Peninsula during the second half of the 20th century. *Geophysical Research Letters*, *32*(19).

Miller, D. J., Macek, M. B., & Shur, B. D. (1992). Complementarity between sperm surface β -l, 4-galactosyl-transferase and egg-coat ZP3 mediates sperm–egg binding. *Nature*, 357(6379), 589-593.

Mirouze, M., & Paszkowski, J. (2011). Epigenetic contribution to stress adaptation in plants. *Current Opinion in Plant Biology*, *14*(3), 267-274.

Mitton, J. B., & Koehn, R. K. (1975). Genetic organization and adaptive response of allozymes to ecological variables in Fundulus heteroclitus. *Genetics*, 79(1), 97-111.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., & Rafalski, A. (2005). Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nature Genetics*, *37*(9), 997.

Mullineaux, L. S., Peterson, C. H., Micheli, F., & Mills, S. W. (2003). Successional mechanism varies along a gradient in hydrothermal fluid flux at deep-sea vents. *Ecological Monographs*, *73*(4), 523-542.

Muths, D., Davoult, D., Jolly, M. T., Gentil, F., & Jollivet, D. (2010). Pre-zygotic factors best explain reproductive isolation between the hybridizing species of brittle-stars Acrocnida brachiata and A. spatulispina (Echinodermata: Ophiuroidea). *Genetica*, *138*(6), 667-679.

Nagylaki, T. (1976). Clines with variable migration. Genetics, 83(4), 867-886.

Navarro, A., & Barton, N. H. (2003). Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. *Science*, *300*(5617), 321-324.

Neal, L., Linse, K., Brasier, M. J., Sherlock, E., & Glover, A. G. (2017). Comparative marine biodiversity and depth zonation in the Southern Ocean: evidence from a new large polychaete dataset from Scotia and Amundsen seas. *Marine Biodiversity*, 1-21.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12), 3321-3323.

Nei, M., (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.

Nevo, E., Ben-Shlomo, R., & Lavie, B. (1984). Mercury selection of allozymes in marine organisms: prediction and verification in nature. *Proceedings of the National Academy of Sciences*, 81(4), 1258-1259.

Nevo, E. (2001). Evolution of genome-phenome diversity under environmental stress. *Proceedings of the National Academy of Sciences*, 98(11), 6233-6240.

Nicholson, S. A., Lévy, M., Llort, J., Swart, S., & Monteiro, P. (2016). Investigation into the impact of storms on sustaining summer primary productivity in the Sub-Antarctic Ocean. *Geophysical Research Letters*, *43*(17), 9192-9199.

Nosil, P., Funk, D. J., & ORTIZ-BARRIENTOS, D. A. N. I. E. L. (2009). Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, *18*(3), 375-402.

Nosil, P., Harmon, L. J., & Seehausen, O. (2009). Ecological explanations for (incomplete) speciation. *Trends in Ecology & Evolution*, *24*(3), 145-156.

Nunney, L., & Campbell, K. A. (1993). Assessing minimum viable population size: demography meets population genetics. *Trends in Ecology & Evolution*, 8(7), 234-239.

Ohta, T. (2002). Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences*, *99*(25), 16134-16137.

Olendorf, R., Rodd, F. H., Punzalan, D., Houde, A. E., Hurt, C., Reznick, D. N., & Hughes, K. A. (2006). Frequency-dependent survival in natural guppy populations. *Nature*, *441*(7093), 633.

Olsen, J. B., Kinziger, A. P., Wenburg, J. K., Lewis, C. J., Phillips, C. T., & Ostrand, K. G. (2016). Genetic diversity and divergence in the fountain darter (*Etheostoma fonticola*): implications for conservation of an endangered species. *Conservation Genetics*, *17*(6), 1393-1404.

Pabis, K., & Sicinski, J. (2010). Distribution and diversity of polychaetes collected by trawling in Admiralty Bay: an Antarctic glacial fiord. *Polar Biology*, *33*(2), 141-151.

Papaleo, E., Riccardi, L., Villa, C., Fantucci, P., & De Gioia, L. (2006). Flexibility and enzymatic coldadaptation: a comparative molecular dynamics investigation of the elastase family. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1764*(8), 1397-1406.

Papot, C., Cascella, K., Toullec, J. Y., & Jollivet, D. (2016). Divergent ecological histories of two sister Antarctic krill species led to contrasted patterns of genetic diversity in their heat-shock protein (hsp70) arsenal. *Ecology and Evolution*, 6(5), 1555-1575.

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. *Nature*, *421*(6918), 37.

Pauls, S. U., Nowak, C., Bálint, M., & Pfenninger, M. (2013). The impact of global climate change on genetic diversity within populations and species. *Molecular Ecology*, 22(4), 925-946.

Payseur, B. A. (2010). Using differential introgression in hybrid zones to identify genomic regions involved in speciation. *Molecular Ecology Resources*, *10*(5), 806-820.

Pearson, R. G. (2006). Climate change and the migration capacity of species. *Trends in Ecology & Evolution*, 21(3), 111-113.

Peck, L. S. (2002). Ecophysiology of Antarctic marine ectotherms: limits to life. In *Ecological Studies in the Antarctic Sea Ice Zone* (pp. 221-230). Springer, Berlin, Heidelberg.
Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS One*, *7*(5), e37135.

Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J. M., Basile, I., ... & Delmotte, M. (1999). Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, *399*(6735), 429-436.

Petit, R. J., Aguinagalde, I., de Beaulieu, J. L., Bittkau, C., Brewer, S., Cheddadi, R., ... & Mohanty, A. (2003). Glacial refugia: hotspots but not melting pots of genetic diversity. *Science*, *300*(5625), 1563-1565.

Piccino, P., Viard, F., Sarradin, P., Le Bris, N., Le Guen, D., & Jollivet, D. (2004). Thermal selection of PGM allozymes in newly founded populations of the thermotolerant vent polychaete *Alvinella pompejana*. *Proceedings of the Royal Society of London B: Biological Sciences*, 271(1555), 2351-2359.

Picton, H., & Linder, S. E. (1897). LVII.—Solution and pseudo-solution. Part III. The electrical convection of certain dissolved substances. Journal of the Chemical Society, Transactions, 71, 568-573.

Piñeda, J., Hare, J. A., & Sponaugle, S. U. (2007). Larval transport and dispersal in the coastal ocean and consequences for population connectivity. *Oceanography*, 20(3), 22-39.

Piry S, Alapetite A, Cornuet, J.-M., Paetkau D, Baudouin, L., Estoup, A. (2004) GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection. *Journal of Heredity* 95:536-539.

Plouviez, S., Faure, B., Le Guen, D., Lallier, F. H., Bierne, N., & Jollivet, D. (2013). A new barrier to dispersal trapped old genetic clines that escaped the Easter Microplate tension zone of the Pacific vent mussels. *PloS One*, *8*(12), e81555.

Plouviez, S., Shank, T. M., Faure, B., Daguin-Thiebaut, C., Viard, F., Lallier, F. H., & Jollivet, D. (2009). Comparative phylogeography among hydrothermal vent species along the East Pacific Rise reveals vicariant processes and population expansion in the South. *Molecular Ecology*, *18*(18), 3903-3917.

Plouviez, S., Le Guen, D., Lecompte, O., Lallier, F. H., & Jollivet, D. (2010). Determining gene flow and the influence of selection across the equatorial barrier of the East Pacific Rise in the tube-dwelling polychaete *Alvinella pompejana*. *BMC Evolutionary Biology*, *10*(1), 220.

Porter, A. H., & Johnson, N. A. (2002). Speciation despite gene flow when developmental pathways evolve. *Evolution*, *56*(11), 2103-2111.

Pörtner, H. O. (2002). Climate variations and the physiological basis of temperature dependent biogeography: systemic to molecular hierarchy of thermal tolerance in animals. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *132*(4), 739-761.

Poulsen, N., Nielsen, E. E., Schierup, M. H., Loeschcke, V., & GrØNkjÆR, P. (2006). Long-term stability and effective population size in North Sea and Baltic Sea cod (Gadus morhua). *Molecular Ecology*, *15*(2), 321-331.

Pradillon, F., Shillito, B., Young, C. M., & Gaill, F. (2001). Deep-sea ecology: Developmental arrest in vent worm embryos. *Nature*, *413*(6857), 698.

Pradillon, F. (2002). Données sur les processus de reproduction et de développement précoce d'un eucaryote thermophile Alvinella pompejana (Doctoral dissertation, Paris 6).

Pradillon, F., & Gaill, F. (2003). Oogenesis characteristics in the hydrothermal vent polychaete *Alvinella pompejana*. *Invertebrate Reproduction & Development*, *43*(3), 223-235.

Pradillon, F., Le Bris, N., Shillito, B., Young, C. M., & Gaill, F. (2005). Influence of environmental conditions on early development of the hydrothermal vent polychaete *Alvinella pompejana*. *Journal of Experimental Biology*, 208(8), 1551-1561.

Pradillon, F., Zbinden, M., Mullineaux, L. S., & Gaill, F. (2005). Colonisation of newly-opened habitat by a pioneer species, *Alvinella pompejana* (Polychaeta: Alvinellidae), at East Pacific Rise vent sites. *Marine Ecology Progress Series*, *302*, 147-157.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., ... & Cagan, A. (2013). Great ape genetic diversity and population history. *Nature*, *499*(7459), 471.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Prout, T. (1968). Sufficient conditions for multiple niche polymorphism. *The American Naturalist*, *102*(928), 493-496.

Raupach, M. J., Thatje, S., Dambach, J., Rehm, P., Misof, B., & Leese, F. (2010). Genetic homogeneity and circum-Antarctic distribution of two benthic shrimp species of the Southern Ocean, Chorismus antarcticus and Nematocarcinus lanceopes. *Marine Biology*, *157*(8), 1783-1797.

Ravaux, J., Hamel, G., Zbinden, M., Tasiemski, A. A., Boutet, I., Léger, N., ... & Shillito, B. (2013). Thermal limit for metazoan life in question: in vivo heat tolerance of the Pompeii worm. *PLoS One*, *8*(5), e64074. Reed, D. H., & Frankham, R. (2003). Correlation between fitness and genetic diversity. *Conservation Biology*, *17*(1), 230-237.

Rieman, B. E., & McIntyre, J. D. (1995). Occurrence of bull trout in naturally fragmented habitat patches of varied size. *Transactions of the American Fisheries Society*, *124*(3), 285-296.

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

Rogers, A. D. (2007). Evolution and biodiversity of Antarctic organisms: a molecular perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1488), 2191-2214.

Romiguier, J., Gayral, P., Ballenghien, M., Bernard, A., Cahais, V., Chenuil, A., ... & Loire, E. (2014). Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, *515*(7526), 261.

Rose, M. R., & Doolittle, W. F. (1983). Molecular biological mechanisms of speciation. *Science*, 220(4593), 157-162.

Rouse, G. W. (2000). Polychaetes have evolved feeding larvae numerous times. *Bulletin of Marine Science*, 67(1), 391-409.

Rouse, G., & Pleijel, F. (2001). Polychaetes. Oxford university press.

Rouse, G. W., & Pleijel, F. (2006). Annelid phylogeny and systematics. In G. W. Rouse & F. Pleijel (Eds.), Reproductive biology and phylogeny of Annelida (Vol. 4, pp. 3–21). Enfield: Science.

Rousset, F. (1995). GENEPOP (Version 1.2): Population genetics software for exact tests and ecumenicalism. *Journal of Heredity*, 83, 239.

Rousset, F. (1997). Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics*, *145*(4), 1219-1228.

Rousset, F. (2008). genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Molecular Ecology Resources*, 8(1), 103-106.

Ruppert, E. E., Barnes, R. D., & Fox, R. S. (2004). *Invertebrate zoology: a functional evolutionary approach*. 7th Ed. Brooks/Cole, Thomson Learning learning, Inc. 990 p.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, *230*(4732), 1350-1354. Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463-5467.

Scher, H. D., & Martin, E. E. (2006). Timing and climatic consequences of the opening of Drake Passage. *Science*, *312*(5772), 428-430.

Selander, R. K. (1970). Behavior and genetic variation in natural populations. *American Zoologist*, *10*(1), 53-66.

Shank, T. M., Fornari, D. J., Von Damm, K. L., Lilley, M. D., Haymon, R. M., & Lutz, R. A. (1998). Temporal and spatial patterns of biological community development at nascent deep-sea hydrothermal vents (9 50' N, East Pacific Rise). *Deep Sea Research Part II: Topical Studies in Oceanography*, 45(1), 465-515.

Shapiro, J. A., Huang, W., Zhang, C., Hubisz, M. J., Lu, J., Turissini, D. A., ... & Chen, Z. (2007). Adaptive genic evolution in the Drosophila genomes. *Proceedings of the National Academy of Sciences*, *104*(7), 2271-2276.

Singhal, S., Huang, H., Donnellan, S. C., Holmes, I., & Rabosky, D. L. (2017). Genetic diversity is largely unpredictable but scales with museum occurrences in a species-rich clade of Australian lizards. *Proceedings of Royal Society of London Publishing*, 284(1854), 2588.

Slatkin, M. (1977). Gene flow and genetic drift in a species subject to frequent local extinctions. *Theoretical Population Biology*, *12*(3), 253-262.

Smith, J. M. (1970). Genetic polymorphism in a varied environment. *The American Naturalist*, *104*(939), 487-490.

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

Smith, J. M., & Hoekstra, R. (1980). Polymorphism in a varied environment: how robust are the models?. *Genetics Research*, *35*(1), 45-57.

Smith, C. R. (1989). Vent fauna on whale remains. Nature, 341, 27-28.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of E. coli. *Nature*, *387*(6634), 703.

Snyder, R. E. (2006). Multiple risk reduction mechanisms: can dormancy substitute for dispersal?. *Ecology Letters*, *9*(10), 1106-1114.

Somero, G. N., & Hochachka, P. W. (1971). Biochemical adaptation to the environment. *American Zoologist*, *11*(1), 159-167.

Somero, G. N., & Soulé, M. (1974). Genetic variation in marine fishes as a test of the niche-variation hypothesis. *Nature*, *249*(5458), 670.

Somero, G. N. (1992). Adaptations to high hydrostatic pressure. *Annual Review of Physiology*, 54(1), 557-577.

Somero, G. N. (2002). Thermal physiology and vertical zonation of intertidal animals: optima, limits, and costs of living. *Integrative and Comparative Biology*, *42*(4), 780-789.

Somero, G. N. (2004). Adaptation of enzymes to temperature: searching for basic "strategies". *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *139*(3), 321-333.

Staton, J. L., Schizas, N. V., Klosterhaus, S. L., Griffitt, R. J., Chandler, G. T., & Coull, B. C. (2002). Effect of salinity variation and pesticide exposure on an estuarine harpacticoid copepod, Microarthridion littorale (Poppe), in the southeastern US. *Journal of Experimental Marine Biology and Ecology*, 278(2), 101-110.

Storch, V., & Gaill, F. (1986). Ultrastructural observations on feeding appendages and gills of Alvinella pompejana (Annelida, Polychaeta). *Helgoländer Meeresuntersuchungen*, *40*(3), 309.

Strass, V. H., Leach, H., Prandke, H., Donnelly, M., Bracher, A. U., & Wolf-Gladrow, D. A. (2017). The physical environmental conditions for biogeochemical differences along the Antarctic Circumpolar Current in the Atlantic Sector during late austral summer 2012. *Deep Sea Research Part II: Topical Studies in Oceanography*, *138*, 6-25.

Strobeck, C. (1974). Sufficient conditions for polymorphism with N niches and M mating groups. *The American Naturalist*, *108*(960), 152-156.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., ... & Bleidorn, C. (2011). Phylogenomic analyses unravel annelid evolution. *Nature*, *471*(7336), 95.

Sverdrup, H. U., Johnson, M. W., & Fleming, R. H. (1942). *The Oceans: Their physics, chemistry, and general biology* (Vol. 7). Prentice-Hall, New York.

Swanson, W. J., Nielsen, R., & Yang, Q. (2003). Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution*, 20(1), 18-20.

Taddei, F., Radman, M., Maynard-Smith, J., Toupance, B., Gouyon, P. H., & Godelle, B. (1997). Role of mutator alleles in adaptive evolution. *Nature*, *387*(6634), 700.

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

Takahata, N. (1983). Gene identity and genetic differentiation of populations in the finite island model. *Genetics*, *104*(3), 497-512.

Thatje, S., Hillenbrand, C. D., & Larter, R. (2005). On the origin of Antarctic marine benthic community structure. *Trends in Ecology & Evolution*, *20*(10), 534-540.

Thormann, C. E., Ferreira, M. E., Camargo, L. E. A., Tivang, J. G., & Osborn, T. C. (1994). Comparison of RFLP and RAPD markers to estimating genetic relationships within and among cruciferous species. *Theoretical and Applied Genetics*, *88*(8), 973-980.

Tivey, M. K., Bradley, A. M., Joyce, T. M., & Kadko, D. (2002). Insights into tide-related variability at seafloor hydrothermal vents from time-series temperature measurements. *Earth and Planetary Science Letters*, 202(3), 693-707.

Tokuriki, N., & Tawfik, D. S. (2009). Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology*, *19*(5), 596-604.

Tunnicliffe, V. (1991). The biology of hydrothermal vents: ecology and evolution. *Oceanographic and Marine Biology Annual Review.*, 29, 319-407.

Tunnicliffe, V., McArthur, A. G., & McHugh, D. (1998). A biogeographical perspective of the deepsea hydrothermal vent fauna. In *Advances in marine biology* (Vol. 34, pp. 353-442). Academic Press.

Van Dover, C. L., & Trask, J. L. (2000). Diversity at deep-sea hydrothermal vent and intertidal mussel beds. *Marine Ecology Progress Series*, 169-178.

Vaughan, D. G., Marshall, G. J., Connolley, W. M., Parkinson, C., Mulvaney, R., Hodgson, D. A., ... & Turner, J. (2003). Recent rapid regional climate warming on the Antarctic Peninsula. *Climatic Change*, 60(3), 243-274.

Vellend, M. (2005). Species diversity and genetic diversity: parallel processes and correlated patterns. *The American Naturalist*, *166*(2), 199-215.

Via, S., & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular ecology*, *17*(19), 4334-4345.

Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them?. *Molecular Ecology*, *23*(8), 2006-2019.

Villemereuil, P., & Gaggiotti, O. E. (2015). A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11), 1248-1258.

Vitalis, R. (2012). Det Sel: An R-Package to Detect Marker Loci Responding to Selection. *Data Production and Analysis in Population Genomics: Methods and Protocols*, 277-293.

Vrijenhoek, R. C. (1997). Gene flow and genetic diversity in naturally fragmented metapopulations of deep-sea hydrothermal vent animals. *Journal of Heredity*, *88*(4), 285-293.

Vrijenhoek, R. C., Shank, T., & Lutz, R. A. (1998). Gene flow and dipersal in deep-sea hydrothermal vent animals. *Cahiers de Biologie Marine*, (3-4).

Wang, G. Z., & Lercher, M. J. (2010). Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evolutionary Biology*, *10*(1), 263.

Watanabe, M., Mochida, K., Kato, T., Tabata, S., Yoshimoto, N., Noji, M., & Saito, K. (2008). Comparative genomics and reverse genetics analysis reveal indispensable functions of the serine acetyltransferase gene family in Arabidopsis. *The Plant Cell*, 20(9), 2484-2496.

Watremez, P., & Kervevan, C. (1990). Origine des variations de l'activité hydrothermale: premiers éléments de réponse d'un modèle numérique simple. *Comptes rendus de l'Académie des sciences. Série 2, Mécanique, Physique, Chimie, Sciences de l'univers, Sciences de la Terre, 311*(1), 153-158.

Watterson, G.A., 1975 On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology*. 7: 256-276.

Whitlock, M. C., & McCauley, D. E. (1990). Some population genetic consequences of colony formation and extinction: genetic correlations within founding groups. *Evolution*, 44(7), 1717-1724.

Widmer, A., & Lexer, C. (2001). Glacial refugia: sanctuaries for allelic richness, but not for gene diversity. *Trends in Ecology & Evolution*, *16*(6), 267-269.

Won, Y., Young, C. R., Lutz, R. A., & Vrijenhoek, R. C. (2003). Dispersal barriers and isolation among deep-sea mussel populations (Mytilidae: Bathymodiolus) from eastern Pacific hydrothermal vents. *Molecular Ecology*, *12*(1), 169-184.

Wong, A., Forbes, M. R., & Smith, M. L. (2001). Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. *Genome*, *44*(4), 677-684.

Wright, S. (1932). *The roles of mutation, inbreeding, crossbreeding, and selection in evolution* (Vol. 1, No. 8). na.

Wright, S. (1943). Isolation by distance. Genetics, 28(2), 114.

Wright, S., (1969) Evolution and the Genetics of Populations, Vol. II. The Theory of Gene Frequencies. University of Chicago Press, Chicago.

Wright, S. (1982). The shifting balance theory and macroevolution. *Annual review of genetics*, *16*(1), 1-20.

Xia, Y., & Levitt, M. (2002). Roles of mutation and recombination in the evolution of protein thermodynamics. *Proceedings of the National Academy of Sciences*, *99*(16), 10382-10387.

Yan, J., Song, Z., Xu, Q., Kang, L., Zhu, C., Xing, S., ... & Sang, T. (2017). Population transcriptomic characterization of the genetic and expression variation of a candidate progenitor of Miscanthus energy crops. *Molecular Ecology*, *26*(21), 5911-5922.

Yang, Z., & Nielsen, R. (1998). Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution*, *46*(4), 409-418.

Yang, Z., & Nielsen, R. (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution*, *17*(1), 32-43.

Yeaman, S., & Otto, S. P. (2011). Establishment and maintenance of adaptive genetic divergence under migration, selection, and drift. *Evolution*, 65(7), 2123-2129.

Young, A., Boyle, T., & Brown, T. (1996). The population genetic consequences of habitat fragmentation for plants. *Trends in Ecology & Evolution*, *11*(10), 413-418.

Zarlenga, D., Wang, Z., & Mitreva, M. (2016). Trichinella spiralis: Adaptation and parasitism. *Veterinary Parasitology*, 231, 8-21.

Zeldovich, K. B., Berezovsky, I. N., & Shakhnovich, E. I. (2007). Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Computational Biology*, *3*(1), e5.

Zuckerkandl, E., & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins* (pp. 97-166).

Résumé :

La question de la variabilité des environnements disponibles pour une population donnée s'est déjà posée par le passé. Levene (1953) a cherché à voir dans quelle mesure il est possible d'avoir un maintien du polymorphisme dans une situation où plus d'une niche écologique est disponible. Par la suite, d'autres se sont intéressés aux conditions (forces de sélection, taille de la niche, etc) dans lesquelles un environnement variable dans l'espace, c'est-à-dire un environnement regroupant plusieurs niches, pouvait induire ou maintenir du polymorphisme. En revanche la variabilité intrinsèque de l'environnement est peu étudiée et donc par là même l'effet que celle-ci a sur les espèces qui y sont exposées. Cette thèse a pu mettre en évidence le pouvoir de la variabilité environnementale sur le maintien de polymorphisme spontané ou introgressé dans les populations. Ainsi des différences de variabilité environnementale entre populations d'une même espèce impliquent le développement d'adaptations et peut générer et renforcer de la différenciation entre celles-ci. Cette variabilité s'additionne aux traits d'histoire de vie et au bagage génétique des espèces pour marquer le génome et laisser des traces caractéristiques des environnements aux variabilités thermiques contrastées. Elle induit aussi des pressions de sélection différentes et plus ou moins fortes sur les systèmes enzymatiques. Ces pressions de sélection induisent la mise en place de différentes stratégies selon les espèces qui doivent exploiter au mieux les possibilités que leur offre leur bagage enzymatique ancestral.

Mots clés : [Diversité génétique – Variabilité environnementale – Annélides – Sélection – Différenciation génétique]

[Evolutionary history and influence of selection on genetic diversity of polychaete annelids from extreme environments]

Abstract :

The question of the variable environments available for a given population has been raised in the past. Levene (1953) sought to see to what extent it is possible to maintain polymorphism in a situation where more than one ecological niche is available. Subsequently, other authors looked for the conditions (selective pressures, niche size, etc.) in which a spatially variable environment, i.e. an environment grouping several niches, could induce or maintain polymorphism. On the other hand, the intrinsic variability of the environment, and therefore the effect it has on the species exposed to it, is little studied. This thesis has been able to highlight the power of environmental variability on the maintenance of spontaneous or introgressed polymorphism in populations. Thus, differences in environmental variability between populations of the same species imply the development of adaptations and can generate and reinforce the differentiation between populations. Along with the life history traits and genetic background of the species, this variability marks the genome and leaves footprints of the environments characterized, among other things, by a contrasting thermal variability. It also induces different and more or less strong selective pressures on enzymatic systems. These selective pressures lead to the implementation of different strategies depending on the species that must make the most of the possibilities offered by their ancestral enzymatic baggage.

Keywords : [Genetic diversity - Environmental variability - Annelids - Selection - Genetic differentiation]