Université de Lille 1 – Sciences et Technologies

Ecole Doctorale : Sciences de la Matière, du Rayonnement et de l'Environnement.
Laboratoire : Evolution, Ecologie, Paléontologie, UMR CNRS 8198

Claire PAPOT

Histoire évolutive et patrons de sélection d'un gène codant un peptide antimicrobien chez deux annélides extrémophiles : le ver côtier *Capitella capitata* et le ver hydrothermal *Alvinella pompejana*

Thèse de doctorat présentée en vue de l'obtention du grade de Docteur de l'Université de Lille – 1 Sciences et Technologies dans la spécialité Géoscience, Ecologie, Paléontologie, Océanographie.

Soutenue le 20 Décembre 2017

Devant le jury composé de :

Pr. Guillaume MITTA	Professeur - Perpignan	*Rapporteur*
Dr. Nicolas BIERNE	Directeur de Recherche - Université de Montpellier 2	*Rapporteur*
Dr. Nicolas MONTAGNE	Maître de Conférences - UPMC	*Examinateur*
Pr. Xavier VEKEMANS	Professeur - Université de Lille 1	*Examinateur*
Dr. Aurélie TASIEMSKI	Maître de Conférences - Université de Lille 1	*Directrice de thèse*
Dr. Didier JOLLIVET	Chargé de Recherche - Station Biologique de Roscoff	*Directeur de thèse*

SOMMAIRE

CHAPITRE 1. INTRODUCTION GENERALE 1

PARTIE I : Présentation générale des peptides antimicrobiens 1
1. Histoire de la découverte des peptides antimicrobiens (PAMs) 1
2. Immunité innée 2
3. Diversité et mode d'action des PAMs dans le règne animal 4
4. Modalités de maturation, transcription et sécrétion des PAMs chez les métazoaires. 14
5. Fonction des PAMs dans le règne animal 19

PARTIE II – Diversification des effecteurs immunitaires : Les processus engagés dans la diversification des agents immunitaires sont-ils les mêmes que ceux qui s'appliquent aux PAMS ? 22
1. Quelques mécanismes de diversification moléculaire chez les invertébrés 22
2. Histoire évolutive des gènes de l'immunité 25

PARTIE III : Diversité génétique et histoire évolutive des peptides antimicrobiens 27
1. Architecture génique du précurseur protéique 27
2. Diversification par duplications géniques : un moteur dans l'évolution des PAMs 30
3. Diversité génétique des peptides antimicrobiens 35
4. Patrons de sélection chez les gènes codant les PAMs 38
5. Evolution rapide des PAMs 41

Objectifs de thèse 49

CHAPITRE 2. DIVERSITE GENETIQUE ET HISTOIRE EVOLUTIVE DU GENE CODANT LE PRECURSEUR PROTEIQUE DU PEPTIDE ANTIMICROBIEN ALVINELLACINE CHEZ *ALVINELLA POMPEJANA*. 51

INTRODUCTION 51

Matériel et Méthodes 59
1. L'échantillonnage 59
2. Le précurseur protéique de l'alvinellacine : la préproalvinellacine. 60
3. Acquisition des données 62
4. Méthodes d'analyses 68
5. Induction du gène de la préproalvinellacine en réponse à un stress abiotique 73
6. Analyse différenciation génétique nord/sud de l'EPR 74

Résultats 76
1. Histoire phylogénétique de l'alvinellacine dans la famille des Alvinellidae 76
2. Amplification du gène et recapture des allèles 77
3. Diversification génique au sein du genre *Alvinella* 78
4. Identification et caractérisation de la famille multigénique chez *A. pompejana*. 82
5. Diversité génétique et tests de neutralité associés aux différents paralogues 89
6. Recherche de sélection positive au sein des différents domaines du préproalvinellacine 90
7. Induction de la préproalvinellacine sous différents stress. 104
8. Différenciation des populations nord/sud au gène de la préproalvinellacine 105

DISCUSSION 108

CHAPITRE 3. DIVERSITE GENETIQUE ET HISTOIRE EVOLUTIVE DU GENE CODANT LE PRECURSEUR PROTEIQUE DE LA PREPROCAPITELLACINE CHEZ L'ANNELIDE COTIER *CAPITELLA SPP.* 124

INTRODUCTION 124

Matériel et méthodes 134
1. Echantillonnage des *Capitella spp.* 134
2. Gènes analysés 135
3. ACQUISITION DES SEQUENCES 138
4. Analyse des données 141

Résultats 151
1. Analyse phylogéographique du complexe *Capitella spp.* avec *Cox-1* – Assignation des espèces françaises par rapport au complexe mondial d'espèces cryptiques. 151
2. Divergence nette inter clades 154
3. Description de la diversité génétique mitochondriale dans les populations Françaises 155
4. Diversité génétique du précurseur protéique de la Capitellacine 164
5. Analyse IMa2- Test de l'hypothèse d'isolement des populations avec migration 178
6. Evolution du ratio d_N/d_S au sein et entre les clades de la capitellacine. 187
7. Polymorphisme et divergence au niveau de la région du PAM. 195

DISCUSSION 198

CHAPITRE 4 : DISCUSSION GENERALE ET PERSPECTIVES 216
1. Comparaison des principaux résultats sur les processus évolutifs ayant façonné la diversité génétique des deux précurseurs protéiques de peptide antimicrobien. 219
2. Comparaison des activités et thermostabilité des 2 peptides antimicrobiens : études préliminaires réalisées au laboratoire dans le cadre du stage de M2 de Lolita Roisin. 226
3. Perspectives 228

BIBLIOGRAPHIE 241

ANNEXES 271
Annexe 1. Caractérisation des transcrits de la preproalvinellacine et mise en évidence de transcrits tronqués 272
Annexe 2. Article publié dans Scientific Reports. 278
Annexe 3. Valeurs de F_{st} calculées pour les clades de la capitellacine au sein des deux régions 5' et 3' par pairs de populations et de façon globale. 316
Annexe 4. Protocole de productions des variants du BRICHOS 317
Annexe 5. Liste des articles et valorisations. 321

Chapitre 1. Introduction générale

PARTIE I : Présentation générale des peptides antimicrobiens

1. Histoire de la découverte des peptides antimicrobiens (PAMs)

Dès la fin du 19eme siècle ont été reportées des activités antimicrobiennes dans les tissus d'organismes sans que les molécules incriminées puissent réellement être purifiées et/ou isolées. Par exemple, chez les plantes, les premières substances à activité antimicrobienne ont été décrites par Jago & Jago en 1895. Alexander Fleming en 1928 découvre pour la première fois qu'un champignon inhibe la croissance des staphylocoques qu'il étudie pour finalement publier ses résultats sur la pénicilline en 1928. Cette substance sera largement étudiée puis utilisée au cours de la seconde guerre mondiale dans un but thérapeutique et vaudra à Fleming le prix Nobel de médecine en 1945. Après les premières découvertes de ces substances, s'en suit l' « âge d'or » des antibiotiques entrainant ensuite l'augmentation des cas de résistance à la pénicilline chez de nombreux agents pathogènes ayant pour conséquence la recherche de nouvelles molécules antibiotiques. La caractérisation plus récente de peptides ayant une activité antimicrobienne à large spectre permettra ensuite d'élargir l'éventail de la réponse antibactérienne des métazoaires, les plantes et même les bactéries, avec un nombre croissant de molécules toutes composées d'un petit nombre d'acides aminés (moins de 100). La recherche sur les peptides antimicrobiens est en plein essor depuis les années 1980 à la suite de la découverte des cécropines par Boman et Steiner, 1981 chez *Hyalophora cecropia*, des défensines chez le lapin par Lehrer et Selsted en 1983 (Lehrer et al., 1983) suivi par leur découverte chez l'homme en 1985 par Ganz et Selsted (Ganz et al., 1985) et finalement de la découverte des magainines chez *Xenopus laevis* par Michael Zasloff en 1987 (Zasloff, 1987). Les PAMs sont depuis largement reconnus comme une nouvelle classe de molécules ayant un rôle clé dans l'immunité d'un grand nombre d'organismes (Bulet et al., 1999, 2004). Ils ont en effet été détectés à la fois chez les vertébrés, les invertébrés, les plantes, les bactéries, les champignons et les archées (Bulet et al., 1999, 2004; Castro and Fontes, 2005; Zasloff, 2002).

A l'heure actuelle, plus de 2700 PAMs ont été découverts (Antimicrobial Peptide Database : http://aps.unmc.edu/AP/main.php) et représentent sûrement une faible proportion de ce qui est produit dans la nature. Dans tous les cas, ces molécules agissent comme première ligne de défense de l'hôte (bien qu'ayant d'autres rôles) en tuant directement une large gamme de procaryotes/microeucaryotes ou encore de virus (Zasloff, 2002).

2. Immunité innée

En premier lieu, des barrières naturelles de l'hôte pour faire face à l'entrée des pathogènes existent que ce soit par la coquille, la peau/les muqueuses chez les vertébrés ou le mucus, les cuticules … chez les invertébrés. La température interne en inhibant la croissance des pathogènes ou encore le pH acide des fluides corporels peuvent également représenter des barrières défensives de l'organisme. L'immunité inné, présente chez tous les métazoaires (Medzhitov and Janeway, 2000), met en œuvre des cellules et des molécules spécialisées dans diverses actions permettant le blocage et l'élimination des pathogènes et elle comprend à la fois une composante cellulaire et une composante humorale (basée sur la production de composés solubles de défense comme les peptides antimicrobiens). Il sera dans cette partie question de présenter rapidement les différents acteurs responsables de l'immunité innée (chez les vertébrés et les invertébrés).

2.1. Immunité cellulaire

La réponse immunitaire innée implique la participation de plusieurs types cellulaires : les acteurs clés sont les neutrophiles, les macrophages ou encore les cellules tueuses naturelles (natural killer NK) ou les monocytes du sang chez les vertébrés. Globalement, les neutrophiles et les macrophages sont des cellules phagocytaires (mais peuvent aussi être des médiateurs de l'inflammation en produisant des cytokines …), les cellules NK sont capables de lyser des cellules infectées par un virus ou des cellules tumorales et peuvent activer les macrophages (Medzhitov and Janeway, 2000).

Chez les invertébrés et plus particulièrement chez les insectes (drosophile : Buchon et al., 2014), les hémocytes sont les cellules spécialisées dans la réponse immunitaire et sont retrouvées dans l'hémolymphe. Chez les annélides, ce sont différentes populations de coelomocytes qui jouent un rôle dans la défense immunitaire : les cellules hyalines participent à l'encapsulation des pathogènes et l'inactivation de ceux-ci, les cœlomocytes

granuleux (appelés granulocytes) sont responsables des activités de phagocytose et activités cytotoxiques et les éléocytes constituent l'équivalent du corps gras des insectes (Cuvillier-Hot et al., 2014).

2.2. Immunité humorale

Les cellules immunitaires peuvent secréter des peptides antimicrobiens ou des protéines à activités antimicrobiennes telles que les lysozymes qui sont des enzymes détruisant les parois des bactéries à gram positif en catalysant l'hydrolyse des peptidoglycanes qui les composent. Les PAMs quant à eux, dont la présentation générale sera effectuée dans la prochaine partie de l'introduction, représentent une composante clé de la réponse immunitaire innée et représentent un des moyens de défense partagé par l'ensemble des espèces vivantes.

La reconnaissance du pathogène est le premier pas permettant l'initiation de la réponse immunitaire innée : la synthèse/sécrétion des effecteurs immunitaires (tels que les PAMs) est conditionnée par cette identification (bien qu'une expression constitutive puisse également être rapportée chez certains organismes).

Cette reconnaissance est basée sur la liaison de motifs structurants (les PAMPs : Pathogen-Associated Molecular Patterns) conservés des pathogènes (virus, bactéries, champignons, protozoaires…) à des récepteurs à large spectre de l'organisme : les PRR (Pattern Recognition Receptor »). Parmi ces PRR, les lectines, les PGRP (Peptidoglycan Recognition Proteins), les GNBP1 (Gram-negative Binding Proteins) ou encore les DsCAM (Down syndrom Cell Adhesion Molecule) ou les FREPs (Fibrinogen Related Proteins) et RLR (RIG-I-Like Receptor) reconnaissent les PAMPs tels que les lipopolysaccharides (LPS), les peptidoglycanes (Leulier et al., 2003) ou encore les mannanes et ARN/ADN viraux. Chez les vertébrés, le plus célèbre des PRR appartient à la famille des TLR (Toll-like receptors) : par exemple TLR4 nécessaire à l'activation de la réponse immunitaire par reconnaissance de LPS.

La liaison PAMPs-PRR déclenche différentes cascades protéolytiques (non détaillées ici) ou voies de signalisations intracellulaires spécifiques du type de microorganisme reconnu, permettant ainsi la production de peptides antimicrobiens et ainsi l'élimination des microorganismes (réponse humorale).

Il est important de noter que différentes molécules très diversifiées (de la famille des immunoglobulines) ont pu être mises en évidence chez les invertébrés (ce qui permettrait de constituer la base d'une immunité spécifique rappelant l'immunité adaptative des vertébrés). Cependant bien que des membres de la superfamille des immunoglobulines aient pu être décrits chez les invertébrés, il apparaît que les immunoglobulines qui sont les effecteurs de l'immunité adaptative manquent chez les invertébrés. C'est ce défaut qui a permis pendant longtemps d'affirmer que les invertébrés étaient dépourvus de réponse adaptative ou, en tout cas, de spécificité et, que l'immunité des invertébrés reposait sur un nombre limité de PRR ne reconnaissant que des structures moléculaires conservées des pathogènes (les PAMPs donc).

Parmi ces molécules peuvent être cités les hémolines, les FREP (pour *fibrinogen-related proteins)* qui sont des protéines hyper-diversifiées constituées à la fois d'un (ou deux) domaine(s) de la superfamille des immunoglobulines (IgSF) en position amino-terminale mais aussi d'un domaine de type fibrinogène (FReD) en position carboxy-terminale (Adema et al., 1997) qui contient le site potentiel de liaison aux résidus glucidiques (CRD). Ces protéines ont été découvertes d'abord chez le mollusque gastéropode *Biomphalaria glabrata* puis chez de nombreux invertébrés parmi lesquels chez les annélides peuvent être cités *Helobdella robusta* et *Capitella spp* (Hanington and Zhang, 2011). Quelques mécanismes de diversification génétique des molécules de l'immunité innée des invertébrés, rappelant des caractéristiques de l'immunité adaptative des vertébrés, seront traités dans la Partie II.

3. Diversité et mode d'action des PAMs dans le règne animal

3.1. Définition

Les PAMs sont des molécules de petite taille : majoritairement entre 12-50 acides aminés, et toujours inférieurs à 200 acides aminés (Bulet et al., 1999). Ceux-ci sont également majoritairement cationiques (charge positive à pH physiologique), et composés de 40 à 50% de résidus hydrophobes leur permettant d'adopter le plus souvent une structure amphipathique (composé d'un pôle hydrophile et un pôle hydrophobe) : leurs résidus cationiques sont spatialement séparés des résidus hydrophobes ce qui leur permet d'interagir avec la bicouche lipidique des membranes bactériennes (Figure 1). C'est cette

nature amphipatique du peptide qui lui permet d'être soluble en milieu aqueux tout en gardant sa capacité à se lier aux membranes bactériennes dont les surfaces sont hydrophobes.

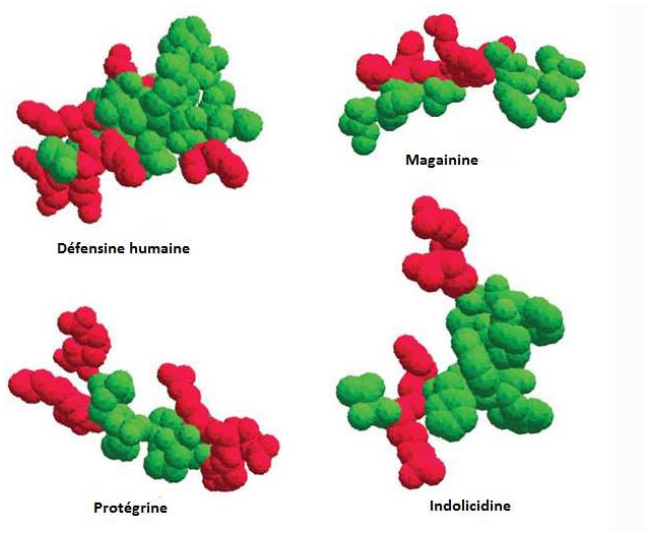


Figure 1. Ségrégation des résidus cationiques et hydrophobes chez plusieurs PAMs de différentes classes illustrant l'amphipatie commune à la plupart des PAMs. (Rouge = acides aminés chargés positivement, vert : acides aminés hydrophobes). De Zasloff, 2002.

3.2. Mode d'action des PAMS

Le mode d'action des PAMs peut être divisé en deux catégories : i) des peptides qui se lient aux membranes bactériennes, créant des pores de façon irréversible ou déstabilisant la membrane et générant l'efflux du contenu cytoplasmique ou ii) des peptides qui traversent la membrane bactérienne et inhibent la division cellulaire en se liant à des composants intracellulaires.

3.2.1. Liaison et destabilisation des membranes bactériennes.

Les PAMs présentent une charge nette positive et un taux élevé d'acides aminés hydrophobes leur permettant de se lier sélectivement aux membranes bactériennes chargées négativement. Cette liaison à la membrane bactérienne conduit à une perturbation non enzymatique de la trame membranaire (Zhang and Gallo, 2016). La capacité des PAMs à tuer les micro-organismes dépend généralement de leur capacité à interagir avec les membranes bactériennes en les distinguant des cellules eucaryotes par la faible proportion de cholestérol et généralement l'absence de phospholipides anioniques. En effet, le cholestérol s'intercale entre les phospholipides avec pour conséquence de réduire les

possibilités d'interaction du PAM ce qui l'empêche d'être actif sur les membranes cellulaires de l'hôte. Cette interaction PAM/phospholipide membranaire connue sous le nom de « modèle de Shai Matsuzaki-Huang » (SMH) est un modèle qui fournit un résumé raisonnable de ce mode d'action vis-à-vis de la membrane plasmique des bactéries.
Certaines étapes de l'activité antimicrobienne ont pu être mises en évidence et modélisées telle que l'attraction du peptide par la surface bactérienne via des forces électrostatiques liées à la complémentarité des charges membranaires externes avec le PAM et sa fixation à la membrane (via les lipopolysaccharides (LPS) pour les bactéries à Gram négatif et les acides teichoïque/lipoteichoïque/lysylphosphatidylglycerol pour les Gram positif). Cette fixation se fait le plus souvent après avoir traversée la capsule polysaccharidique, l'insertion du peptide entraînant la perméabilité membranaire et *in fine* la fuite du contenu cellulaire de la bactérie cible. Cette dernière étape fait l'objet de plusieurs modèles théoriques (Figure 2) :

A) le modèle dit du "tonneau" (ou « barrel-stave model » : Oren and Shai, 1998) qui se traduit par l'agrégation des peptides jusqu'à une certaine concentration qui conduit à la formation de pores transmembranaires. Ici, les peptides se positionnent en cercle créant le pore dans la bicouche lipidique de la membrane du pathogène. Pour ce modèle, les régions hydrophobes s'alignent avec les têtes des lipides de la bicouche lipidique alors que les régions hydrophiles des peptides forment la région intérieure du pore.

B) le modèle dit du "tapis" (ou « carpet model » : Pouny *et* al. 1992) qui se traduit par la déformation et la déstabilisation de la bicouche lipidique, les peptides étant en contact avec les têtes polaires des lipides mais sans jamais s'insérer dans la membrane. Les peptides vont se fixer les uns aux autres pour former un tapis qui, à forte concentration, provoquera ensuite la perturbation de la structure de la membrane et la formation d'une micelle (agrégat sphéroïdal de molécules possédant une tête polaire hydrophile dirigée vers le solvant et une chaîne hydrophobe dirigée vers l'intérieur).

C) la création de pores toroïdaux. Ce modèle ressemble au premier modèle sauf que les peptides s'agrègent et s'insèrent en forant la membrane sans déformation de la double couche lipidique. Plus précisément, les peptides dans ce modèle sont associés avec les têtes des lipides même lorsque les PAMs sont insérés perpendiculairement à la bicouche lipidique.

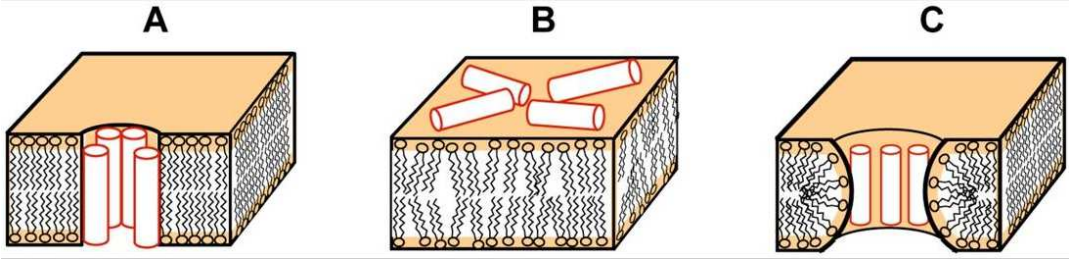


Figure 2. Différents modèles de fixation du peptide à la membrane plasmique (de Tang and Hong, 2009).

D'autres modèles (Figure 3) sont désormais proposés et décrivent de nouveaux procédés de disruption ou représentent une complexification d'un modèle préexistant (exemple : « disordered toroidal pore model »).

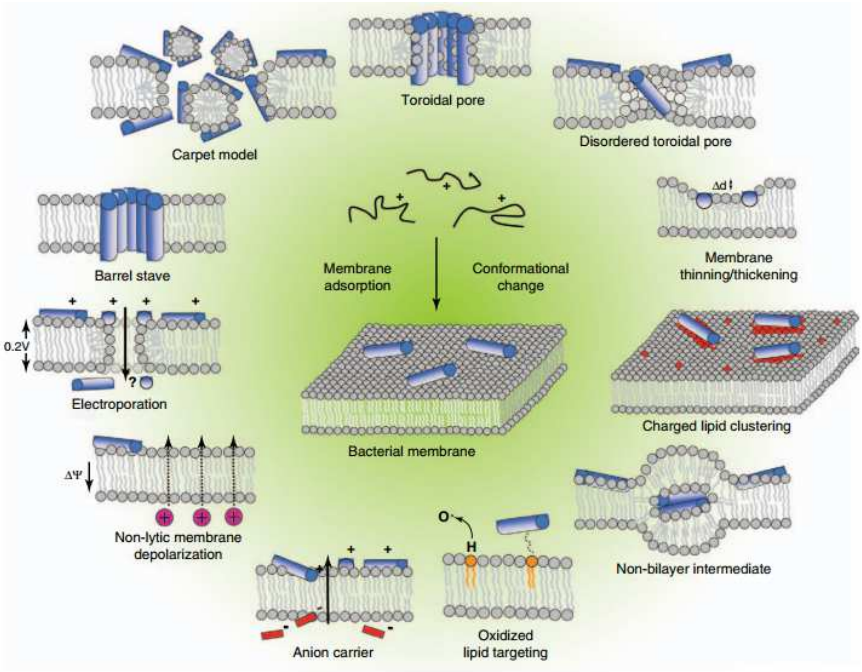


Figure 3. Autres modèles de disruption de la membrane cytoplasmique proposés par (Nguyen et al., 2011).

Par exemple, le modèle dit « anion carrier » est un nouveau modèle documenté dans le cas de l'indolicine dans lequel le peptide se couple à des anions au travers de la membrane pour finalement entraîner leur efflux vers le milieu extracellulaire (Rokitskaya et al., 2011). Dans le modèle révisé du pore toroïdal désordonné, la formation de pores est plus aléatoire et

nécessite moins de peptides, l'absorption du peptide sur la membrane peut être améliorée en ciblant les phospholipides oxydés …

3.2.2. Liaison avec des éléments du contenu intracellulaire

Certains peptides sont également capables de traverser la bicouche lipidique membranaire sans en altérer la structure et peuvent ainsi tuer les bactéries en inhibant certaines fonctions intracellulaires directement dans le cytoplasme/noyau (Figure 4). Ce mode d'action intracellulaire est dans l'ensemble documenté comme pouvant se lier à l'ADN et empêcher la synthèse de macromolécules nécessaires au bon fonctionnement de la cellule, en inhibant la réplication de l'ADN, la synthèse d'ARN ou la traduction (Brogden, 2005). C'est par exemple le cas des apidaecines, drosocines et pyrrhocoricines qui s'associent à la chaperone/heat-shock protein DnaK, chez *Escherichia coli* (Kragol et al. 2001). La bactérie synthétise alors des protéines non-fonctionnelles parce qu'elle ne peut plus les conformer correctement. Cette dernière stratégie est moins susceptible de causer le relargage d'endotoxines et, peut être considérée comme plus sélective contre les micro-organismes (Sperstad et al., 2011).

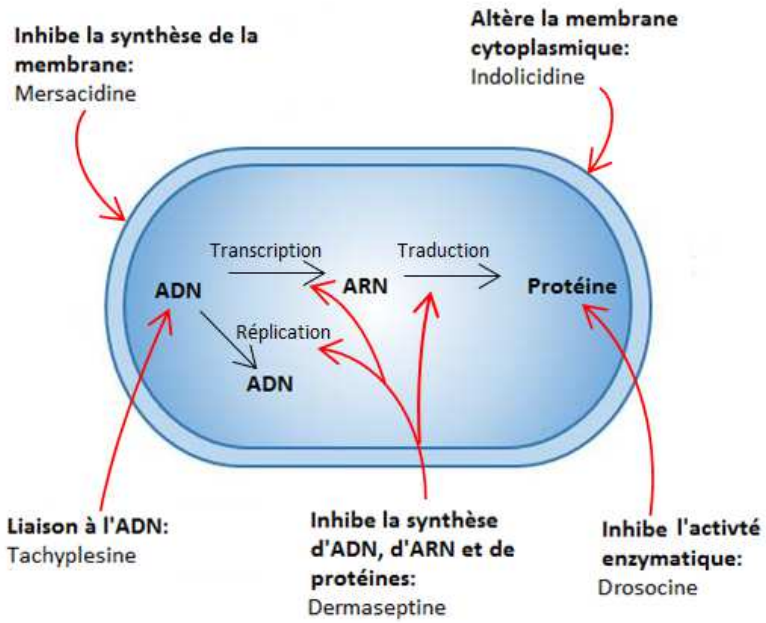


Figure 4. Schéma récapitulatif des possibilités d'interaction à des composants intracellulaires pour inhiber la croissance bactérienne proposé par (Brogden, 2005).

3.2. Diversité des PAMs dans le règne animal

Ainsi, les PAMs sont très diversifiés, et sans histoire phylogénétique commune, malgré le fait que la majorité d'entre eux partagent des caractéristiques structurales telles que la structure en épingle, l'amphipatie et la charge positive de la molécule. Les PAMs cationiques par exemple ne présentent que très peu d'homologie dans leur structure primaire et une large variété de structures secondaires (Zasloff, 2002).

Les PAM peuvent être classés en quatre groupes principaux, basés sur leur structure secondaire et/ou tertiaire, sans aucun lien avec une histoire évolutive, ou l'appartenance à un groupe phylogénique: on peut alors parler de convergence fonctionnelle.

Ces 4 groupes sont (Figure 5) :

(1) les peptides linéaires à hélices alpha (e.g. cécropines, magainines),
(2) les peptides linéaires en épingle à cheveux ou feuillet beta (tachyplésine, gomésine, brévinines).
(3) les peptides riches en hélices alpha et feuillet beta tels que les défensines
(4) les peptides linéaires riches en un ou deux types d'acides aminés (tryptophane, histidine, arginine, glycine).

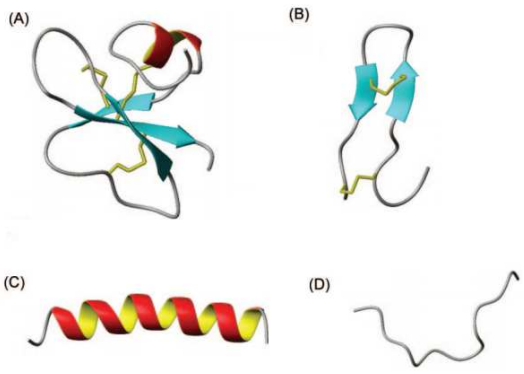


Figure 5. Différents types de peptides antimicrobiens. (A) beta-défensine humaine (B) polyphémusine à feuillet beta, (C) magainine à hélice alpha, et (D) indolicine. Les ponts disulfures sont représentés en jaune.

Ainsi, les peptides peuvent adopter une structure en hélices alpha (magainine) ou peuvent être riches en cystéines impliquées dans la formation de ponts disulfures. Ces derniers

peuvent être constitués de deux feuillets beta repliés en épingle à cheveux par l'intermédiaire d'un seul (arénicine) ou plusieurs ponts disulfures (polyphémusine-alvinellacine). Ces peptides, lorsque plus longs, peuvent également présenter une structure mixte composée de feuillet beta et d'hélices alpha stabilisées par des ponts disulfures (ex : défensine). Les « extended peptides » montrent une structure secondaire qui n'est pas classique et contiennent le plus souvent une grande proportion en acides aminés du type Arg, Trp ou encore Pro (indolicine).

La suite de cette introduction s'intéressera à la diversité des peptides antimicrobiens caractérisés chez les annélides polychètes marins ainsi qu'à leur spécificité au sein des environnements auxquels sont inféodés ces organismes. A ce jour, quelques 40 PAMs différents sont décrits chez les invertébrés marins dont seules les défensines sont communes avec les espèces terrestres.

3.3.1. Les PAMs d'invertébrés marins

Des PAMS ont été caractérisés chez les cnidaires, les annélides, les arthropodes, les mollusques, et les échinodermes (Zasloff, 2002). Une structure riche en ponts disulfures constitue la caractéristique prédominante des PAMs d'invertébrés marins (Sperstad et al., 2011) et la plupart des peptides découverts paraissent inféodés à une espèce ou bien caractéristiques de quelques taxons. Cette spécificité pourrait indiquer soit des lacunes dans l'échantillonnage des PAMs chez les invertébrés marins, soit que ces espèces ont développé des molécules uniques adaptées à leur habitat. Par exemple, chez la moule *Mytilus edulis*, les mytilines (isoformes A et B : Charlet *et* al. 1996) n'ont été retrouvées ensuite que chez l'espèce sœur *M. galloprovincialis* (Mitta *et* al. 2000) et d'autres mytlidae comme les bathymodioles (A. Tanguy, *comm. pers*.). A l'inverse, des peptides ayant des similarités avec les défensines des arthropodes ont également été découverts chez *Mytilus* avec 2 isoformes chez *M. edulis* (contenant 6 cystéines : Charlet et al., 1996) et 2 isoformes plus éloignées chez *M. galloprovincialis* (contenant 8 cystéines : Mitta *et* al. 1999b). Cette famille de PAMs a depuis été caractérisée chez d'autres mollusques (Gestal et al., 2007).

Une autre famille de PAMs, les crustines, a également été mise en évidence chez le crabe *Carcinus maenas* (Relf et al., 1999) mais aussi chez plusieurs espèces de crevettes *Litopenaeus vannamei* et *Litopenaeus setiferus* (Bartlett *et* al.*2002*; Smith *et* al. 2008 pour

review) ou encore chez le homard (Hauton et al., 2006) ou chez la crevette hydrothermale *Rimicaris exoculata* (Zhang et al., 2017). Chez les ascidies, les espèces *halocynthia papillo*sa et *H. aurantium* synthétisent quant à elles des peptides qui leur sont propres (halocyntine et halocidine respectivement : Galinier *et* al. 2009, Jang et al., 2002), ce qui indiquerait que deux espèces proches du point de vue phylogénétique peuvent aussi synthétiser un arsenal de PAMs différent (perte/gain d'isoformes) selon l'environnement associé et/ou le stress biotique reçu. Cependant encore une fois par manque d'investigation, il reste difficile de confirmer de façon fiable le caractère spécifique de ces observations.

Les PAMs d'invertébrés marins peuvent présenter plusieurs isoformes au sein d'une même espèce (cf. partie 3: duplication et/ou épissage) : les pénéidines chez les crevettes pénéides (Cuthbertson *et* al. *2002*; Kato *et* al. 2002), les crustines, les styelines, clavanines, tachystatines, mytilines et arénicine 1 et 2 chez *Arenicola marina* sont d'autres exemples de PAMs exprimés sous différents isoformes (Bartlett et al.; Lehrer et al., 2003; Mitta et al., 2000b; Osaki et al., 1999). En plus de la multiplicité des isoformes d'un PAM, une espèce peut également co-exprimer un cocktail de PAMs selon le stress microbien (Sperstad et al., 2011) comme chez les crabes *Hyas araneus (*Hyastatine et Arasine) *; Tachypleus tridentatus* (tachcitines, tachystatines, big defensines), le tunicié *Styela clava* (Clavanines, Clavaspirines, Styelines) ; les moules *Mytilus edulis* et *M. galloprovincialis* (mytilines, myticines, mytimycines) ou encore l'huître creuse *Crassostrea gigas* (Cg-Prp et défensine).

3.3.2. Les différentes familles de peptides antimicrobiens chez les annélides

Historiquement, la première étude qui a permis de détecter une activité antimicrobienne à partir de fragments d'un annélide remonte à l'étude réalisée par Lassègues et al. (1989) qui ont décrit cette activité bactéricide chez l'oligochète *Eisenia fetida,* activité confirmée ensuite par Pan et al. (2003). Deux petits peptides (6 acides aminés) aux activités antibactérienne et anti tumorale ont ainsi été caractérisés chez cette espèce et appelé F-1 et F-2 (Zhang et al., 2001). Chez les annélides polychètes, des PAMS ont été décrits chez *Perinereis aibuhitensis* (Pan et al., 2004), *Arenicola marina* (Ovchinnikova et al., 2004), *Hediste diversicolor* (Tasiemski et al., 2007), *Alvinella pompejana* et *Capitella capitata* (Tasiemski et al., 2014) ou encore chez *Marphysa sanguinea* (Seo et al., 2016). Des PAMs ont également pu être caractérisés chez les nématodes par exemple avec le cas des PAMs de la

famille des ABF (Antibacterial factor) mais aussi chez les annélides achètes avec les macines (PAM riches en cystéines).

Les macines sont une famille de PAMs riche en cystéines et caractérisés dans un premier temps chez les annélides achètes (sangsues : *Theromyzon tessulatum* et *Hirudo medicinalis* Tt-theromacine et Hm neuromacine et Hm-theromacine : Tasiemski *et* al. 2004; Schikorski *et* al. 2008). Un PAM de cette famille a ensuite été isolée à partir d'*Hydra magnipapillata* (hydramacine -1 : Jung et al., 2009). Toutes ces molécules possèdent 8 cystéines leur permettant de former 4 ponts disulfures, sauf pour la théromacine qui en possède un pont disulfure additionnel (Hung et al., 2014).

Une autre famille de PAM, riche en cystéines, est la famille des arénicines. Cette famille est composée de quatre membres : l'arénicine (deux isoformes : arénicine-1 et arénicine-2), l'alvinellacine et la capitellacine, tous retrouvés chez des annélides polychètes, respectivement *Arenicola marina, Alvinella pompejana et Capitella teleta*.

Le premier PAM de la famille des arenicine-like long de 21 acides aminés (arénicine) a été caractérisé à partir des coelomocytes de l'annélide polychète *Arenicola marina* (Ovchinnikova et al., 2004). Une étude a ensuite caractérisé un PAM de 22 acides aminés (alvinellacine) isolés à partir des coelomocytes de l'espèce hydrothermale *Alvinella pompejana* (Tasiemski et al., 2014). Un orthologue de cette famille a également pu être caractérisé chez l'annélide polychète côtier *Capitella sp* (Tasiemski et al., 2014).

Du point de vue structural, il s'agit de PAM présentant un pont disulfure (arénicine) ou deux ponts disulfures (alvinellacine et capitellacine) et se repliant en feuillet beta double brin antiparallèle. La structure tertiaire est conservée malgré une faible identité de la structure primaire entre ces PAMs. Les trois PAMs sont codés par un gène ayant une architecture exonique et intronique conservée et sont tous issus de la maturation d'un précurseur présentant la structure caractéristique des protéines à domaines BRICHOS, à savoir un peptide signal (région hydrophobe) suivi d'une pro-région (qui contient le domaine BRICHOS) associée à une région C terminale ayant une propension à former des feuillets beta (ici, le PAM) (Willander et al., 2011).

Figure 6. Alignement des prorégions des membres de la famille des arenicine-like (en bleu : région du PAM hautement variable ; trait noir : délimitation du domaine BRICHOS) et structure tridimensionnelle pour (de haut en bas), l'arenicine, l'alvinellacine et la capitellacine. La présence d'une étoile indique une conservation des acides aminés.

L'alignement de la Figure 6 permet d'illustrer la diversité de séquences au sein de la famille des « arénicine-like » notamment au sein de la région codant pour le peptide antimicrobien où un alignement des séquences est impossible. Ainsi, ces peptides sont tous des peptides riches en cystéines (et en arginines) avec une structure secondaire en épingle à cheveux (beta-hairpin) qui ne montre pas de conservation en terme de séquence primaire (Berlov and Maltseva, 2016).

D'autres PAMs ont pu être caractérisés chez les annélides tels la périnerine, l'hédistine et la lumbricine I, les deux premiers étant les seuls représentant de leur famille (seule la lumbricine est présente chez *Alvinella pompejana*). La périnerine est un PAM riche en cystéines caractérisé chez l'espèce *Perinereis aibuhitensis*, long de 51 acides aminés (dont 4 cystéines) et ne possédant aucune similarité avec les autres PAMs décrits (Pan et al., 2004). L'hédistine est un peptide antimicrobien uniquement trouvé chez *Hediste diversicolor* qui adopte quant à lui une structure linéaire qui comprend des résidus bromotryptophanes (Tasiemski et al., 2007). Plus récemment, un peptide antimicrobien (msHémérycine) a également été décrit chez *Marphysa sanguinea* d'une taille de 14 acides amines et dérivé de

la partie N-terminale de l'hémérythrine (Seo et al., 2016). La lumbricine I (peptide riche en prolines) est le seul PAM ayant été caractérisé chez *Lumbricus rubellus* (Cho et al., 1998) *mais retrouvé* chez tous les représentants des annélides *comme A. pompejana* ou *A. marina* mais aussi chez *Pheretima tschiliensis (PP-1), Pheretima guillelmi* (Lombricine-PG) et *Hirudo medicinalis* (Hm-lombricine) (Schikorski et al., 2008; Wang et al., 2003).

4. Modalités de maturation, transcription et sécrétion des PAMs chez les métazoaires.

4.1. Les PAMs avec maturation d'un précurseur protéique

Hormis les diptéranes et lépidoptéranes, la plupart des PAMs d'invertébrés sont isolés à partir des cellules immunitaires (hémocytes ou coelomocytes) et les PAMs sont largement documentés comme étant clivés à partir d'un précurseur protéique plus long et inactif du point de vue de son action antibactérienne (appelé prépropeptide) qui contient un peptide signal en région N-terminale responsable de l'adressage cellulaire vers le réticulum endoplasmique (Zhang and Gallo, 2016). Un peptide synthétisé n'est pas actif sous sa forme non épissée. Le précurseur sera en effet clivé par maturation protéolytique, ce qui va générer deux régions distinctes : la prorégion et le peptide mature qui pourra ensuite exercer ses activités biologiques *in vivo*. Alors que le peptide signal a un rôle dans l'adressage cellulaire du peptide, la prorégion peut avoir différentes fonctions et possède une charge nette négative (Balandin and Ovchinnikova, 2016). En effet, celle-ci joue tout d'abord un rôle de transporteur et d'inhibiteur du PAM dans le but d'éviter toute cytotoxicité vis-à-vis de l'hôte. Ceci a par exemple été documenté chez les insectes avec la cécropine P4 qui est inactive lorsqu'elle est liée au précurseur (Boman et al., 1989; Ueno et al., 2008). Les prorégions d'autres peptides antimicrobiens (alpha défensines de mammifères et cathélicidines par exemple) ont également été décrites comme inhibant les activités des peptides matures (Michaelson et al., 1992). Les cécropines synthétisées chimiquement étant actives, Ueno et al., 2008 en déduisent qu'il est donc peu probable que la prorégion puisse avoir un rôle de chaperone pour prévenir une mauvaise conformation du PAM.

4.2. Modalités d'expression des PAMs d'invertébrés

4.2.1. Une expression inductible

La plupart des PAMs produits le sont de façon inductible suite à une infection bactérienne. Par exemple chez les arthropodes et plus particulièrement chez l'espèce la plus étudiée, la drosophile, les gènes codant pour les PAMs sont induits suite à une infection bactérienne par le corps gras (Lemaitre *et* al. 1997). Les PAMs sont ensuite secrétés dans l'hémolymphe entraînant une réponse systémique à l'infection.

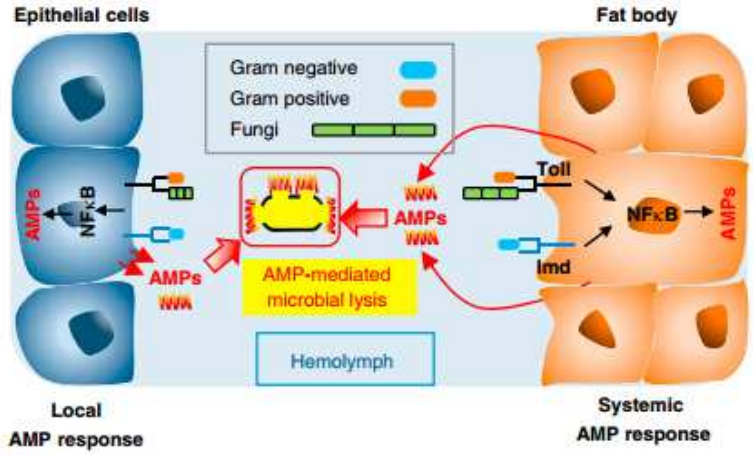


Figure 7. Sécrétion de PAM de façon systémique ou locale chez la drosophile (Zhang and Gallo, 2016).

La Figure 7 permet d'illustrer la synthèse des PAMs chez la drosophile en réponse à une infection microbienne par différentes voies métaboliques qui peut être donc soit systémique (produite par le corps gras et secrété dans l'hémolymphe) soit locale (au niveau des épithéliums). Dans ce dernier cas, les PAMs préviennent alors l'entrée des pathogènes (au sein des cellules épithéliales).

Chez les mammifères, en plus des infections pouvant induire l'expression des gènes codant pour les PAMs, l'expression des cathélicidines par les épitheliums peut également être induite par différentes substances (butyrate, vitamine D3 : voir Wassing et al., 2015) ayant toujours pour rôle principal de renforcer la protection antimicrobienne de la barrière épithéliale. Chez la moule, la MGD2 montre quant-à-elle une surexpression en cas de stress

thermique mais cette surexpression est documentée comme étant réduite en cas d'infection bactérienne (Mitta et al., 2000b).

Ainsi, chez les invertébrés, mais aussi chez les animaux et les plantes, (Thomma & Broekaert 1998), une expression inductible au sein de l'épithelium en contact direct avec l'environnement a pu être décrite. Chez *Bombyx mori* par exemple, les ARNm de la cécropine sont retrouvés au sein des cellules épithéliales et la surexpression de PAMs peut être documentée sous la cuticule lorsque celle-ci est légèrement abrasée et en présence de bactéries (Brey et al., 1993).

Les PAMs, peuvent donc fournir également une réponse locale surtout au niveau des épitheliums et des muqueuses en étant surexprimés en réponse à une infection.

4.2.2. Des formes constitutives

Une production de PAMs à un niveau basal par certains tissus sans infection bactérienne préalable a pu être montrée. Cette expression de PAMs est dite constitutive et se localise surtout au sein des épithéliums à la fois chez les insectes et les mammifères (Selsted and Ouellette, 2005; Tzou et al., 2000). Elle confèrerait alors une protection à la fois contre la flore commensale mais aussi un 'pool' de peptides pré-mobilisés contre les bactéries pathogènes ou opportunistes.

Chez la drosophile, cette expression constitutive concerne peu de tissus et un nombre limité d'agents anti-microbiens : seule l'andropine est exprimée constitutivement au sein de l'appareil reproducteur spécifiquement des drosophiles mâles pour protéger cet organe (Samakovlis et al., 1991) et n'est pas surexprimé en cas d'infection (il s'agit là d'une néo-fonctionnalisation d'un gène codant initialement pour la cécropine du même organisme). Chez les termites *Pseudocanthothermes spiniger*, les termicines et les spinigérines sont constitutivement présentes au sein des hémocytes et des glandes salivaires (Lamberty et al., 2001). Les mytilines chez la moule *Mytillus galloprovincialis* (comme tous les PAMs décrits chez les moules) sont quant-à-elle exprimées constitutivement au sein des hémocytes (Mitta et al., 2000c), ce qui est également vrai dans le cas des crustines (Smith et al., 2008).

4.2.3. L'expression des PAMs chez annélides : une expression modulable

Chez *Hediste diversicolor*, l'hédistine est exclusivement exprimée au sein des cellules G3 qui sont des granulocytes à partir desquels sont synthétisées les molécules de l'immunité humorale (Cuvillier-Hot et al., 2014; Tasiemski et al., 2007). Ce PAM n'est pas documenté comme étant surexprimé suite à une infection microbienne. A la place, les coelomocytes (contenant l'hédistine) s'accumulent autour du site d'infection et, relarguent l'hédistine dans l'environnement extracellulaire (réponse locale). L'expression des neuromacines et theromacines a quant à elle été montré comme étant induite suite à une infection bactérienne (Schikorski et al., 2008; Tasiemski et al., 2004). De plus, l'expression des gènes codant pour la Hm-lombricine (tout comme la neuromacine) au sein du système nerveux central (SNC) de la sangsue est induit en présence de composés microbiens et permet de conclure quant au rôle essentiel de ces PAMs dans l'immunité du SNC (Schikorski et al., 2008). Les gènes codant la lombricine chez *Lumbricus rubellus* ont cependant été décrits comme n'étant pas induits suite à une infection microbienne mais montre une expression constitutive illustrant la difficulté de faire des généralités quant à l'expression des peptides antimicrobiens (ce qui est vrai pour tous les organismes) (Cho et al., 1998).

Les différentes isoformes de l'arénicine ont été montrés comme étant constitutivement exprimées et leurs expressions ne dépendent pas de la stimulation par un agent infectieux. L'expression de l'arénicine-1 a été détectée dans de nombreux compartiments (intestin supérieure et inférieur, glande salivaire…) bien que le signal le plus fort ait été décrit au sein des coelomocytes (Maltseva et al., 2014; Berlov and Maltseva, 2016). Les arénicines sont stockées dans les coelomocytes et relarguées au sein du phagolysosome pour tuer les bactéries phagocytées et leur sécrétion au sein de l'intestin et de l'épithélium (body wall) permet de leur supposer un rôle de première ligne de défense contre les infections pour l'organisme. L'alvinellacine sécrétée par l'annélide polychète *Alvinella pompejana* est induite à partir d'un stock de peptides dans les coelomocytes en réponse à une infection microbienne (Tasiemski et al., 2014). Ce PAM a également été montré comme étant constitutivement exprimé par les cellules épithéliales du tégument ou l'on trouve les micro-organismes épibiotiques (rôle de défense et de contrôle/sélection des épibiotes).

Les données actuelles ne permettent pas de faire des généralités quant à l'inductibilité des gènes codant les PAMs que ce soit chez les annélides ou chez d'autres organismes tant la dynamique est différente en fonction des organismes étudiés mais aussi puisque des PAMs de la même famille peuvent être tantôt inductible tantôt constitutif selon l'espèce considérée.

4.3. Modifications post-traductionnelles des PAMs

La plupart de ces modification incluent des mécanismes de traitement protéolytique des précurseurs protéiques des peptides antimicrobiens (Zasloff, 2002) mais il existe également des modifications post-traductionnelles : amidation de la région C-terminale ou des modifications plus complexes (Taylor et al., 2000) telles que glycosylation, halogénation, isomérisation de certains acides aminés (Bulet et al., 1999; Simmaco et al., 1998). Il a par exemple été montré chez la styéline D que les modifications post traductionnelles permettraient au peptide de préserver son activité anti gram positif à pH faible (pH physiologique) et à forte salinité puisque le peptide synthétique non modifié possède une activité moindre par rapport au peptide naturel qui lui, a subit les modifications post-traductionnelles. L'hédistine d'*Hediste diversicolor*, quant à elle, peut montrer des cas de modifications post-traductionnelles avec la présence de résidus bromotryptophanes (Tasiemski et al., 2007).

Le traitement protéolytique peut amener lui aussi à créer plusieurs types de PAMs à partir du même précurseur. Pour les cathélicidines humaines par exemple, le peptide appelé LL37 issu des neutrophiles présente une activité antimicrobienne à large spectre « classique » après son clivage par une protéase neutrophile (protéinase 3) alors que sur la peau humaine ce même peptide est dégradé en plusieurs dérivés appelés RK-31, KS-30 par des protéases produites par la microflore épidermique conférant à ses nouveaux peptides des activités distinctes du peptide LL37 (López-García et al., 2005).

5. Fonction des PAMs dans le règne animal

5.1. Multifonction des PAMs

II est désormais admis que le rôle des PAMs ne se cantonne pas uniquement à une action antimicrobienne (Figure 8) : ces molécules peuvent avoir d'autres fonctions telles que des activités antivirales et chimiotactiques (Myticine C : Balseiro et al., 2011), des activités anticancéreuses, un rôle dans l'apoptose cellulaire, des activités neurotrophiques (régénération du système nerveux : Schikorski et al., 2008), et dans le recrutement des cellules immunitaires (Choi et al., 2012; Lai and Gallo, 2009; Yeung et al., 2011). Par exemple chez les annélides, en plus d'une activité antibactérienne, la neuromacine et la *Hm*-lumbricine favorisent la régénération du cordon nerveux de la sangsue, ce qui est corrélée au fait que les cellules gliales et les neurones expriment également ces peptides antimicrobiens (Schikorski *et* al. 2008). L'immuno-modulation par les PAMs permet également un contrôle de l'inflammation, favorise la cicatrisation et initie la réponse immunitaire adaptative chez les vertébrés. Il a été par exemple montré que les défensines permettent le recrutement des cellules immunitaires sur le site infectieux (Selsted and Ouellette, 2005) ; les cathélicidines font de même et induisent aussi la prolifération des cellules immunitaires et la production de cytokine (Schauber and Gallo, 2008; Soehnlein, 2009). Les histatines auraient un rôle dans la cicatrisation cellulaire (Oudhoff et al., 2008) et les dermicidines, un rôle dans le chimiotactisme (Harris et al., 2009).

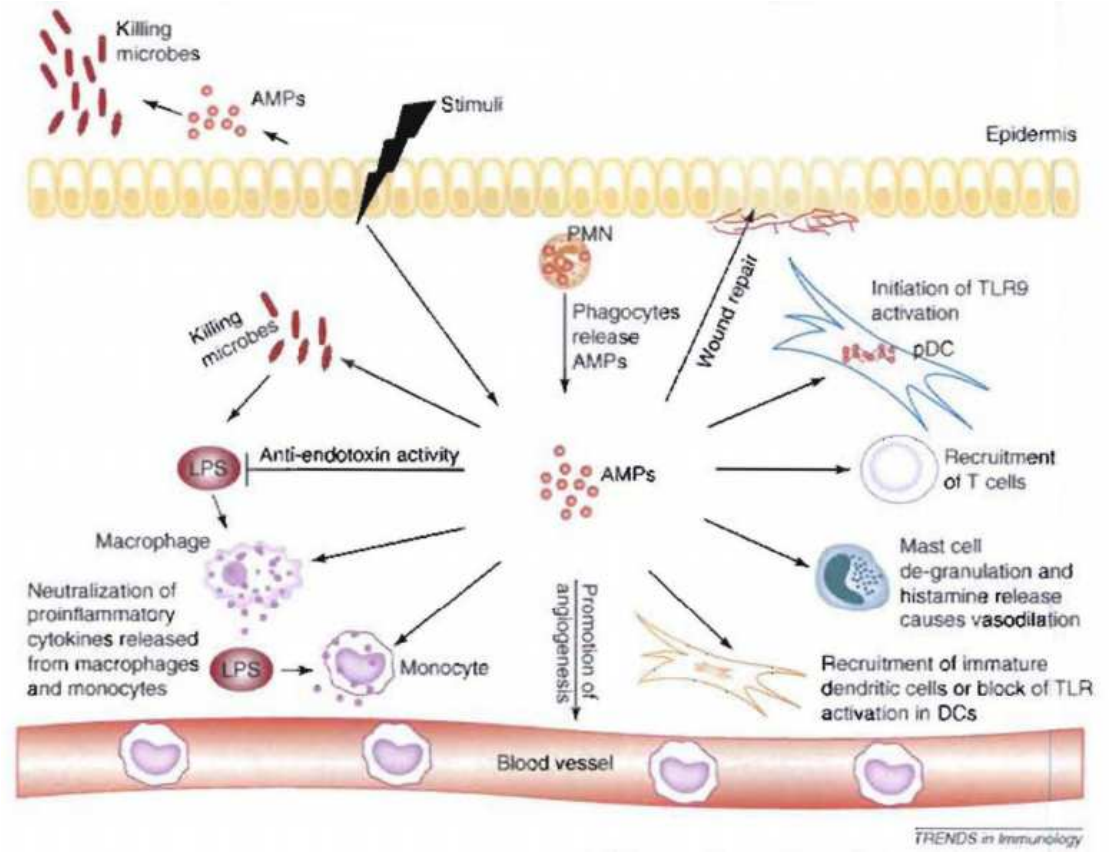


Figure 8. Fonctions multiples des peptides antimicrobiens dans la défense de l'hôte notamment l'immuno-modulation. De Lai and Gallo, 2009.

5.2. Immunité externe des espèces

L'implication des PAMs dans la résistance naturelle aux infections est confortée par leur localisation stratégique au sein des fluides corporels, au niveau des cellules épithéliales et/ou du mucus ou encore au sein des phagocytes (i.e. aux interfaces entre les organismes et leurs environnements) (Selsted and Ouellette, 2005; Zasloff, 2002). Ceux-ci s'expriment entre autres dans les tissus en contact avec l'environnement : les épithéliums et les muqueuses sont donc typiquement le siège d'une grande production de PAMs (Frohm et al., 1997). Chez les batraciens par exemple, les dermaseptines en forte concentration dans le mucus assurent la protection de la peau (Vanhoye et al., 2003) alors que les cathélicidines protègent la peau des mammifères (Dorschner et al., 2001; Nizet et al., 2001). Chez les végétaux, les PAMs peuvent également se retrouver à la surface des feuilles ou dans les exsudats racinaires (Oard and Enright, 2006). C'est ainsi qu'ils assurent la protection de l'hôte en inhibant la multiplication des micro-organismes sur la peau, lesquels pourraient devenir invasifs en cas de lésion de la barrière cutanée.

5.3. Rôle dans le contrôle des symbioses microbiennes

Il a été montré à plusieurs reprises que les PAMs, en plus de leur rôle dans la lutte antimicrobienne, jouent un rôle très important dans la médiation de l'interaction bénéfique sur le long terme entre un hôte et ses microbes commensaux/mutualistes. Les alpha-défensines ont, par exemple, un rôle important dans le contrôle et la sélection du microbiote intestinal chez la souris (Salzman et al., 2010). En effet, des différences de composition du microbiote intestinal ont pu être montrées entre des souris synthétisant normalement le PAM DEFA5 et des souris déficientes en MMP7 (aussi appelée matrylisine) qui est l'enzyme permettant de cliver (et donc d'activer) le peptide antimicrobien en une forme mature. Ces résultats ont permis de conclure quant à un rôle du PAM pour la régulation du microbiote. En plus d'un rôle de régulation de la composition du microbiote digestif, les PAMs peuvent également être impliqués dans le confinement des symbiotes dans des organes dédiés. Deux études récentes (Login et al., 2011; Masson et al., 2016) ont ainsi montré, chez le charançon du maïs *Sitophilus zeamais,* qu'un PAM (coleoptéricine A) est responsable du confinement des endosymbiotes au sein du bactériosome avec un dialogue moléculaire régulé localement. Ce PAM est responsable non seulement du confinement des symbiotes mais régule également leur croissance en inhibant la division cellulaire permettant à l'hôte de contrôler la densité et l'expansion des endosymbiotes en les empêchant de coloniser d'autres tissus (Login and Heddi, 2013). Ces résultats suggèrent donc une co-évolution à long-terme qui module la réponse de l'effecteur immunitaire pour le maintien de la relation hôte-symbionte.

D'autres études montrent que les fonctions des PAMs peuvent permettre de rendre favorables les conditions pour l'instauration d'une relation mutualiste et agissent comme modulateur de la réponse immunitaire (Mylonakis et al., 2016; Nguyen et al., 2011; Tasiemski et al., 2015). Chez l'hydre, les périculines, bien que montrant une forte activité bactéricide qui leur attribue un rôle de défense, sont également responsables de la sélection spécifique de bactéries ayant un rôle bénéfique pour le développement de l'hydre (Bosch et al., 2009; Fraune and Bosch, 2010).

Les PAMs peuvent donc jouer des rôles cruciaux dans l'interaction hôte-symbiotes que ce soit dans le confinement des bactéries mais aussi dans la mise en place ou le maintien de cette interaction.

PARTIE II – Diversification des effecteurs immunitaires : Les processus engagés dans la diversification des agents immunitaires sont-ils les mêmes que ceux qui s'appliquent aux PAMS ?

La diversification des effecteurs immunitaires permet de mettre en place des défenses spécifiques qui peuvent nécessiter, entre autre, le réarrangement de différentes régions de gènes (intra et intergénique) permettant à partir d'un nombre limité de segments géniques de créer un nombre de variants quasi illimité. Ceci serait en effet une réponse aux pathogènes qui présentent une grande diversité d'antigènes et, dans le cadre d'une course à l'armement, ont développé/développent en permanence des mécanismes de contournement de la défense immunitaire. C'est donc cette coévolution qui entraîne un fort niveau de diversification et/ou de polymorphisme des molécules de reconnaissance chez l'hôte vertébré.

Dans partie, nous nous intéresserons à différents mécanismes générant de la diversité des molécules de l'immunité innée chez invertébrés et les vertébrés en effectuant quelques comparaisons avec les mécanismes de diversification du système de l'immunité adaptative qui, lui a fait l'objet de beaucoup plus d'études sur le sujet. Un focus plus particulier sur les mécanismes de diversification des PAMs, sujets de cette thèse, fait l'objet de la Partie III.

1. Quelques mécanismes de diversification moléculaire chez les invertébrés

Pendant longtemps l'immunité des invertébrés était supposée être limitée à un système inné sans reconnaissance spécifique des microorganismes. Ainsi, cette immunité était fondée uniquement sur des récepteurs analogues au type Toll et à un nombre restreint d'autres récepteurs de reconnaissance des PAMPs.

Chez les invertébrés, les mécanismes permettant la diversification de différentes molécules de reconnaissance (PRR) ou des effecteurs immunitaires ont été partiellement décrits (Tableau 1).

	Groupe ou espèce	Diversification génétique	Eléments reconnus	Références
FREPs	Mollusques et arthropodes	Famille multigénique (14 familles de protéines contenant jusqu'à 8 locus différents), épissage alternatif, recombinaison somatique	anti-trématodes, levure, gram négatif, gram positif	Zhang & Loker, 2003; Zhang et al., 2004
VCBPs	Amphoxius: *Branchiostoma floridae*	Famille multigénique, Copy Number Variation, recombinaison et conversion génique, fort allélisme du domaine Immunoglobuline	Reconnaissance chitine	Dishaw et al., 2010
DsCAM	Arthropodes (insectes et crustacés)	Duplication d'exons et épissage alternatif somatique	Anti-bactérien	Ghosh et al., 2011 pour review
PGRP	Répandu mais décrit chez *Drosophila melanogaster*	Epissage alternatif, famille multigénique	gram négatif, Anti-gram positif	Lemaitre & Hoffman, 2007; Werner et al., 2000

Tableau 1. Mécanismes de diversification de plusieurs effecteurs de l'immunité et PRR chez les invertébrés. FREPs: Fibrinogen related proteins, DsCAM: Down syndrom cell adhesion molecule, VCBPs: Variable domain-containing chitin binding.

Les invertébrés peuvent présentent une forte diversité génique dans un répertoire de gènes codant des protéines à domaine Ig.

1.1. Le cas des FREPs : immunité anticipative chez les invertébrés

Des études réalisées sur le gastéropode *Biomphalaria glabrata,* qui est l'hôte intermédiaire du trématode parasite *Schistosoma mansoni,* ont permis de renseigner les bases moléculaires de la reconnaissance immunitaire de l'hôte avec son pathogène (Adema et al., 1997). Il a été montré que ces protéines sont capables de reconnaître une large gamme de molécules provenant de divers pathogènes aussi bien procaryotes qu'eucaryotes. Le précurseur protéique de ces protéines possède un peptide signal, une région N terminale composée de domaines de type immunoglobuline (Ig1 et Ig2, rappelant ceux trouvés chez

les vertébrés) ainsi qu'une région C-terminale composée d'un domaine fibrinogène (FReD ou FBG) (voir Ghosh et al., 2011 pour synthèse).

Chez *B. glabrata* plus longuement étudiée, actuellement 14 sous-familles du gène FREP (FREP1 à FREP14) ont été identifiées. Elles diffèrent par leur structure exons-introns et la présence d'un ou deux domaines Ig arrangés en tandem en amont du domaine FBG. Chaque sous-famille est constituée de 1 à 8 locus (Zhang and Loker, 2004). Les séquences complètes de 4 sous-familles de FREPs ont été décrites et montrent qu'il existe une forte variabilité au sein des domaines codant le domaine Ig et celui codant le domaine FBG (Zhang et al., 2004). De plus, il a été également montré que 45 séquences nucléotidiques différentes ont pu être retrouvées au sein d'un seul individu pour l'exon 2 du gène FREP3 et 37 séquences pour un autre individu dont une seule séquence est commune aux deux individus. Ce nombre impressionnant de séquences n'est pas prédit par le nombre de locus identifiés (entre 3 et 5 par la même étude). Les auteurs ont donc suggéré que le fort niveau de diversification des isoformes s'explique par 2 mécanismes : (1) l'accumulation de mutations somatiques à partir de séquences dites « séquences alléliques initiales » dont le nombre correspondrait à 2 fois le nombre de locus correspondant à la sous-famille concernée (ici FREP3), et (2) un processus de recombinaison entre les différentes copies du gène (Zhang and Loker, 2003). Ces auteurs ont en effet permis de décrire un phénomène d'épissage alternatif permettant de promouvoir la diversité chez ces protéines en générant des transcrits tronqués des domaines Ig et FBG. La diversité des FREP chez de nombreuses espèces d'invertébrés et son rôle fonctionnel fait l'objet d'un nombre accru d'études comme le montrent plusieurs publications récentes (Dai et al., 2017; Hou et al., 2016). Les protéines de la famille des FREPs représentent donc une famille hyper-diversifiée qui serait capable de se lier à des antigènes de pathogènes variés et dont le répertoire est potentiellement illimité. De nombreuses études continuent de s'intéresser aux différents mécanismes qui peuvent générer une spécificité de réponse chez les effecteurs de l'immunité des invertébrés considérée en premier lieu comme non spécifique (Ghosh et al., 2011; Huang et al., 2015; Pees et al., 2015; Schulenburg et al., 2007). Cette diversification permet ainsi aux invertébrés de réagir plus rapidement aux changements antigéniques des pathogènes.

1.2. Le complexe majeur d'histo-incompatibilité (CMH) chez les vertébrés

Le CMH des vertébrés est constitué d'un ensemble de gènes distribués le long d'un fragment d'ADN qui s'étend sur environ 4 millions de paires de bases chez l'homme et contiendrait plus de 260 gènes (Kelley et al., 2005), il est donc polygénique mais présente également plusieurs allèles à chaque gène (De Bakker et al., 2006; Horton et al., 2004). Les gènes CMH sont parmi les gènes étudiés les plus polymorphes actuellement connus (Trowsdale and Parham, 2004) Pour certains locus, au niveau populationnel, presque 200 allèles ont en effet été répertoriés (Vogel et al., 1999). De plus la création de nouveaux allèles par recombinaison est également documentée pour ce système mais constitue un processus s'appliquant à tous les gènes eukaryotes (Belich et al., 1992; Watkins et al., 1992; Zhao et al., 2013). Deux types de recombinaison sont décrits : (1) le brassage d'exons qui est caractérisé par la recombinaison de régions exoniques entières avec des 'hotspots' de recombinaison dans les régions introniques et, (2) des évènements de recombinaison entre des séquences plus courtes intra-exonique ont également pu être décrits. Par exemple, Hughes et al. (1993) ont mis en évidence des évènements de recombinaison incluant de petites portions d'exons dans la région responsable de la reconnaissance des antigènes (ARS). De plus, il apparaît que ces sites sont sous sélection balancée montrant que ces évènements de recombinaison favorisent l'exploration du potentiel adaptatif de l'espèce vis-à-vis des mécanismes de défense antimicrobien (Hughes et al., 1993).

2. Histoire évolutive des gènes de l'immunité

Chez les invertébrés, des études menées à l'échelle du génome chez la drosophile par Sackton *et* al*.* (2007) et Obbard *et* al. (2009) ont d'abord permis de montrer que seule une petite portion du génome était soumise à une forte pression de la sélection positive et que celle-ci se regroupait dans un petit nombre de voies de signalisation (IMD pathway) qu'ils décrivent comme des « hot spots » de la coévolution. Ces voies de signalisation sont rapportées comme étant la cible du parasite pour supprimer la réponse immunitaire de l'hôte. Ainsi alors qu'intuitivement des molécules telles que les PRRs (PGRP, DsCam…) ou même les PAMs (en contact direct avec les membranes des bactéries pathogènes) pourraient être plutôt soumises à des pressions de sélection positive pour la reconnaissance et/ou le ciblage spécifique de bactéries, à l'inverse, les molécules impliquées dans les voies

de signalisation déclenchées par les PRRs, moins en contact avec le pathogène, ne devraient pas être à même de coévoluer avec le pathogène. Lazzaro (2008) et Obbard et al. (2009) ont mis en évidence, chez la drosophile, que contrairement à ces hypothèses de départ, les gènes codant pour les PRRs (par exemple les PGRPs) ne présenteraint pas d'évolution adaptative spécifique entre les espèces *D. melanogaster* et *D. simulans* (une exception a été trouvée sur deux acides aminés sous sélection positive dans la région codant le domaine PGRP-LCa). Les molécules impliquées dans les voies de signalisation déclenchées (facteurs de transcription de type NF-κB - Relish voir Lazzaro 2008 pour review) sont, elles, sous l'action de la sélection positive.

L'observation selon laquelle les PRRs n'évoluent pas sous sélection positive a pu également être retrouvée chez d'autres invertébrés tels que les moustiques du genre *Anopheles* ou encore les termites *Nasutitermes* (Bulmer and Crozier, 2006; Little and Cobbe, 2005) bien qu'il est difficile de détecter l'action de la sélection positive lorsque la séquence codante est courte (cas des PAM). Ainsi, une absence d'évolution adaptative des gènes codants les PAM a, dans le même sens d'abord été décrite par Clark and Wang, (1997), mais de récentes études mettant en lumière des mécanismes d'évolution non neutre ont donc contredit ces premiers travaux (cf. partie 3) laissant supposer que l'action de la sélection sur ces effecteurs supposés généralistes reste à élucider (Unckless and Lazzaro, 2016).

PARTIE III : Diversité génétique et histoire évolutive des peptides antimicrobiens

1. Architecture génique du précurseur protéique

Les PAMs constituent une famille de molécules ayant une origine ancienne polygénique. De ce fait, ils présentent une grande variété de structures tant au niveau du gène qu'au niveau de la protéine et une large gamme d'histoires évolutives. Pour rappel, les PAMs sont clivés à partir d'un précurseur protéique plus long (appelé prépropeptide) qui contient un peptide signal en région N-terminale qui est responsable de l'adressage cellulaire vers le réticulum endoplasmique (Zhang and Gallo, 2016). Les Figures 9 et 10 illustrent la diversité de structure des gènes codant les prépropeptides de différentes familles de PAM intra et inter-espèces.

La structure des gènes codant les différentes familles de PAMs n'est pas, ou que très rarement, conservé (Diamond et al., 2009). Chez l'homme par exemple, alors que le gène codant le précurseur protéique des alpha-défensines est constitué de trois exons et deux introns, celui codant pour le précurseur protéique des beta-défensines est constitué de deux exons et d'un intron. De plus alors que pour les alpha et beta-défensines, la prorégion et le PAM sont codés chacun par un exon différent, pour les cathélicidines le pro-domaine est codé par trois exons et le PAM LL-37 par le dernier exon (Diamond et al., 2009).

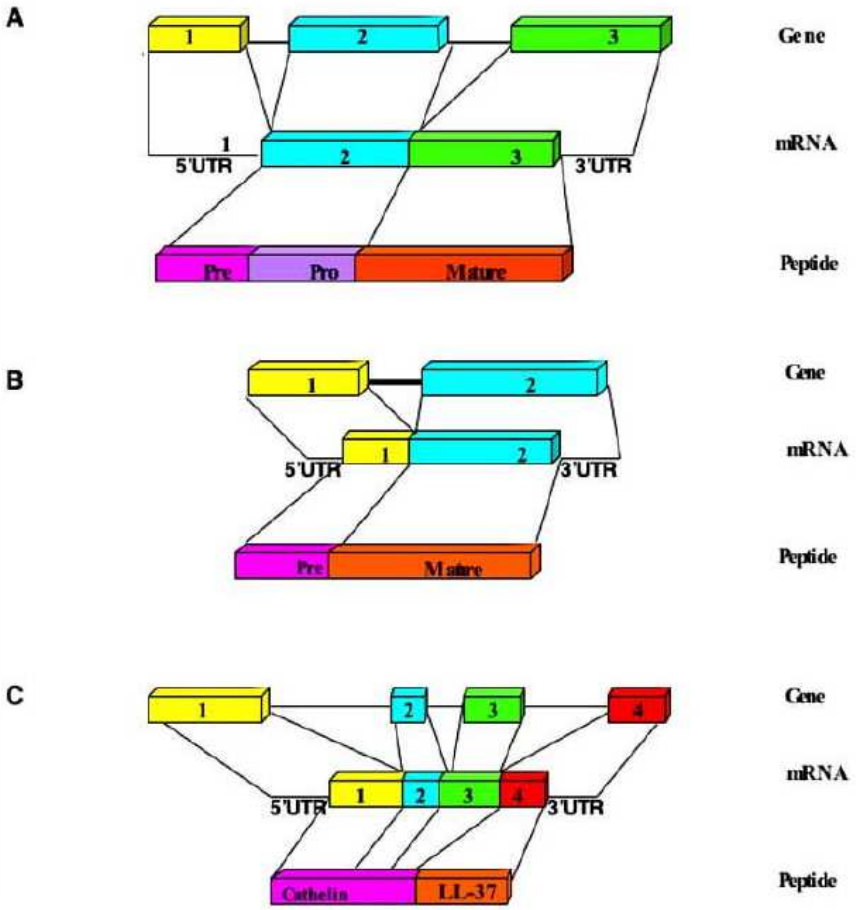


Figure 9. Structure des gènes (avec différents exons en couleurs) codant pour le précurseur protéique des alpha-défensines humaines (A), des beta-défensines (B) et des cathélicines (C) (d'après Diamond et al., 2009). En première ligne, la position des exons, puis la structure de l'ARN messager et finalement la structure du prépropeptide.

En plus de ne pas montrer de conservation en termes de positionnement des régions introniques/exoniques (triangles gris dans la Figure 10), des architectures géniques des prépropeptides sont variées d'une espèce à l'autre pour un gène donné. Par exemple chez les défensines de mollusques et d'arthropodes, alors que chez *Mytilus galloprovinciallis* la défensine MGD2 est constituée dans l'ordre d'un peptide signal suivi du PAM et d'une région semblable à la prorégion, chez le moustique, le précurseur protéique de la défensine A est constitué d'un peptide signal, d'une prorégion et enfin du PAM. La structure génique des défensines est donc très différente entre les mollusques et les arthropodes. Cependant, comme tous les introns de défensines sont de phase I et que ces introns ne sont jamais trouvés entre les exons codant pour le motif CS alpha-beta (motif responsable de la structure 3D nécessaire à l'activité des défensines), ces différences dans le positionnement des différents domaines pourrait être simplement la conséquence d'un brassage d'exons

(« exon shuffling »), les domaines pris séparément étant génétiquement apparentés entre mollusques et arthropodes (Du Pasquier, 2009).

Finalement, même au niveau intra-spécifique (cf. défensines (v) et (vi) de l'espèce *Stomoxys calcifrans* : Figure 10), cette architecture peut être variable entre deux PAMs de la même famille avec des tailles d'exon différents (Froy and Gurevitz, 2003).

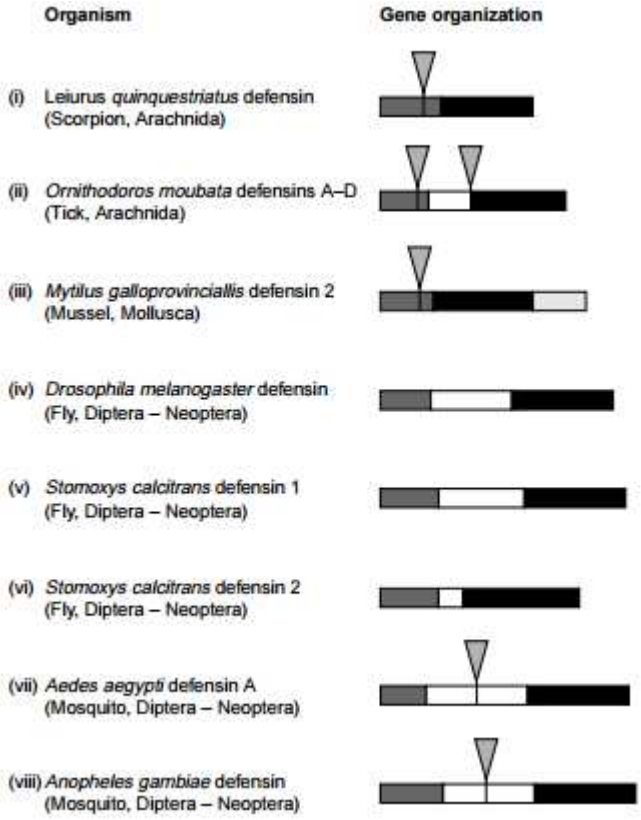


Figure 10. Comparaison de la structure du gène codant les défensines entre espèces. Gris= peptide signal ; Noir= PAM ; Blanc= prorégion. Triangle = position des introns.

Malgré cette diversité de structure génique, il existe quand même des cas où la structure génique est conservée entre espèces. C'est le cas des cécropines chez qui le gène présente une structure conservée avec deux exons séparés par un intron avec une taille variable de l'intron entre les espèces (Tassanakajon et al., 2015).

Cependant, en dépit d'une structure/organisation des gènes codant le précurseur protéique du PAM (longueur, positionnement des introns, organisation des exons) assez variée entre PAMs et entre espèces pour un PAM donné, il est remarquable de noter que le prépropeptide possède en général un peptide signal, une prorégion et la région du PAM mature.

2. Diversification par duplications géniques : un moteur dans l'évolution des PAMs

2.1. Les duplications géniques : définition

Les duplications géniques sont observées au sein des trois règnes du vivant (Lenormand et al., 1998) et désignent le processus par lequel un gène, un fragment de génome, voire un génome entier (polyploïdisation) va passer d'une à deux copies initialement identiques. Après fixation dans l'espèce (l'événement de duplication apparaît en général au sein d'un individu et se propage ensuite dans la population), ces gènes dupliqués (paralogues) vont pouvoir diverger l'un de l'autre au cours du temps tout en gardant des possibilités d'échanges par recombinaison et/ou conversion génique dans la première phase de séparation (Fawcett and Innan, 2011). La théorie d'évolution des gènes dupliqués (Lynch and Conery, 2000; Ohno, 1970) suggère trois alternatives puisqu'une seule des deux copies est suffisante pour remplir la fonction ancestrale. L'une des 2 copies restant sous sélection purifiante pour assurer la fonction initiale du gène, l'autre peut alors évoluer plus ou moins librement par relâchement des pressions de sélection.

(1) Une copie évolue vers la perte de sa fonction et accumule des mutations de façon neutre, par dérive génétique, pour aboutir à la formation d'un pseudo-gène (pseudogénisation) avec, par exemple, l'apparition d'un codon stop ou d'une délétion partielle du gène.

(2) Une copie acquiert une fonction nouvelle et avantageuse (néo-fonctionnalisation) par accumulation de nouvelles mutations non synonymes (certaines mutations rares avantageuses peuvent être retenues sous l'action de la sélection naturelle/positive) par rapport à l'autre copie qui elle assure la fonction initiale. Dans ce cas, on peut également envisager que ces mutations affectent les régions de régulation de l'expression du gène et puissent conduire à des niveaux d'expression du gène qui diffèrent entre paralogues selon les conditions du milieu, le type tissulaire ou le stade de développement.

(3) Les fonctions des deux copies sont altérées partiellement par l'accumulation de mutations légèrement délétères mais la fonction globale des 2 protéines dérivées reste semblable à celle de la protéine initiale ou adaptée à un type cellulaire donné (sous-fonctionnalisation). Ce modèle de sous-fonctionnalisation (modèle DDC: Duplication-Dégénérescence-Complémentation) a été proposé au début des années 2000 par Force et

collaborateurs pour expliquer pourquoi tant de gènes dupliqués persistent dans le génome. Ce modèle est basé sur le fait que de nombreux gènes ont des expressions spécifiques dans différents types cellulaires ou à des moments clés du développement de l'organisme. En résumé, dans ce modèle, chacune des copies ont accumulé de façon indépendante des mutations (légèrement délétères) de façon à ce que les deux copies soient nécessaires pour obtenir une fonction équivalente (complémentation) à celle du gène d'origine.

Les événements de duplication se classent aussi en fonction du positionnement des copies du gène dans le génome. Il existe 2 modes de translocation des gènes :

i) la duplication en tandem - les copies du gène se suivent et se «corrigent» par conversion génique ou 'crossing over' inégal entre elles. C'est ce qu'on appelle le mécanisme de l'évolution concertée qui a été très bien décrit par (Coen et al., 1982a, 1982b) sur les sous-unités ribosomiques 18S et 28S et qui permet d'éliminer les différences inter-copies du gène dupliqué et ;

ii) la duplication inter-chromosomique dans laquelle les gènes dupliqués se retrouvent dans des régions distinctes du génome – les copies du gène ne sont pas 'corrigés' par recombinaison ou conversion génique et une divergence entre duplicats apparaît progressivement résultant d'une évolution séparée des gènes nouvellement dupliqués (Du Pasquier, 2006).

2.2. Les PAMs : une évolution par duplication

Les gènes codant les PAMs, tout comme ceux codant les effecteurs du système immunitaire, constituent souvent des familles multigéniques, la plupart du temps dupliqués en tandem sur une portion du génome (Bulmer and Crozier, 2004; Lynn et al., 2004a; Maxwell et al., 2003; Schutte et al., 2002; Semple et al., 2003).

Par exemple, Tennessen and Blouin, 2007 ont décrit chez *R. chiricahuensis* que la ranatuerine est codée par au moins deux gènes issus d'un évènement de duplication et par trois gènes dupliqués chez *R. pipiens*. Les gènes codant les alpha-défensines humaines HNP1 & 4 et HD5 & 6, ainsi que les beta-défensines HNP1 & 2 sont toutes localisées sur la même portion de chromosome (8p23) ce qui est également décrit chez les cathélicidines regroupées quant à elles au sein du même chromosome (3p21.3) (Yang et al., 2004). Les

attacines chez la drosophile sont codées, elles aussi, par une famille de gènes dupliqués en tandem (attacines A et B séparées par 1.1kb) et sont soumis à la conversion génique (96% d'identité en termes d'acides aminés). A l'inverse, l'attacine C, située à 1.3 Mb des attacines A et B est beaucoup plus divergente (67% d'identité en acides aminés) et ne présente aucune séquence portant la marque d'une conversion génique (Lazzaro and Clark, 2001). Ces auteurs ont également noté l'existence dans le génome d'un pseudo-gène (gène non fonctionnel) révélant une histoire évolutive complexe et ancienne pour cette famille multigénique.

Ces évènements de duplication en tandem peuvent à la fois conduire à une accumulation d'isoformes dans le génome pour le même PAM avec présence d'évolution concertée entre les isoformes (49 beta-défensines chez la souris : Schutte et al., 2002) mais aussi générer 2 copies récentes du peptide antimicrobien comme c'est le cas pour l'androcine et la cécropine chez la drosophile. En effet, ces deux peptides, en plus d'être physiquement proches au sein du génome (1kb de distance), possèdent la même structure de gène (en nombre et positions des introns) et des structures secondaires similaires mais ne possèdent pas les mêmes activités antibactériennes et, par-dessus tout, l'andropine est un peptide antimicrobien qui est exprimé uniquement au sein de l'appareil reproducteur mâle (Samakovlis et al., 1991). Ces observations ont donc permises aux auteurs de conclure quant à un évènement de duplication ancestral avec néo-fonctionalisation.

2.3. Une diversité génétique des PAMs complexifiée par la recombinaison et des mécanismes d'épissage alternatif

2.3.1 Recombinaison dans les familles multigéniques de PAMs

Comme vu précédemment, il existe 2 modes de translocation des gènes dont la duplication en tandem pour laquelle les copies du gène se suivent et se «corrigent» entre elles par conversion génique ou 'crossing over' inégal.

La conversion génique, définie comme un transfert non-réciproque d'information génétique entre deux gènes possédant un important degré de similitude (Radding, 1978) permet également la duplication de petites portions d'ADN entre ces copies. Ce processus, proche de la recombinaison, est documenté comme étant un élément important dans la dynamique des familles multigéniques. Cette homogénéisation de quelques paires/centaines de paires de bases (mais également les 'crossing over' inégaux : homogénisation de plus grandes

régions) représentent les deux mécanismes principaux qui génèrent l'évolution concertée de deux gènes dupliqués (Elder Jr and Turner, 1995) et peuvent être à la base de nouveaux allèles (avec –ou non– maintien de ceux-ci par sélection balancée). Ainsi, les paralogues peuvent encore « communiquer » entre eux via des évènements de recombinaison surtout lorsque ceux-ci sont proches dans le génome.

Chez les peptides antimicrobiens, dans le cas des gènes codant les défensines Cg-def et Cg-Prp (Proline-rich peptide) chez l'huître *Crassostrea gigas*, des évènements de recombinaison entre gènes dupliqués ont été décrits comme un mécanisme permettant de créer de la diversité génétique et de tester de nouveaux variants vis-à-vis de la sélection naturelle (Schmitt et al., 2010). Des évènements de recombinaison entre gènes dupliqués codant pour la myticine-C (au moins deux gènes) ont également pu être observés mais les auteurs précisent qu'il s'agirait d'évènements somatiques puisque ces évènements ne sont retrouvés que chez les ADNc et absents des données génomiques (gDNA) (Vera et al., 2011). Ces résultats sont de plus contradictoires avec une précédente étude (Padhi and Verghese, 2008) qui sur la même espèce et le même peptide antimicrobien n'avait pas pu mettre en évidence d'évènements de recombinaison.

Chez *M. galloprovinciallis* et *M. edulis*, des évènements de recombinaison avérés ont pu être mis en évidence chez deux peptides antimicrobiens (Mytilin B et défensine MGD2) faisant suite à un contact secondaire entre les deux espèces (Boon et al., 2009). Pour la Mytilin B, des recombinants intra- et inter-clades ont pu être décrits dont un recombinant inter-clade au niveau de l'exon codant le PAM mature entraînant l'existence d'un nouveau peptide antimicrobien. La MGD2 quant à elle montre un taux de recombinaison et aucune différenciation génétique entre populations Nord Atlantique et de la Péninsule Ibérique (cette absence de différenciation est également retrouvée pour la Mytiline B entre les populations Nord Atlantique où les deux espèces sont en mélange). La recombinaison dans ce cas est supposée brouiller le signal de la différenciation (surtout en région 3' du gène codant pour le peptide antimicrobien) puisqu'une légère structure génétique peut être détectée dans la région 5' du gène.

2.3.2. Mécanime post-transcriptionnel de diversification : l'épissage alternatif

L'épissage alternatif en éliminant ou non certains exons de l'ARN mature représente également un autre moyen supplémentaire de créer de la diversité fonctionnelle chez les PAMs. A partir d'un seul gène, l'épissage alternatif permettra de produire plusieurs ARNm qui pourront ensuite être traduits en autant de protéines avec des propriétés/fonctions potentiellement distinctes. Ce processus est donc largement contributeur de la diversité du protéome et il fournit une explication au différentiel entre le nombre de gènes et le nombre de protéines traduites.

Ce mécanisme a été décrit comme étant un des mécanismes permettant de générer une grande spécificité/diversité fonctionnelle dans les molécules de reconnaissance des pathogènes ou directement sur les effecteurs de la réponse innée des invertébrés (Watson et al., 2005; Zhang and Loker, 2003). Chez *D. melanogaster* par exemple, l'épissage alternatif des DsCAM permet de générer près de 30 000 isoformes à partir d'un nombre réduit de gènes. Dans le cas des PAMs, ce mécanisme a pu être mis en évidence sur les penaeidines chez deux espèces de crevettes *Litopenaeus vannamei* et *L. setiferus* pour lesquelles la variabilité des isoformes serait due à des mécanismes transcriptionnels. Par exemple chez *L. vannamei* trois transcrits appartenant à l'isoforme Lv3a (peneidine de classe 3) ont été retrouvés : ceux-ci sont exactement identiques dans la région codante du précurseur protéique du PAM mais possèdent une région 3' UTR spécifique non homologue (Cuthbertson et al., 2002). Ces auteurs émettent donc l'hypothèse d'un épissage alternatif sur la région 3'UTR. Chez le nématode *C. elegans*, des mécanismes d'épissage alternatif sont également observés pour expliquer la présence de plusieurs isoformes d'un PAM (Kato et al., 2002). En effet, il apparaît dans ce cas que des transcrits tronqués sont produits pour les PAMs *abf-1* et *abf-2* (en plus des transcrits « normaux » à la taille attendue). Les auteurs suggèrent cependant que ces transcrits plus courts pourraient être le produit de la transcription de gène puisque des motifs TATA-box et CCAAT ont été identifiés en amont de la séquence des transcrits tronqués.

Une autre étude a également permis de mettre en évidence qu'un PAM de la famille des cathélicidine (appelé Bac4) est généré à partir de l'épissage alternatif de deux gènes contigus (Bac 7 et « Bac 4») par excision d'exons des deux gènes contigus chez le bovin (*Bos taurus*). Mais alors que le PAM Bac 7 (issu de l'expression du gène Bac7) dispose d'une

activité antimicrobienne, le potentiel PAM issu du gène appelé « Bac 4» n'en possède pas (il est non fonctionnel). C'est en fait l'action de l'épissage alternatif (épissage du dernier exon de Bac7 et premier exon du gène appelé Bac4) qui permet au PAM appelé Bac 4 (qui n'est donc pas issu de la transcription et traduction usuelle du gène Bac4 mais de celui de la forme épissée Bac7-Bac4) d'être *in fine* fonctionnel (Scocchi et al., 1998).

Malgré ces quelques exemples, le mécanisme d'épissage alternatif, bien décrit pour d'autres molécules de l'immunité des invertébrés, est un mécanisme plutôt rare pour créer de la diversité fonctionnelle au niveau des peptides antimicrobiens que ce soit chez les vertébrés ou les invertébrés.

3. Diversité génétique des peptides antimicrobiens

Le niveau de polymorphisme des PAMs à l'échelle du locus peut varier énormément entre des espèces différentes, au sein d'une même famille de gènes pour une espèce donnée, et même, entre régions du précurseur protéique. Pour reprendre l'exemple des défensines chez l'huître *C. gigas* (Cf-defhs et Cgdefm), les gènes codant ces isoformes dupliquées diffèrent de manière importante le long du gène au niveau de leur polymorphisme respectif (Schmitt et al., 2010). Cette étude comparative a en effet permis de mettre en évidence que chez ces prépropeptides (diversité nucléotidique global de Cg-defhs=0,022 : Cg-defm=0,090), un polymorphisme «élevé » est similaire entre la région du peptide signal (relativement conservée entre les différents locus) et celle du peptide mature alors que le PAM Cg-defh1 de cette même espèce est monomorphe aux mêmes régions. De plus, la diversité nucléotidique du PAM Cg-prp est largement inférieure dans la région codant pour le domaine cationique (considéré comme étant le domaine actif) chez cette même espèce par opposition aux deux autres régions du prépropeptide (région anionique et peptide signal, diversité=0,044 ; Schmitt et al., 2010). Chez *M. galloprovinciallis*, un fort niveau de polymorphisme est également observé dans le cas de la myticine C avec des valeurs de diversité nucléotidique de 0,046 pour tout le précurseur contre 0,022 pour le peptide signal, et 0,058 pour le peptide mature (Padhi and Verghese, 2008).

Partant du constat que de nombreuses familles de PAM montrent des niveaux de polymorphisme élevés au niveau de la région codant pour le PAM, certaines études ont cherché à savoir si cette forte diversité nucléotidique s'accompagnait ou non, d'une forte

diversité de structure primaire du PAM. Même si cette diversité nucléotidique n'a été étudiée que dans un nombre réduit de cas, les descriptions de la variation de la structure en acide aminés des isoformes de PAMs sont nombreuses.

Par exemple, chez les défensines, les séquences en acides aminés des Cg-defms et Cg-defhs sont variées (Figure 11).

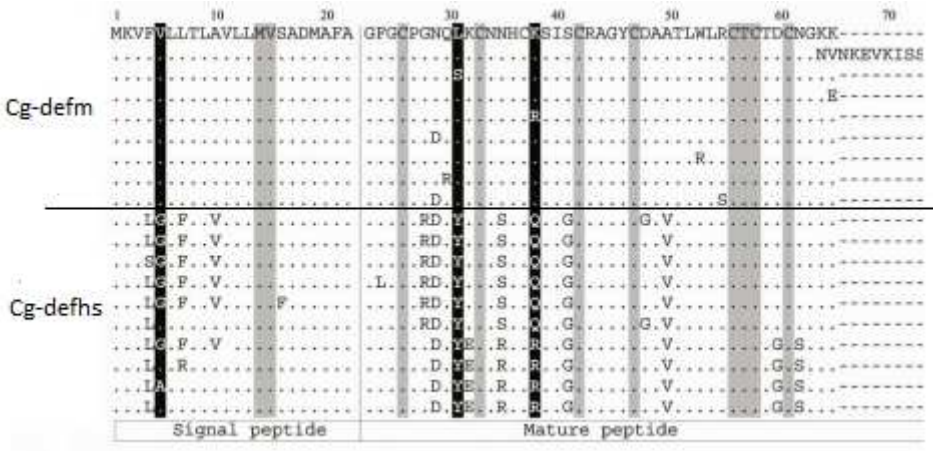


Figure 11. Alignement des séquences peptidiques des isoformes de Cg-defh et Cg_defm chez *Crassostra gigas* pour le peptide signal et le peptide mature. Un point indiquent que l'acide aminé d'une séquence est similaire à la séquence consensus (première séquence) et les tirets à l'absence de l'acide aminé. De Schmitt et al. (2009).

L'alignement de la Figure 11 illustre le fait qu'un polymorphisme non synonyme est retrouvé entre paralogues de ces 2 PAMs et, également, que la diversité nucléotidique décrite précédemment correspond à une diversité importante d'isoformes peptidiques, même pour le peptide antimicrobien dont la diversité nucléotidique est la plus faible (Cg-defm).

Chez les invertébrés, comme expliqué précédemment, les orthologues d'une même famille de peptides antimicrobiens sont rarement décrits entre espèces, illustrant la difficulté de connaître le niveau de divergence moyen des PAMs entre 2 espèces proches. Cependant, lorsque cela est possible, on remarque que la divergence des séquences entre espèces est particulièrement élevée au sein de chaque famille de gènes comme illustré dans la Figure 12.

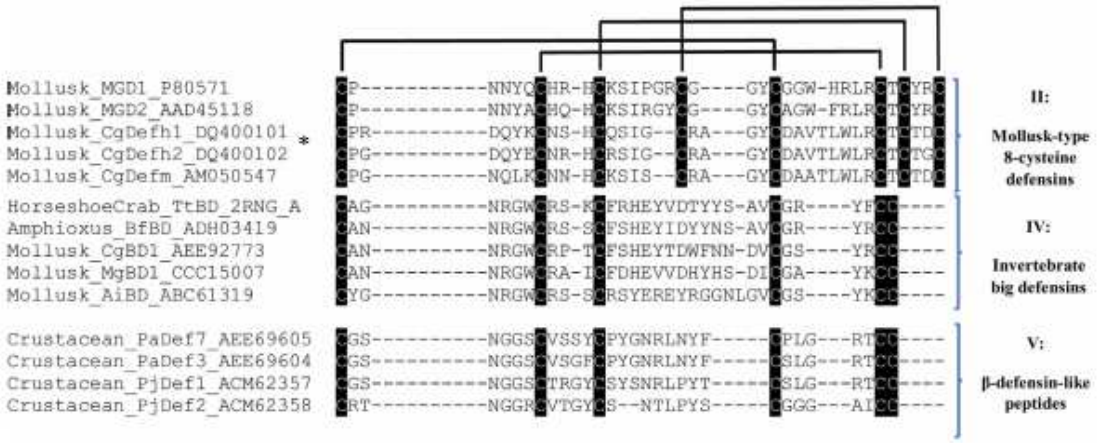


Figure 12. Alignement des séquences peptidiques de plusieurs orthologues de PAM d'invertébrés. Les traits noirs correspondent aux ponts disulfures. De Tassanakajon et al. (2015).

Par exemple, les défensines décrites chez *C. gigas* sont nettement divergentes de celles retrouvées chez *M. galloprovinciallis* bien que la structure 3D soit conservée (conservation des cystéines impliquées dans les ponts disulfures). A l'inverse, bien qu'une forte diversité des séquences primaires soit la règle (même entre espèces proches), il existe quand même des cas où les séquences en acides aminés codant les PAMs soient conservées entre espèces. Par exemple, Rahnamaeian et al. (2015) ont décrit une seule et même séquence de l'hymenoptaecine entre deux espèces de bourdon, *Bombus pascuorum* et *Bombus terrestris* et, dans le cas de l'abaecine, des PAMs qui ne divergent que d'un seul acide aminé chez ces mêmes espèces. Il est cependant nécessaire de conditionner cette absence de divergence au temps de divergence faisant suite à l'événement de spéciation ayant donné lieu à ces 2 espèces.

Différentes études se sont intéressées aux mécanismes de sélection qui agissent sur le polymorphisme de ces effecteurs de l'immunité (son maintien ou non dans les différentes régions du précurseur protéique) mais aussi de déterminer si l'action de la sélection peut conduire à promouvoir la divergence des orthologues d'un même gène après spéciation ou entre les paralogues d'une même espèce. La suite de cette introduction générale s'intéressera donc, dans un premier temps, à définir les différents modes de sélection qui peuvent agir sur le polymorphisme d'un gène et les attendus de ces modes en termes de diversités non synonyme et synonyme. Nous nous intéresserons ensuite aux études de cas ou différents mécanismes sélectifs ont pu être incriminés pour expliquer la diversification allélique/divergence des orthologues/paralogues de ces effecteurs immunitaires.

4. Patrons de sélection chez les gènes codant les PAMs

4.1. Modes de sélection : définitions

Il existe 2 types de sélection opérant sur les produits des gènes : la sélection négative, purifiante qui élimine des gènes les mutations délétères ou légèrement délétères et la sélection positive qui favorise un allèle ou un génotype dans un environnement particulier.

La sélection purifiante (ou négative) diminue la diversité génétique neutre attendue à un gène mais de façon diffuse et constante au cours du temps puisque ce processus tend à éliminer les mutations délétères (donc non-synonymes) sur les parties codantes du gène. Cette sélection s'applique sur tous les gènes dès lors que leur produit est contraint par sa structure et sa fonction, et aura pour conséquence de diminuer seulement la diversité génétique non-synonyme (mutation entraînant un changement de l'acide aminé).

En termes de sélection positive, il existe plusieurs modalités de cette force selon son effet sur la diversité génétique de l'espèce au gène visé (i.e. balayage sélectif vs sélection balancée).

Sélection directionnelle ou balayage sélectif

Cette sélection favorise la montée en fréquence d'un allèle particulier dans les populations d'une espèce selon l'environnement (biotique et/ou abiotique) rencontré. Dans ce cas, elle se traduit par la fixation d'un allèle avantageux dans la population ce qui diminue la diversité génétique de l'espèce au gène impacté avec l'augmentation en fréquence de l'allèle sélectionné dans un laps de temps court : c'est le balayage sélectif (sélection dirigée). De nouvelles mutations vont ensuite apparaître dans le polymorphisme après fixation en fonction du temps écoulé depuis le balayage puisque ce processus a réduit/éliminé presque toute la variation génétique aux sites liés à/aux (la) mutation(s) sous sélection positive.

Sélection balancée

La sélection balancée pourra avoir un effet sur le génotype hétérozygote lorsque les individus possédant ce génotype ont une meilleure valeur sélective que les individus homozygotes : c'est la surdominance. Dans le cas des gènes de l'immunité, le type de

sélection balancée le plus répandu reste néanmoins la sélection fréquence-dépendante lorsque l'aptitude (fitness) d'un génotype dépend de sa fréquence dans la population (Takahata and Nei, 1990). Celle-ci peut favoriser les allèles rares jusqu'à ce que l'augmentation de leur fréquence réduise leur valeur sélective et entraîne une contre-sélection. Cette dynamique favorise le maintien d'une multitude d'allèles dans la population. Ainsi, les gènes sous sélection balancée sur une longue durée présentent un polymorphisme synonyme et non-synonyme supérieur à l'attendue sous neutralité, ce qui peut être présenté par un coalescent avec des branches plus longues pendant lesquelles de nombreuses mutations ont pu être acquises au sein des branches séparant les 2 lignées alléliques. (Charlesworth, 2006). En effet, cette sélection fréquence-dépendante négative ne permet pas la fixation d'allèles mais résulte plutôt en un maintien balancé du polymorphisme (pouvant entraîner du polymorphisme trans-spécifique : TSP), avec des valeurs élevées de d_N/d_S (mutations non synonymes conservées sur un long temps évolutif) et/ou d'un pic de polymorphisme associé à la(les) mutation(s) sous sélection dû au fait qu'autour du site sous sélection la recombinaison est contre-sélectionnée (Charlesworth et al., 1997; Hudson and Kaplan, 1988).

Qu'elle soit dirigée ou balancée, la sélection pourra être qualifiée de 'diversifiante' dans un environnement hétérogène et/ou temporellement instable et conduira à maintenir un polymorphisme adaptatif spatialisé à l'échelle de l'espèce, en maintenant les nouveaux mutants en fréquence élevée selon l'avantage qu'ils confèrent aux individus dans un environnement donné.

4.2. Cas de la coévolution hôte-pathogène

Les gènes codant pour le système immunitaire sont classés suite à de nombreuses études parmi les gènes évoluant les plus rapidement dans le génome (Hughes and Nei, 1988; Lazzaro, 2008; Sackton et al., 2007a; Schlenke and Begun, 2003). En effet, alors que les hôtes sont en contact direct avec une large gamme de parasite *sensu lato* auxquels ils doivent s'adapter, les parasites sont reconnus comme étant l'une des plus importantes composantes de la sélection naturelle dans les populations. Ces pathogènes sont en effet ubiquitaires, très abondants avec un temps de génération court - comparés aux hôtes (Rolff and Schmid-Hempel, 2016).

Ceci entraîne une forme de compétition apparentée à une course à l'armement (stratégie de la reine rouge), qui constitue un mécanisme majeur de la coévolution hôte-pathogènes (Dawkins and Krebs, 1979). Un des attendus de cette stratégie du point de vue génétique consiste en une succession de balayages sélectifs puisque lorsque l'hôte fixe une mutation qui confère une résistance à un pathogène améliorant sa fitness, ceci impose en retour une pression de sélection sur ce même pathogène qui va ré-évoluer à son tour *ad finitum*. Le modèle « gene- for-gene » peut aussi être décrit puisqu'il désigne une interaction directe entre chaque gène de virulence d'un agent pathogène et le gène de résistance correspondant de l'hôte.

Les pathogènes, avec leurs potentialités d'adaptation rapides vis-à-vis des populations hôte sont également de très bons candidats pour favoriser un régime de sélection fréquence-dépendante négative sachant que pour ce type de sélection (balancée donc), la valeur sélective d'un allèle est inversement proportionnelle à sa fréquence (Charlesworth, 2006). Des mécanismes complexes ont en effet été proposés pour mieux comprendre la dynamique évolutive de ce type d'interactions : le « time-lagged churning allele model (Dybdahl and Lively, 1998)» qui met en avant le fait que le pathogène s'adapte au génotype le plus commun, ce qui a pour conséquence de diminuer sa fréquence dans la population au bénéfice des allèles les plus rares (préservant donc le polymorphisme à la fois intra- mais aussi inter-population). Ces derniers auront alors un avantage compétitif jusqu'à ce que leur fréquence soit suffisamment élevée pour être de nouveau la cible du pathogène. A ce moment-là, la pression des pathogènes est telle que ces allèles vont de nouveau diminuer en fréquence et remonter en valeur sélective (Decaestecker et al., 2007).

Ainsi, en termes d'évolution et de dynamique des populations, une autre hypothèse expliquant la multiplicité des variants d'un PAM au sein d'une population (et/ou le fait de détecter de la sélection balancée) pourrait être la résultante d'un équilibre entre adaptation locale et migration si les communautés microbiennes diffèrent entre habitats/zones géographiques séparées et/ou à l'histoire évolutive propre des populations hôtes lorsque les flux de gènes sont fortement restreints entre populations.

Cas du CMH des vertébrés

L'action de la sélection balancée – qui a été également décrite pour d'autres effecteurs immunitaires tels les immunoglobulines (Su and Nei, 1999)- a été démontrée pour expliquer le maintien du polymorphisme existant au sein du Complexe Majeur d'Histocompatibilité. Plus particulièrement, il s'agirait de l'action d'une sélection de type fréquence-dépendante négative imposée par les pathogènes (Borghans et al., 2004) bien qu'un avantage à l'état hétérozygote (« overdominance » : les hétérozygotes reconnaissent plus de pathogènes que les homozygotes) ait d'abord été avancé (Hughes and Nei, 1988). La très forte diversité observée serait donc le résultat d'un avantage aux allèles ayant des résidus sous sélection spécifiques pour détecter différents oligopeptides provenant de divers pathogènes (Ejsmond and Radwan, 2015). Ce type de sélection qui favorise plutôt la co-existence d'un maximum d'allèles adapter à chaque oligopeptide exogène peut maintenir des allèles pendant des millions d'années pouvant résulter dans des polymorphismes transpécifiques ou la divergence entre allèles est plus grande que celle entre espèces (Zhao et al., 2013). Les processus de co-évolution hôte pathogène conditionnent ce type de sélection en modifiant constamment les oligopeptides microbiens et donc la production et la conservation d'allèles spécifiques.

5. Evolution rapide des PAMs

En leur supposant uniquement un rôle antimicrobien, le mécanisme d'action des PAMs en ciblant les membranes bactériennes ne paraît pas dans un premier temps être très compatible avec une évolution rapide du PAM puisqu'il est peu probable que les bactéries changent, d'un point de vue biochimique, la composition de leur membrane à travers l'évolution (Lazzaro, 2008; Obbard et al., 2009; Sackton et al., 2007b; Tennessen and Blouin, 2008). Zasloff (2002) a par exemple émis l'hypothèse que la mise en place d'une résistance par un micro-organisme à un PAM est trop complexe pour être rapidement acquise et donc peu probable. Malgré ceci, de nombreux exemples montrent désormais que les peptides antimicrobiens se diversifient rapidement après spéciation et il est souvent difficile de les aligner dès lors que les taxons incriminés divergent trop. De plus comme exposé précédemment, au niveau populationnel, le niveau de diversité génétique non synonyme est souvent élevé dans la région codant le peptide antimicrobien (c'est également vrai pour la

prorégion). La suite de l'introduction s'attachera donc à exposer comment la sélection peut être mise en jeu –ou non– pour expliquer que la divergence inter-espèces de ces peptides antimicrobiens soit élevée.

5.1. Evolution différentielle des PAMs après spéciation

Quelques études phylogénétiques se sont intéressées à l'action de la sélection sur des lignées orthologues (bien que l'évolution a pu favoriser la disparition de certaines lignées d'orthologues après spéciation) de peptides antimicrobiens dans le but de détecter la trace d'une sélection positive qui aurait agi sur des espèces ayant des trajectoires évolutives distinctes et des environnements différents.

Hollox & Armour (2008) ont détecté l'action de la sélection positive dans l'évolution des 17 beta-défensines de primates catarhiniens en identifiant les orthologues de ces gènes chez les humains, les chimpanzés, orang outans, gorilles, gibbons et les macaques rhésus. L'analyse bayésienne ayant permis de rechercher des codons sous sélection positive en comparant des modèles de sélection et d'évolution 'presque neutre', ces auteurs ont pu montrer que cinq des 17 beta-défensines (DEFB1, DEFB118, DEFB120, DEFB127 and DEFB132) - avec des activités antimicrobiennes avérées - présentaient la trace d'une sélection positive sur au moins un codon par gène. Pour le PAM DEFB127, le 71^{e} codon sous sélection positive varie entre espèces et la reconstruction des séquences ancestrales montre que cette position a muté indépendamment cinq fois pendant l'évolution des primates. De plus, dans l'un des gènes (DEFB1), le site sous sélection positive a un rôle sur l'activité antimicrobienne du PAM et les auteurs en déduisent que la sélection a modulée la charge et la polarité de la molécule en fonction des différents environnements microbiens.

Etudier ensuite le polymorphisme de ces beta-défensines chez les humains leur a permis de mettre en évidence l'action d'une sélection balancée sur certains variants et notamment le maintien de deux clades alléliques majoritaires qui sont discriminés majoritairement par deux positions décrites comme étant sous sélection positive dans l'évolution des primates (et donc potentiellement maintenu sur de longue période évolutive par sélection balancée).

Chez les grenouilles, Duda et al. (2002) ont observé quant-à-eux de forts taux de substitution en acides aminés ($d_N/d_S>1$) chez plusieurs familles de PAMs, ce qui leur a permis de mettre en évidence l'action d'une sélection diversifiante sur le peptide mature pouvant être

expliquée également par une réponse à des environnements microbiens distincts selon la nature de l'habitat dans lequel les espèces ont évoluées. L'action de la sélection positive a également pu être mise en évidence sur les alpha défensines lors de l'évolution des mammifères (Lynn et al., 2004a). Cette sélection est là aussi plus spécifiquement orientée sur le peptide antimicrobien lui-même et non sur la prorégion suggérant des modifications fonctionnelles importantes du PAM. Duda et al. (2002) ont également suggéré qu'une co-évolution entre différentes régions du prépropeptide (prorégion et PAM) aurait pu avoir eu lieu chez une famille d'amphibiens, les Hylidae, pour faire en sorte que la prorégion fournisse des charges électriques complémentaires au peptide afin de l'inhiber avant clivage et éviter ainsi sa cytotoxicité au sein de l'hôte. Une co-évolution de même nature entre prorégion et domaine mature avait déjà été rapporté chez les défensines de mammifères (Hughes and Yeager, 1997a).

En plus de ces exemples d'évolution des PAMs par sélection positive, l'action de la sélection purifiante a quand même aussi pu être mise en évidence sur l'évolution de plusieurs défensines orthologues de primates (DEFB103) permettant aux auteurs d'en déduire que sa fonction de défense a dû être fixée très tôt dans l'évolution des mammifères (Crovella et al., 2005).

Chez les invertébrés, les premières études n'ont pas réussi à démontrer de façon probante l'action d'une sélection positive comme une récurrence dans l'évolution des PAMs notamment chez la drosophile plus particulièrement étudiée (Jiggins and Kim, 2005; Lazzaro and Clark, 2003; Sackton et al., 2007a). Sackton et al. (2007) ont même suggéré que la sélection diversifiante attendue sur le PAM aurait plutôt lieu sur les éléments régulateurs que sur le peptide lui-même. Erler et al. (2014) ont quand même montré que l'hyménoptaecine (un PAM inductible suite à une infection décrite pour la première fois chez *Apis mellifera*) avait évolué sous sélection positive en comparant 12 espèces du genre *Bombus*. Les sites sous sélection positive de l'hyménoptaecine sont distribués aléatoirement le long du gène et ne sont pas spécialement décrits dans la région C-terminale codant le PAM. L'évolution de la défensine-1 dans cette même étude est également étudiée mais les auteurs ne détectent pas de signe d'évolution adaptative pour ce gène illustrant la difficulté, là encore, de faire des généralités quant à l'histoire évolutive des gènes codant les PAMs

tant la dynamique évolutive semble être complexe (notamment quant à la recherche/détection d'orthologues qui peuvent avoir été perdus au cours de l'évolution).

5.2. Evolution des paralogues par sélection positive

Chez les insectes sociaux, l'action d'une sélection positive a été décrite sur les termicines des termites australiennes du genre *Nasutitermes.* Cette sélection a pour conséquence de diminuer la charge positive du PAM dans les lignées actuelles par rapport à la molécule ancestrale (Bulmer & Crozier, 2004 : les PAM des groupes d'espèces frères montrant substantiellement des charges plus élevées). Le rôle de la sélection positive a également été montré dans l'évolution de certains duplicats codant pour les PAMs chez les termites (Bulmer & Crozier, 2004). Chez ces insectes, des évènements de duplication ont été montrés avant mais aussi après spéciation (dans 2 des 11 cas) et l'action d'une sélection diversifiante a été mise en évidence sur plusieurs paralogues selon les différentes espèces analysées. Une première étape a été de regarder si certaines lignées de paralogues avaient évolués sous sélection positive. Cette analyse n'a pas été conclusive mais la comparaison de modèles de sélection sur les codons leur a permis de montrer que certains codons sont effectivement sous sélection positive au niveau du peptide mature ($d_N/d_S>>1$) entre les différents paralogues. A l'inverse, le peptide signal est sous forte sélection purifiante ($d_N/d_S<1$). Les auteurs en déduisent que la divergence significative entre les différents paralogues serait une réponse des espèces aux nouveaux pathogènes (diversification en termes d'abondances et de diversité des pathogènes). Ces pathogènes sont en effet différents selon l'environnement dans lesquelles les espèces ont évoluées après leur radiation : de la savanne (dry savannah) aux forêts tropicales humides. De plus, puisqu'il apparaît que la modification de charge du PAM soit plus forte entre duplicats d'une même espèce qu'entre les espèces, les auteurs en concluent que l'avantage à avoir deux PAMs avec deux charges différentes serait une contre-réponse à l'évolution de la résistance des pathogènes fongiques chez les insectes sociaux particulièrement vulnérables à ces épidémies.

Chez l'huître *C. gigas*, une étude s'est intéressée également à détecter l'action de la sélection sur le maintien/création de diversité génétique au sein de 2 PAMs : les Cg-défensines (Cg-defs : Cg-defh1-2 et Cg-defm) et la Cg Proline rich peptide (Cg-Prp) (Schmitt et al., 2010). Tout d'abord pour les Cg-defs, le peptide signal est conservé au sein des trois

classes de défensines, tout comme le sont les cystéines impliquées dans les ponts disulfures de la région du PAM et un motif Pro-Arg ayant une importance dans l'activité de celui-ci, L'action de la sélection positive a quant à elle été mise en évidence sur certains codons changeant la polarité et/ou la charge positive des PAMs pour quelques paralogues (affectant l'affinité des peptides pour les membranes bactériennes selon les auteurs).

De la même façon, des indices de sélection positive entre paralogues d'une même espèce ont été trouvés chez les séquences codant les PAMs de vertébrés. Maxwell et al. (2003) et Morrisson et al. (2003) ont étudié l'évolution des beta-défensines chez le modèle murin (sous espèces de *Mus musculus*) et montré que l'action de la sélection positive diversifiante après duplication s'effectuait essentiellement au sein d'une petite région du peptide mature avec des séquences relativement conservées au sein du peptide signal et de la prorégion. Chez l'homme, pour ces mêmes beta-défensines, il a été montré que les isoformes de cette famille multigénique étaient issues d'évènements de duplication suivi d'une relaxation de la sélection purifiante et/ou de l'action de la sélection positive/diversifiante sur certains codons du peptide mature avant la séparation humain-babouin il y a 23 millions d'années (Maxwell et al., 2003; Semple et al., 2003). Cette étude révèle que le précurseur protéique des beta-défensines est composé de deux régions : un peptide signal (codé par le premier exon) et une petite prorégion suivie du peptide antimicrobien (codé par le deuxième exon). La divergence entre paralogues se fait uniquement au sein de ce deuxième exon et correspond à un excès de mutations non-synonymes au niveau du PAM mature sur 9 sites entourant une cystéine conservée.

5.3. Maintien du polymorphisme par sélection positive à un locus donné : une nécessité du système de défense?

Comme cela a pu être montré dans le cas du CMH, l'action de la sélection balancée agit pour le maintien du polymorphisme sur de longues échelles temporelles. Quid du maintien du polymorphisme chez PAMs ?

Dans le cas de la brévinicine-1.1 chez les grenouilles léopard du genre *Rana,* Tennessen & Blouin (2008) ont montré que la sélection positive permettait le maintien du polymorphisme du peptide mature (forte variabilité des mutations non synonymes dans cette région et forte diversité haplotypique). En effet, ces auteurs montrent l'existence de plusieurs lignées alléliques divergentes pour cette région du PAM avec des allèles partagés

entre les génomes de plusieurs espèces. A l'inverse, chez l'espèce *R. pipiens*, il a été montré pour la ranatuerine-2 l'existence d'un balayage sélectif (Tennessen and Blouin, 2007) bien que ces auteurs (Tennessen and Blouin, 2008) n'excluent ensuite pas l'hypothèse d'une « fluctuating selection » (sélection changeante dans le temps et l'espace) dans le cas où les autres allèles aient disparu récemment ou soient trop rares pour être échantillonnés dans les populations étudiées. Chez les beta-défensines de la mésange charbonnière et de la mésange bleue, un cas de polymorphisme trans-spécifique avec des taux d'hétérozygotie élevés pour plusieurs gènes a permis aux auteurs de suggérer l'action de la sélection balancée agissant pour le maintien du polymorphisme sur du long terme, même après spéciation (Hellgren and Sheldon, 2011). Un même cas de polymorphisme trans-spécifique a été révélé dans le cas de la morue atlantique et le colin d'Alaska qui serait également révélateur de l'action d'une sélection balancée sur le maintien de polymorphisme puisque présent avant la spéciation (Halldórsdóttir and Árnason, 2015).

Chez les invertébrés, des polymorphismes trans-spécifiques chez les PAMs ont été récemment détectés chez les arthropodes (Unckless and Lazzaro, 2016; Unckless et al., 2016). En effet, dans le cas de la diptéricine chez la drosophile, ces auteurs notent la présence d'un remplacement Ser->Arg qui est apparu cinq fois dans le sous genre *Sophophora*. Ces derniers ont également pu mettre en évidence un polymorphisme d'indels (3 codons) chez l'attacine qui est apparu chez trois espèces de drosophile *D. melanogaster*, *D. sechellia* et *D. mauritiana* suggérant, là encore, soit une convergence évolutive, soit le maintien de cette mutation par sélection balancée à travers les événements de spéciation. Finalement, Unckless *et al*. (2016) montrent dans une revue de synthèse que des exemples de polymorphismes trans-spécifiques de PAMs sont assez courants dans la littérature. Par exemple, chez la cécropine, une alanine et une valine ségrégent en 6^{e} position du peptide signal dans le polymorphisme de deux espèces proches *Bombyx mori* et *B. mandarina* (Guo et al., 2011). Dans le cas de la termicine, une valine et une arginine, d'une part, et une histidine et une arginine, d'autre part, ségrégent également en 13^{e} position du peptide signal et en 14^{e} position du peptide mature dans le polymorphisme des espèces *Reticulitermes chinensis* et *Odontotermes formosanus*. L'existence de polymorphisme partagé entre espèces proches pourraient donc être le reflet d'une sélection naturelle qui fluctue dans le temps et l'espace en réponse à des différences de diversité des pathogènes (Unckless and

Lazzaro, 2016). Ainsi, l'action de la sélection balancée pourrait se faire sur un nombre restreint d'acides aminés qui ne permet pas *in fine* d'être détectée en réalisant des analyses de criblage génomique comme cela est souvent le cas pour étudier le rôle de la sélection sur les gènes codant les effecteurs du système immunitaire.

Bien que l'action de la sélection balancée soit largement évoquée, il reste tout de même important de garder en tête que ces polymorphismes trans-spécifiques peuvent également être le produit d'évènement d'introgression/hybridation entre espèces (Hedrick, 2013) et/ou de mutations qui seraient apparues indépendamment au sein de deux espèces proches (convergence évolutive). Chez la moule *M. edulis*, un polymorphisme sur un codon (leucine/arginine L31R) a été mis en évidence chez le peptide antimicrobien MGD2 avec une différenciation génétique entre populations en Europe et qui serait le résultat d'un maintien adaptatif d'un polymorphisme pré-existant à des fréquences différentes au sein des populations étudiées (Gosset et al., 2014). Ce polymorphisme pourrait être le résultat d'un contact secondaire et hybridation entre les espèces *M. edulis* et *M. galloprovincialis* qui serait désormais maintenu soit par sélection balancée selon les communautés microbiennes différentes rencontrées par les populations de moules, soit à de la sélection « intermittente » qui résulte en des fluctuations aléatoires des fréquences alléliques dans le temps et l'espace.

Pour résumer, comme le dit Tennessen (2005): "*Positive selection on AMP genes is very common and has resulted in an enormous functional diversity of these molecules among species and among loci in many taxa*". Cet auteur précise de plus qu'une des tendances générales de l'évolution des PAMs est que les cystéines (nécessaire à la structure 3D du peptide) sont conservées et les sites sous sélection positive trouvés au plus loin desdites cystéines. Le peptide signal a tendance à être sous sélection purifiante de même que la prorégion (Tennessen, 2005). Lorsque les implications fonctionnelles d'un changement en termes de structures secondaires ont été étudiées, il a été montré à plusieurs reprises que la modification d'un seul acide aminé changeait l'activité du PAM. La défensine humaine, HNP1, qui diffère d'un seul acide aminé par rapport à HNP3, possède une activité anti *C. albicans* qui n'est pas détectable chez HNP3 (Raj et al., 2000). Chez *Drosophila melanogaster* également, la drosomycine présente six isoformes, issues d'évènements de duplication, avec des activités fongicides différentes (Yang et al., 2006). De la même façon, les cécropines ont

évolué en plusieurs paralogues dont les activités antimicrobiennes sont différentes. Chez *H. cecropia*, la cécropine B est celle qui montre l'activité anti Gram négative la plus forte alors que toutes les cécropines (A, B et D) montrent des activités contre les bactéries à Gram positif et Gram négatif (Tassanakajon et al., 2015). Très récemment, Unckless *et* al. (2016) ont également montré que le changement d'un seul acide aminé sur un PAM de la drosophile avait une forte influence vis-à-vis de leur susceptibilité à l'infection montrant que de telles modifications peuvent en effet bien avoir un effet sur la valeur sélective de l'hôte. Ainsi, l'action de la sélection positive que ce soit entre paralogues au sein d'une espèce ou dans l'évolution des orthologues –après spéciation- pourrait se faire le plus souvent pour changer la charge de la molécule en fonction des différents environnements biotiques (présence de nouveaux pathogènes par exemple) mais aussi abiotiques (l'habitat ayant un impact sur l'immunocompétence et la susceptibilité aux infections notamment chez les insectes sociaux; Fuller et al., 2011) dans lesquels les espèces ont évoluées. De plus, une forte diversité non-synonyme de ces molécules peut être considéré comme un bénéfice sélectif dans un environnement dans lequel les pathogènes évoluent rapidement en permettant à la population d'être plus flexible vis-à-vis de changements spatio-temporels des communautés microbiennes (Du Pasquier, 2006).

OBJECTIFS DE THESE

La large gamme d'habitats dans lesquels sont retrouvées les annélides, et donc d'environnements microbiens associés, offre un contexte idéal pour comprendre l'évolution des adaptations immunitaires au sein de de ces environnements particuliers. Dans cette thèse ont été considérés (1) le milieu hydrothermal profond caractérisé par des régimes fluctuants de température et de pH, une pression particulièrement élevée (250 fois la pression atmosphérique) et une forte hypoxie au sein duquel les espèces ont un long passé d'adaptation (Little and Vrijenhoek, 2003) (2) les environnements envasés portuaires qui constituent également des environnements contraignants dans lesquels les organismes évoluent dans des sédiments pollués, riches en matière organique et en sulfures et peuvent rencontrent largement des conditions anoxiques (Cuvillier-Hot et al., 2014). Les peptides antimicrobiens sont des composants clefs des systèmes de la défense immunitaire innée et ils sont répandus dans tous les phylums du règne vivant, incluant les vertébrés, invertébrés (Zasloff, 2002). Les gènes codant pour le système immunitaire sont largement décrits comme évoluant sous la pression sélective des environnements microbiens, et la détection de la sélection positive sur les familles multigéniques codant les peptides antimicrobiens chez vertébrés et invertébrés peut être décrite (Duda et al., 2002; Tennessen, 2005; Unckless and Lazzaro, 2016; Unckless et al., 2016). Ces molécules se diversifient donc rapidement par différents mécanismes que ce soit pour créer de la diversité fonctionnelle pouvant permettre d'affronter de nouveaux nouveaux mutants pathogène dans la population ayant évolués pour échapper à la réponse immunitaire de l'hôte et/ou vis-à-vis de fluctuations des communautés microbiennes dans le temps et l'espace (habitats nouveaux). Dans cette thèse, des PAM de la famille des arenicine-like possédant la structure typique des protéines à domaine BRICHOS (domaine chaperon) ont été étudiés. Ce plus, ces PAMs sont les seuls représentants, jusqu'à aujourd'hui, de cette famille de PAM décrite uniquement chez les polychètes.

Ainsi, cette thèse se propose d'apporter un éclairage sur les patrons de diversité génétique et de sélection agissant sur les différents domaines qui composent le précurseur protéique de ces effecteurs immunitaires. De plus, les deux espèces étudiées montrent des cas d'association symbiotique: de l'épibiose diversifiée et obligatoire dans le cas *d'Alvinella pompejana* à la symbiose facultative pour *Capitella sp* avec une seule souche microbienne

Thiomargarita. Il pourra donc également s'agir d'effectuer des inférences quant au rôle de ces effecteurs immunitaire dans la mise en place de ces associations.

Le manuscrit s'articulera en 2 chapitres principaux qui traitent (1) de l'évolution moléculaire du gène codant le précurseur protéique de l'alvinellacine (preproalvinellacine) et la structure génétique du/des locus dans les populations d'*Alvinella pompejana* et (2) la même approche concernant le gène codant la preprocapitellacine et structure génétique du locus dans les populations de *Capitella sp* en Atlantique Nord Est et en Manche.

Notamment, il s'agira dans ces deux chapitres d'étudier si les précurseurs protéiques sont codés par des familles multigénique (duplication en tandem qui s'influencent par recombinaison ?), si l'évolution des (potentielles) duplications est adaptative (action de la selection positive sur la divergence des paralogues et/ou pour le maintien d'un polymorphisme), si les pressions de sélections éventuelles sont les mêmes au sein de toutes les régions du précurseur (mise en évidence d'une coévolution entre proregion et région du PAM ?) et si le passage d'allèles à travers une barrière semi perméable aux flux de gène peut être mise en évidence (avantage sélectif à partager un pool d'allèle issus d'évènements de spéciation via introgression/hybridation ?)

Chapitre 2. Diversité génétique et histoire évolutive du gène codant le précurseur protéique du peptide antimicrobien alvinellacine chez *Alvinella pompejana*.

INTRODUCTION

L'environnement hydrothermal

En milieu profond, les premières communautés associées aux sources hydrothermales ont été découvertes par Lonsdale sur la ride des Galapagos en 1977. Cette découverte a ensuite été suivie par le recensement de nombreuses communautés d'organismes symbiotiques, toutes inféodées aux environnements chimiosynthétiques sur l'ensemble des dorsales explorées mais également le long des arcs volcaniques et les zones de subduction (Chevaldonné et al., 1997; Desbruyères et al., 2006; Rogers et al., 2012). Depuis, l'environnement hydrothermal a pu être retrouvé au sein de tous les océans et la faune particulière qui lui est associée se répartit en 6 provinces biogéographiques (variation entre 5 à 8 selon les auteurs) puisque retrouvée aussi bien sur les dorsales du Pacifique nord-est que de l'Atlantique, l'océan Indien, les bassins arrière arc où les arcs volcaniques du Pacifique ouest (Figure1 : Van Dover et al., 2002; Tyler et al., 2002; Moalic et al., 2011). Ces communautés ont également été retrouvés sur les dorsales Antarctique (Brandt et al., 2012; Rogers et al., 2012) et Arctique (Edmonds et al., 2003). Cet écosystème se distribue d'une façon quasi-linéaire sur l'équivalent d'une chaîne de montagne de 70 000 km de long à l'échelle du globe et les émissions sont en général regroupées au sein de sites de quelques centaines de mètres, regroupés ensuite en champs hydrothermaux de quelques kilomètres de longueur eux même connectés – ou non – sur une échelle de l'ordre de la dizaine voir de centaines de kilomètres de longueur (Chevaldonné et al., 1997).

L'environnement hydrothermal, en plus d'être dépourvu de lumière, est caractérisé par la décharge dans l'eau de fond froide (2-5°C) et oxygénée d'un fluide hydrothermal anoxique et riche en sulfure d'hydrogène, métaux lourds (Cd, Cu, Fe, Hg, Zn), ammoniac, CO_2 et

méthane, pouvant atteindre jusqu'à 400°C avec un pH de 3,5. A ce titre, il peut être considéré comme un milieu particulièrement contraignant/pollué/toxique pour la vie dans des zones profondes (notamment pour les espèces marines classiques). La pression hydrostatique est forte et peut également être une pression de sélection sur les organismes en tant que telle (Siebenaller and Somero, 1978) - en plus de pouvoir jouer sur la composition chimique du fluide. Cet environnement est hétérogène, fragmenté et temporellement instable à toutes les échelles d'espace et présente des variations thermo-chimiques importantes à micro-échelle spatiale (Sarradin et al., 1999). Le mélange dynamique entre le fluide hydrothermal et l'eau de fond (2-5°C, pH 7,8) conduit à une forte précipitation des minéraux du fluide hydrothermal à l'origine de la formation des cheminées hydrothermales : édifices polymétalliques composés principalement de fer, manganèse, cuivre et zinc (Fisher et al., 2007). La présence de sources dépend principalement de la présence/proximité du réservoir magmatique et de zones de fracture qui conditionnent l'entrée de l'eau de fond vers lithosphère. En effet, c'est la fracturation du plancher océanique au niveau de nombreuses fissures près des dorsales qui va permettre à l'eau de mer de fond (dense et froide) de s'infiltrer jusqu'à la chambre magmatique où elle va se réchauffer entraînant une diminution de sa densité. L'eau réchauffée sous pression va ainsi remonter, se charger en différents éléments qui constituent les roches qu'elle va rencontrer et va s'acidifier (pH 3,5). C'est cette acidification qui permettra à des nombreux éléments métalliques de se dissoudre et va donner naissance au fluide hydrothermal. A sa sortie, le fluide va jaillir de la croute océanique, et crée une zone de mélange avec l'eau de mer froide et, de par leur différence de nature chimique, va entraîner la précipitation de nombreux composés du fluide hydrothermal (caractérisé par le terme « fumeur noir »). C'est la précipitation de divers sulfures métalliques et de sulfate de calcium anhydre qui entraîne la constitution progressive des cheminées hydrothermales (Zierenberg et al., 2000).

L'initiation ou l'arrêt d'une source est intimement lié(e) aux éruptions volcaniques et aux réarrangements tectoniques dans la vallée axiale de la dorsale mais aussi au colmatage progressif en sub-surface des fissures ayant donné lieu à la résurgence hydrothermale. Ce milieu est donc caractérisé comme un environnement instable et fragmenté puisqu'il est délimité par de nombreuses ruptures/décalages de la dorsale (failles transformantes ou discontinuités géologiques) qui affectent sa linéarité et modifient la circulation des masses

d'eau dans et au-dessus de la vallée (Chevaldonné et al., 1997). Ces segments de dorsale peuvent influencer la connectivité et le renouvellement des communautés associées puisque cette segmentation représente souvent une entrave à la dispersion larvaire et donc aux flux géniques lorsque les espèces se distribuent sur l'ensemble de la dorsale. Dans la Figure 1, l'exemple B représente la dorsale est Pacifique pour laquelle la zone équatoriale avec le Hess Deep et les zones de fracture Quebrada/Discovery/Gofar représentent une barrière semi-perméable aux flux géniques (Plouviez et al., 2010). En effet, ces auteurs (mais voir aussi Hurtado et al., 2004; Jang et al., 2016) ont pu montrer que les populations de l'espèce *A. pompejana* au nord et au sud de cette zone pouvaient être considérées comme des isolats géographiques en voie de spéciation avec une barrière semi perméable aux flux de gènes et la présence d'hybrides à 7°S/EPR.

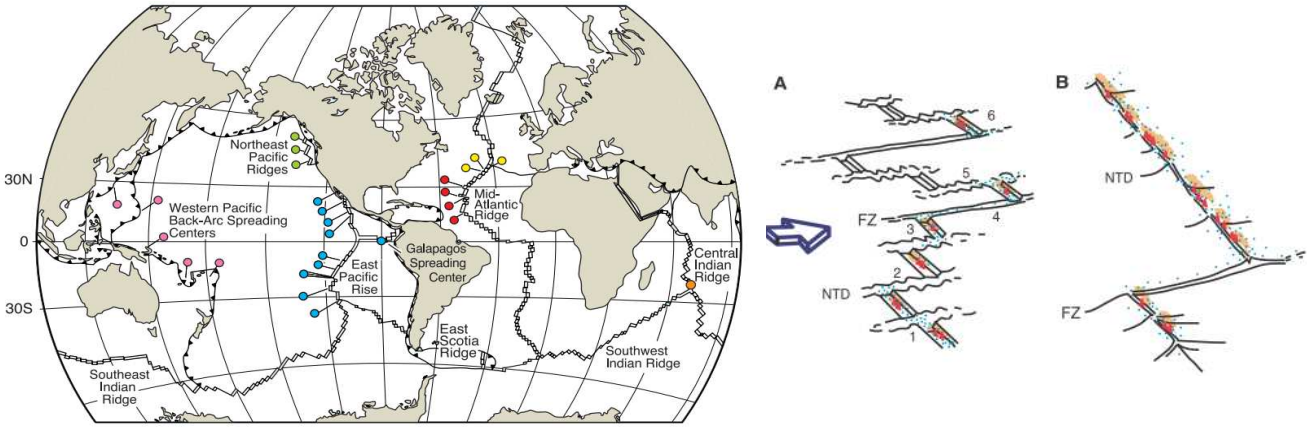


Figure 1. A gauche : distribution en 6 provinces biogéographiques des communautés hydrothermales et à droite : schémas de discontinuité d'une dorsale océanique selon leur taux d'accrétion lent ou rapide (adapté de Van Dover et al., 2002).

Les communautés associées

Dans cet environnement hétérogène fragmenté et instable, d'importantes communautés microbiennes se développent et sont constituées d'Archaea et de Bacteria thermophiles et extrêmophiles. Les espèces inféodées à ces milieux sont caractérisées par un fort endémisme mais aussi par leurs associations symbiotiques avec des bactéries chimioautotrophes dont le métabolisme met à profit l'oxydation des composés réduits présents dans l'environnement pour réaliser une production primaire qui sera ensuite exploitée par les espèces hétérotrophes (procaryotes ou eucaryotes) (Dubilier et al., 2008; Kiel and Tyler, 2010; Orcutt et al., 2011). La biomasse animale, bien qu'abondante, est peu

diversifiée et est dominée par trois groupes taxonomiques d'invertébrés marins qui représentent la majorité de la biomasse des sources hydrothermales : annélides, arthropodes et mollusques dont la plupart dépendent d'associations symbiotiques (Cavanaugh et al., 2006; Tunnicliffe et al., 1998).

Le ver de Pompéi : *Alvinella pompejana*

Les annélides symbiotiques associées à ces environnements font partie de deux familles de polychètes : les Siboglinidae et les Alvinellidae (appartenant à l'infraclasse des Canalipalpata) (Bright and Giere, 2005). Une des espèces emblématiques de l'écosystème hydrothermal est l'annélide polychète tubicole *Alvinella pompejana* (Desbruyères & Laubier, 1980). Ce « ver de Pompéi » colonise exclusivement la paroi des cheminées des sources hydrothermales de la dorsale Est-Pacifique (EPR) et de Guaymas de 23°N à 38°S (Hurtado et al., 2004). Les adultes qui vivent en effet en colonie dans des tubes organiques fixés à la paroi de la cheminée, présentent de manière systématique une association avec des épibiotes qui recouvrent les expansions dorsales inter-segmentaires de cet annélide (Le Bris and Gaill, 2006). Son espèce sœur du point de vue phylogénétique, *A. caudata,* vit en syntopie avec *A. pompejana* et a pour le moment été retrouvée dans tous les échantillons géographiques où l'espèce *A. pompejana* a été échantillonnée, faisant croire, à sa découverte, à une forme épitoque du ver de Pompéi (Desbruyères & Laubier, 1980). L'analyse génétique a permis ensuite de trancher sur leur statut d'espèce à part entière (Autem et al., 1985), les 2 espèces présentant une très forte divergence évolutive (Fontanillas et al., 2017). Les deux espèces montrent des cas d'association avec des épibiotes ce qui n'est pas le cas des espèces du genre voisin *Paralvinella* (Cary et al., 1997). L'espèce *A. pompejana* se place parmi les métazoaires les plus thermotolérants (5-80°C) et les plus eurythermes rencontrés à ce jour (Desbruyères et al., 1998; Le Bris & Gaill, 2006; Ravaux et al., 2013). Cependant, l'environnement immédiat à l'intérieur des tubes dans lequel l'animal vit se caractérise par des températures n'excédant pas 50°C (Desbruyères et al., 1998) mais la température à l'extérieur des tubes varie beaucoup plus même si elle n'atteint que rarement plus de 60°C (Le Bris and Gaill, 2006). (Ravaux et al., 2013) ont de plus montré que des expositions prolongées (>2heures) au-dessus de 55°C étaient létales pour le ver en conditions simulées (expérimentations sous pression à bord) et que l'optimum thermique de cet annélide se trouverait autour de 42°C. Il a été suggéré que les bactéries insérées sur le

tégument dorsal (et présentes à l'intérieur des tubes de façon libre, Figure 2) fourniraient une source de nourriture au ver et interviendraient dans la détoxification de l'environnement immédiat du ver (Alayse-Danet et al., 1987). Ce rôle reste très spéculatif et encore largement méconnu. Des analyses de métagénomique ont permis de caractériser la microflore épibiotique qui est composée d'un complexe d'espèces de 12 à 15 phylotypes dont 98% appartiennent au groupe taxonomique des epsilon-protéobactéries. Ces epsilon-protéobacteries, retrouvées aussi bien en association avec des invertébrés hydrothermaux qu'à l'état libre représentent le groupe prédominant des micro-organismes vivant en association avec les espèces hydrothermales. Ils sont également prépondérants dans les 'mattes' bactériennes retrouvées sur de nombreux substrats et considérées, de façon générale, comme un maillon clé des environnements sulfurés (Cary et al., 1997; Grzymski et al., 2008). Cependant, deux phylotypes (5A et 13B) sont principalement retrouvés dans le consortium d'épibiontes du tégument dorsal *d'Alvinella pompejana* et ont été décrits comme des bactéries filamenteuses chimio-litho-autotrophes. Grzymski et al. (2008) suggèrent pour conclure sur la nature de ces épibiontes que « *The success of* Epsilonproteobacteria *as episymbionts in hydrothermal vent ecosystems is a product of adaptive capabilities, broad metabolic capacity, strain variance, and virulent traits in common with pathogens*" et que ces propriétés leur permettent de prospérer dans cet environnement aux régimes thermique et chimique changeants. Cette analyse a également révélé la présence de gènes codant certaines enzymes impliquées dans le cycle rTCA, la sulfo-oxydation, la dénitrification ainsi que dans les métabolismes hétérotrophes et aérobies. Ainsi, les épibiontes présentent une large flexibilité métabolique qui leur permet de s'adapter de manière optimale aux importantes fluctuations de l'habitat d'*Alvinella pompejana* tout en gardant un potentiel virulent pour l'hôte. Il a été montré que *A. caudata* et *A. pompejana* possèdent au moins un phylotype dominant en commun dans leur consortium épibiotique (Cary et al., 1997). Au niveau morphologique chez *A. pompejana*, l'établissement de cette symbiose est rendue possible par la sécrétion de gouttes de mucus par des protubérances insérées entre le neuropode et le notopode. C'est dans cette partie que s'intègrent l'extrémité de nombreuses bactéries, principalement filamenteuses. D'autres types bactériens sont également retrouvés : forme en bacille, coccoïdale, et en spirale… C'est cette diversité microbienne qui donne cet aspect « chevelu » au ver. *A caudata* quant à elle porte ses épibiontes seulement sur la partie postérieure effilée de son

corps (expansions sur les parapodes de la région caudale) qu'elle met en contact direct avec le fluide hydrothermal dans la partie basse de son tube attaché à la cheminée.

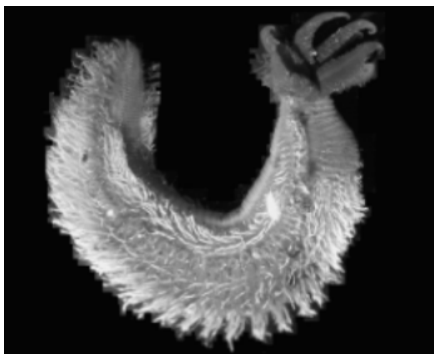


Figure 2. Micrographie d'*A. pompejana* et ses ectobactéries symbiotiques.

Un peptide antimicrobien (PAM) nommé alvinellacine a été isolé et identifié à partir du liquide coelomique d'*Alvinella pompejana* (Tasiemski et al., 2014). Cette étude a permis de mettre en évidence que le PAM est exprimé de façon constitutive par les cellules épithéliales du tégument de l'animal et se retrouve dans l'environnement hydrothermal en contact direct avec les épibiontes. De plus, il a été montré que l'alvinellacine possédait une activité bactéricide contre les bactéries de type filamenteuse dominantes trouvées sur son tégument dorsal ainsi qu'une activité bactérienne anti-gram négatif. Chez l'espèce hydrothermale, ce PAM participe à l'immunité anti-bactérienne mais aussi au contrôle et à la sélection des épibiotes (Tasiemski et al., 2014). Ce rôle de contrôle de l'épibiose est renforcé par le fait que (1) le gène est sélectivement induit dans les coelomocytes par les épibiontes de l'animal et (2) qu'il est secrété par ces mêmes coelomocytes lorsqu'une infection par les bactéries du tégument de l'animal est mimée.

Comme présenté dans l'introduction générale, les PAMs sont documentés comme ayant un rôle très important dans la modulation/maintien/mise en place/contrôle de l'interaction bénéfique entre l'hôte et ses microbes commensaux/mutualistes (Login et al., 2011; Masson et al., 2016). Il existe en effet un nombre croissant d'évidences suggérant que le système immunitaire inné interagit différemment avec les symbiotes et les pathogènes chez les invertébrés. Un autre exemple - en plus de ceux traités dans l'introduction - existe chez l'hydre pour lequel il a pu être montré que les périculines, bien que montrant une forte activité bactéricide permettant de leur attribuer un rôle de défense, sont responsables de la sélection de bactéries dont le rôle est bénéfique et obligatoire pour le développement de l'hydre (Bosch et al., 2009; Fraune and Bosch, 2010). Ainsi la mise en place de telles

symbioses, comme celle documentée chez *Alvinella pompejana*, nécessite un système de domestication des bactéries et de reconnaissance de celles-ci par l'hôte à travers son arsenal de défense immunitaire qu'il est particulièrement intéressant d'étudier notamment en ce qui concerne les peptides antimicrobiens.

Objectifs du chapitre

La colonisation progressive de l'environnement hydrothermal et les fluctuations physico-chimiques du milieu (environnement instable) va entraîner des pressions de sélection dirigée sur certains processus physiologiques des organismes inféodés à ces environnements et notamment de nombreuses adaptations (pigments respiratoires, branchies, système de détoxification de l'H_2S, et même conditionner l'évolution des protéines… (Powell and Somero, 1986; Hourdez and Lallier, 2007; Jollivet et al., 2012). Par exemple, le milieu hydrothermal est un milieu riche en divers métaux lourds et éléments radioactifs et il a été montré chez les annélides hydrothermaux une capacité d'accumuler ces éléments dans leur tissus qui est 1000 fois supérieurs à celle des organismes côtiers (Cherry et al., 1992), ce qui a une incidence sur les systèmes de réparation des acides nucléiques (Pruski & Dixon 2003). D'autres études ont également pu mettre en évidence par exemple une plus forte expression de l'enzyme superoxide dismutase (qui constitue la première ligne de défense contre les dommage liés aux ROS « reactive oxygen species ») chez les crabes hydrothermaux (famille des Bythograeidae) par rapport aux crabes littoraux (de différentes familles dont les Cancridae, Dromiidae…) (Marchand et al., 2009). Chez les crevettes, les protéines des chocs thermiques sont elles aussi surexprimées en réponse à un stress thermique (30°C) plus fortement chez *Rimicaris exoculata* que celle mise en place par l'espèce *Palaemonetes varians* (monophyletic shallow water relative) (Cottin et al., 2010). Les auteurs en déduisent que, puisque *P. varians* subit des stress thermiques de 30°C (pendant l'été à marée basse), cette réponse chez l'espèce hydrothermale serait dû à une absence d'acclimatation préalable due aux fluctuations extrêmement rapides de cet environnement (Cottin et al., 2010). Ainsi, s'adapter à cet environnement hypervariable et toxique a constitué un défi évolutif pour les espèces qui lui sont inféodées et notamment dans la mise en place de symbioses trophiques. Les symbioses (endo/ecto) ont en effet un rôle clé pour supporter la vie des métazoaires adaptée à ces environnements et notamment l'exploitation des bactéries chimiolithoautotrophes souvent anaérobiques stricts (Alayse-

Danet et al., 1987), ce qui leur imposent de pallier au manque d'oxygène du milieu (Bernhard et al., 2000). Les annélides polychètes alvinellidae (Terebellida) sont les premières espèces à coloniser les cheminées hydrothermales et les deux espèces *A. pompejana* et *A. caudata* de ce groupe sont retrouvées en association avec des bactéries épibiotiques. L'alvinellacine, PAM synthétisé par l'espèce *A. pompejana*, est impliquée dans le contrôle et la régulation des populations des bactéries épibiotiques. C'est une des molécules clés de l'immunité externe du ver en étant secrétée directement dans le milieu externe pour entrer en contact direct avec les épibiontes. A ce titre, elle subit directement les conditions abiotiques de l'environnement hydrothermal et doit être efficace quelles que soit les conditions du milieu rencontrées (Tasiemski et al., 2014). De plus, les polychètes, de par leur longue évolution, produisent des PAMs qui sont souvent très spécifiques d'une seule famille (comme pour la préproalvinellacine) ou même d'un genre ou d'une espèce particulière (préprohedistine). Ceci permet de faire l'hypothèse que l'évolution de ce type de molécules est doublement contrainte par l'environnement biotique et abiotique de l'espèce. A ce titre, il est intéressant de noter que cette famille de molécules est la seule à coupler l'agent anti-microbien à une molécule chaperonne de type BRICHOS. Ces molécules sont donc des modèles particulièrement intéressants pour étudier comment les pressions de sélection mises en jeu ont conduit à une telle association et quels sont les mécanismes génétiques qui assurent le bon fonctionnement de cet effecteur immunitaire dans un contexte biotique/abiotique original.

Dans ce contexte, le présent chapitre décrit en détail l'évolution d'un gène codant pour le précurseur protéique de l'alvinellacine. Par l'étude de la diversité génétique et l'évolution du/des gène(s) codant pour le précurseur protéique de l'alvinellacine, il s'agira tout particulièrement de répondre aux questions suivantes :

- A-t-on à faire à une famille multigénique complexe comme cela est largement documenté pour les gènes codant des effecteurs immunitaires (CMH)?
- Si le PAM appartient à une famille multigénique, les duplications sont-elles récentes ou postérieures à la séparation des espèces du genre *Alvinella* ?
- Si tel est le cas, l'évolution des paralogues est-elle adaptative? A-t-elle conduit à une forte diversification allélique du gène depuis l'acquisition de la symbiose ?

- Quelles sont les forces de sélection qui s'exercent sur le peptide antimicrobien et peut-on s'attendre à mettre en évidence l'action d'une sélection balancée qui est la sélection la plus décrite chez les effecteurs immunitaires (tels le CMH) surtout dans un contexte d'environnement changeant au moins du point de vue des paramètres physico-chimiques ?
- Les pressions de sélection sont-elles les mêmes sur tous les domaines du précurseur protéique, et y a-t-il une coévolution moléculaire entre le précurseur protéique et sa fonction chaperonne et le peptide antimicrobien lui-même ?
- Puisque l'échantillonnage inclus des populations de part et d'autre d'une barrière semi perméable aux flux de gènes, le/les gène(s) codant ce précurseur protéique présente(nt)-il(s) une différenciation génétique nord/sud ce qui indiquerait l'absence d'une introgression adaptative des allèles de part et d'autre de la barrière, et donc une ségrégation plutôt neutre des allèles existant.

Matériel et Méthodes

1. L'échantillonnage

Deux populations géographiquement différenciées (Figure 3) ont été échantillonnées pour cette étude chez les 2 espèces syntopiques d'*Alvinella,* au Nord et au Sud de la barrière équatoriale située au point triple des Galapagos (Hess Deep/Quebrada/Discovery/Gofar) :

(1) Les individus de la population nord proviennent du champ hydrothermal 9°50N/EPR à 2550m sur trois habitats très proches du site pVent (échantillons Périscope, SnowBall et pVent répartis sur une zone d'environ 100 m^2).
(2) Les individus de la population sud proviennent d'une zone située à plusieurs milliers de kilomètres du point précédent sur le champ hydrothermal 18°25S à 2640m à partir de l'échantillonnage d'une seule cheminée (site Fromveur).

Les deux populations sont séparées par 3000 km et une série de failles transformantes au niveau de l'équateur. Cette dorsale est composée d'un dôme régulier et est considérée comme une dorsale rapide, c'est-à-dire que l'épaisseur de la lithosphère à l'axe de la dorsale est comprise entre 1 à 2 km au-dessus du réservoir magmatique (Lagabrielle, 2005) et que le taux d'accrétion est supérieur à 10 cm/an. Du nord au sud plusieurs barrières à la dispersion ont été identifiées et plus ou moins localisées géographiquement. Dans le cas de l'espèce *A. pompejana*, (Plouviez et al., 2010) a montré que les populations nord et sud représentent

des espèces cryptiques potentielles avec cependant l'existence d'une barrière semi-perméable aux flux de gènes. Cette hypothèse a été récemment confirmée à l'échelle génomique par une analyse de données RADseq (A. Bioy, pers. comm.) et par une analyse multilocus (Jang et al., 2016).

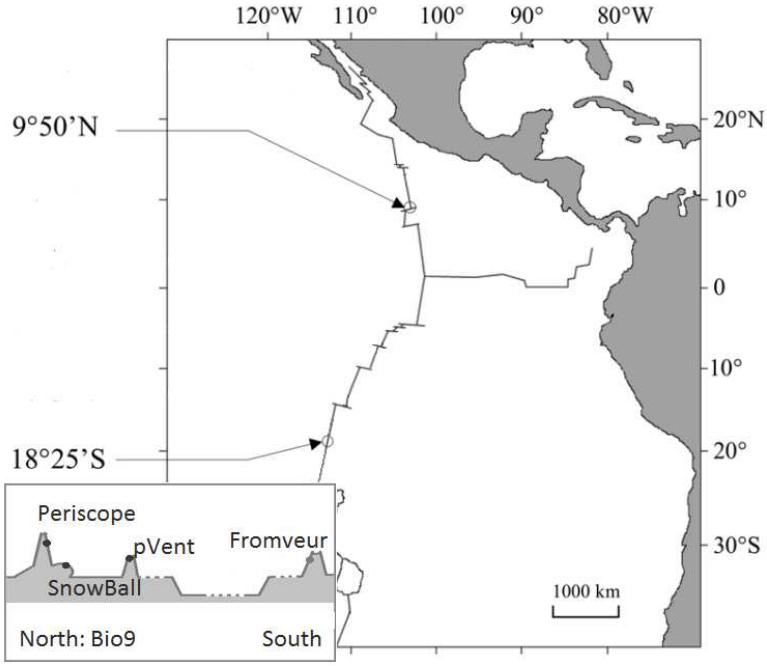


Figure 3. Localisation des populations d'*A. pompejana* et *A. caudata* échantillonnées le long de la dorsale est pacifique. Au Nord : les sites Périscope, SnwBall et pVent et au Sud de la barrière équatoriale : Fromveur.

Les animaux ont été prélevés à l'aide de la pince télémanipulée du Nautile durant les campagnes océanographique Mescal 2012 et Biospeedo 2004 à bord du navire hauturier de l'IFREMER : N/O l'Atalante. Les animaux ont été disséqués directement à leur remontée à bord et conservés à -80°C jusqu'à l'extraction de l'ADN génomique en laboratoire à l'aide du protocole CTAB (Doyle and Doyle, 1987).

2. Le précurseur protéique de l'alvinellacine : la préproalvinellacine.

Les peptides antimicrobiens matures sont clivés d'un précurseur protéique plus large qui contient un peptide signal et une prorégion (Boman, 2003). Cette prorégion pourrait empêcher l'activité antimicrobienne (et donc la cytotoxicité) du peptide associé et, en

conséquence, ce serait le clivage de ce précurseur protéique par une peptidase qui entraînerait son activation. L'étude du gène de l'alvinellacine par Tasiemski et collaborateurs (2014) à l'aide de PCR emboîtés ont permis de décrire la structure du gène de la préproalvinellacine, de caractériser son ADN complémentaire et de modéliser sa structure tertiaire. Celles-ci sont résumées dans la Figure 4. Au niveau de la prorégion, on trouve un domaine conservé appelé domaine BRICHOS (en jaune). Jusqu'à présent, ce domaine d'une centaine d'acides aminés n'a jamais été identifié dans un précurseur de PAM autre que celui de la préproarénicine qui est un PAM de la même famille que l'alvinellacine (à savoir les arenicine-like avec la capitellacine et arénicine) (Willander et al., 2011). Le gène codant pour le précurseur protéique est constitué de 6 exons et 5 introns, il fait 2 kb et la région codant le peptide mature de 22 acides aminés se trouve en région 3'. Les régions exoniques font respectivement, 57, 180, 282, 45 et 66 paires de bases pour respectivement le peptide signal, la prorégion, le domaine BRICHOS, le linker et la région du PAM. Ceci correspond à 19, 60, 94, 15 et 22 acides aminés pour chacune des régions.

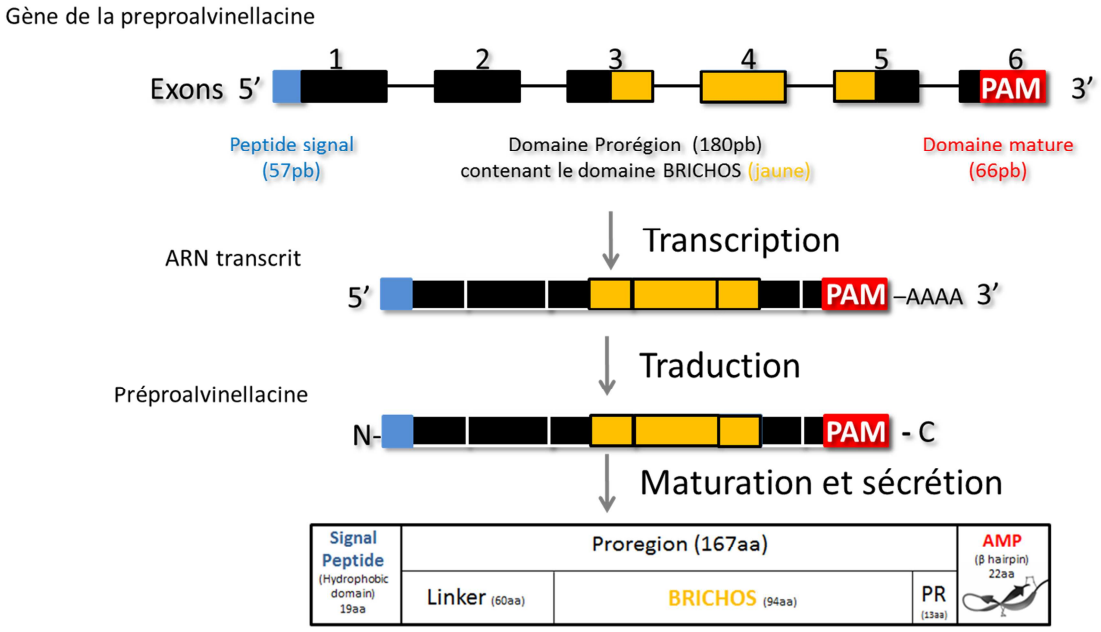


Figure 4. Structure du gène codant la préproalvinellacine et structure du précurseur protéique (en bleu : peptide signal, en noir : prorégion, en jaune : domaine BRICHOS, en rouge : peptide antimicrobien).

3. Acquisition des données

3.1. Obtention des séquences de la préproalvinellacine à partir de l'ADNg

Pour les deux espèces (4 et 96 individus pour *A. caudata* et *A. pompejana* respectivement), le gène d'intérêt (faisant 2 kb) a été divisé en deux parties égales pour être amplifié en PCR dans sa globabilité. Pour cela, des amorces ont été définies pour chaque partie (5' en gris et 3' en vert ; Figure 5 et Tableau 1).

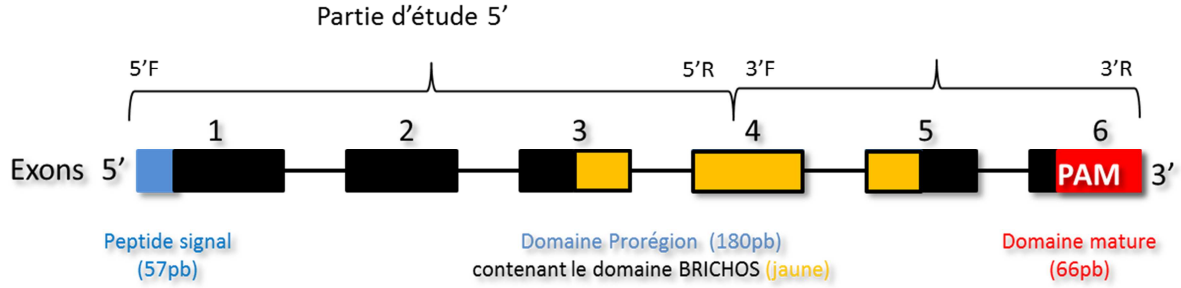


Figure 5. Emplacement des amorces utilisées pour amplifier les 2 fragments du gène codant la préproalvinellacine chez les deux espèces d'Alvinellidae

La partie 5' (3 exons + 3 introns) comprend donc le peptide signal et le début de la prorégion avec une partie de la région du BRICHOS (26 acides aminés) alors que la partie 3' comprend quant-à-elle la fin du domaine BRICHOS, la fin de la prorégion (aussi appelé linker) et la région codant pour le peptide antimicrobien (3 exons + 2 introns).

Comme le gène codant pour la préproalvinellacine est un gène nucléaire autosomal, les individus séquencés présentent deux allèles pour chaque copie du gène. Pour obtenir la séquence complète d'au moins un des deux allèles présents au sein de chaque individu, les produits d'amplification par PCR de chaque individu doivent être clonés. La technique de clonage des allèles étant trop longue et trop coûteuse pour être effectuée individuellement, la technique de marquage-clonage-recapture des allèles développée par (Bierne et al., 2007) a été utilisée pour l'espèce *A. pompejana*. Cette technique permet d'identifier les allèles *a posteriori* pour chaque individu après un clonage unique des produits de PCR obtenus sur l'ensemble des individus. Pour cela, le gène cible est amplifié séparément sur chaque

individu à l'aide d'un couple d'amorces présentant un tag spécifique de 5 bases supplémentaires dans leur partie 5'. Après amplification, les produits de PCR sont mélangés de façon équimolaire et clonés en une seule fois (Tableau 1 et Figure 6).

Pour maximiser la recapture des allèles, par partie du gène, un clonage a été effectué pour un mélange de 16 individus (soit 32 allèles) avec un effort de recapture qui a été fixé à 3 (soit trois fois plus de clones à séquencer que d'allèles présents dans le mélange). Ainsi, pour chaque partie du gène, 6 pools de 16 individus ont été effectués et 96 clones ont été sequençés par pool. Ceci donne 96*2 (F et R), 192 séquences par pool donc 1152 séquences par région (1152*2=2304 séquences pour tout le jeu de données).

Pour l'espèce *A. caudata*, un clonage individuel a également été effectué pour les 4 individus (2 au nord, 2 au sud) avec un effort de recapture de 4 (32clones pour 4 individus : 64 F et R, 128 séquences pour les deux régions au final).

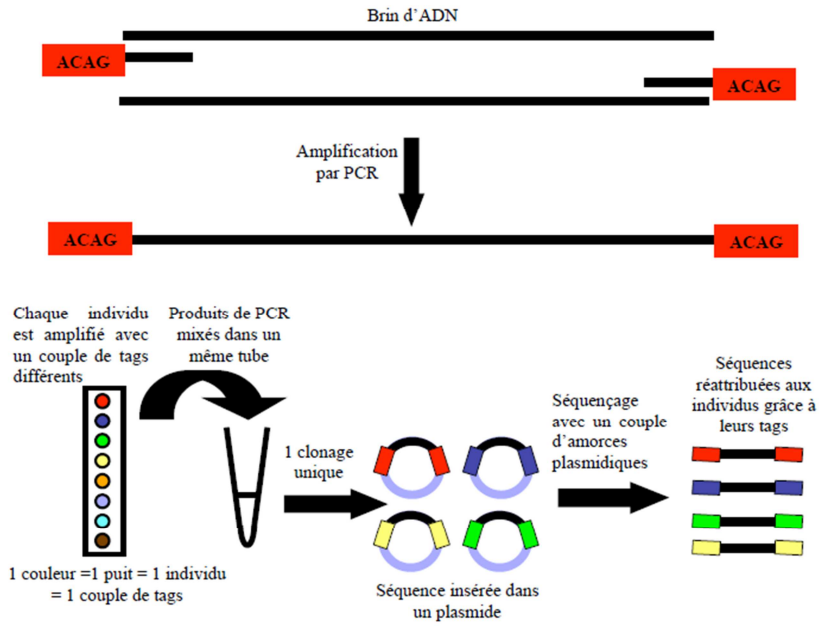


Figure 6. Schéma de la technique Marquage-Clonage-Recapture (MCR) de Bierne et al., 2007. Repris de la thèse de doctorat de S. Plouviez (2008).

Les amplifications ont été réalisées dans un volume final de 25µL avec: 1X de tampon, 2mM $MgCL_2$, 0.05mM dNTP, 0.4µM de chaque primer, 1U de Taq polymerase (Uptitherm, Interchim^{TM}). Les cycles d'amplification ont été les suivants : 96°C pendant 4 min, suivi de 40 cycles à 96°C pendant 30 s, 60°C pendant 45 s et 72°C pendant 2 min, avec une élongation

finale de 10 min à 72°C. Le vecteur de clonage utilisé est le TA cloning kit (Invitrogen) utilisant des bactéries compétentes *E. coli* top10. Le clonage a été effectué selon le protocole fourni du fournisseur (TA cloning kit, INVITROGEN). Après un criblage bleu/blanc, l'insert contenu dans le plasmide des clones positifs obtenus a été amplifié avec les amorces M13 et M13rev (Tableau 1) pour vérifier qu'il possède bien la taille attendue. Après vérification, les produits PCR ont été purifiés avant séquençage en utilisant des plaques « Multiscreen 96-plate purification kit » de Millipore puis séquencés dans les deux sens sur le séquenceur à capillaires ABI3100 suivant le protocole du BigDye Terminator du fournisseur.

3.2. Traitements des échantillons avant analyse

Après séquençage, les séquences sont alignées, corrigées et réattribuées aux individus sur la lecture des tags. La qualité des séquences est vérifiée visuellement à l'aide des chromatogrammes en utilisant le module 'De Novo Assemble' du logiciel Geneious (Drummond et al., 2010) qui permet d'assembler les séquences obtenues dans les 2 sens afin d'établir une séquence consensus pour chaque allèle. Les alignements entre individus sont ensuite effectués en utilisant le module 'PairWise/Multiple Alignment' de Geneious.

3.3. Elimination des recombinants artéfactuels dans le jeu de données

La technique MCR possède l'inconvénient de créer des recombinants artéfactuels au moment de la PCR d'amplification (création de chimères entre les 2 allèles d'un même locus, ou entre des allèles de locus différents dans le cas de famille multigénique) ou lors de la ligation de l'insert (cas beaucoup plus rare). Chaque individu étant identifié *a posteriori* par une combinaison de tags (2 x 5 bases) préétablie, il est en premier lieu facile d'éliminer toute combinaison de tags non attendue qui constitue alors une recombinaison artéfactuelle de 2 allèles lors de la ligation et d'enlever ceux-ci du jeu de données.

Dans un deuxième temps, il convient de traquer les recombinants artéfactuels entre allèles d'un même individu par un traitement des lots de séquences recapturées au sein de chaque individu à l'aide du logiciel RDP 4.0 : Recombinant Detection Program (Martin and Rybicki, 2000). La procédure Chimera, méthode préconisée dans le but de détecter des évènements de recombinaison entre séquences d'ADN (Posada and Crandall, 2001), a été utilisé avec les

paramètres laissés par défaut. Cette étape est effectuée pour des individus ayant plus de 16 recaptures d'allèles et exposée plus en détails dans les résultats.

La technique est la suivante :

Mode opératoire pour l'analyse et la suppression des recombinants artéfactuels.

1. Identification des individus les plus recapturés : au final, 10 individus avec plus de 20 séquences).
2. Alignement des séquences Forward et Reverse pour chaque allèle recapturé.
3. Production de la séquence consensus pour chaque allèle puis alignement des allèles entre individus
4. Alignement passé sous RDP : ce logiciel permettant de détecter des recombinants entre séquences d'ADN présente dans le jeu de données.

Pour les 10 individus séparément, lorsque le logiciel détecte pour une séquence des allèles parentaux qui existent dans le jeu de données, celles-ci sont supprimées du jeu de données – après vérification visuelle sur l'alignement « brut » en parallèle sur GENEIOUS. Il existe des cas où le logiciel suggère des évènements de recombinaisons avec des allèles parentaux inconnus dans le jeu de données « unckown », ceux-ci ne sont pas supprimés mais notés pour vérification plus tard dans le but de déterminer si l'allèle parental existe chez d'autres individus mais n'a juste pas été recapturé chez l'individu en question.

Les allèles des 10 individus (purgés donc des recombinants artéfactuels) sont ensuite alignés tous ensemble et cet alignement est également passé sous RDP avec le même paramétrage pour détecter d'éventuels recombinants restants avec des points de recombinaisons différents. En effet, si des recombinants montrent les mêmes points de recombinaisons entre deux individus différents, ceux-ci sont gardés. Sinon, ils sont supprimés du jeu de données.

Dans le cas des séquences recapturées presque à l'identique en double (à un ou deux singletons près), une seule a été gardée et le nombre de mutations/singletons qui séparent ces deux séquences a ensuite permis de calculer un pourcentage et de l'appliquer à tout le jeu de donnée final. Ainsi, une analyse de suppression des mutations artéfactuelles est

également réaliséé manuellement sur la base du nombre de singletons attendus au sein d'un individu.

3.4. Caractérisation des gènes paralogues : étape de génotypage

Chez l'espèce *A. pompejana* pour laquelle la diversité génétique a été plus spécifiquement étudiée, plusieurs allèles sont généralement retrouvés au sein d'un même individu. Ceci laisse supposer que ce peptide est codé par un mélange de plusieurs gènes paralogues issus d'évènements de duplication en tandem. Une dernière étape de validation/caractérisation des différents gènes est nécessaire. En général, les différentes lignées alléliques d'un même locus se regroupent par clade, ces derniers présentant plus de ressemblance entre eux qu'avec les allèles d'une autre copie du gène (sauf dans le cas de la sélection balancée et/ou d'une rétention de polymorphisme ancestraux lorsque les lignées alléliques pré-datent les évènements de duplication). Néanmoins, il convient de vérifier si chaque clade correspond bien à un gène particulier en effectuant un génotypage des individus à partir d'amorces spécifiques desdits clades (signatures divergentes sur le gène) qui amplifieront des régions contenant des mutations diagnostiques au locus présupposé (Génotypage : Tableau 1).

Cette étape de vérification basée sur la recherche d'individus homo- et hétérozygotes aux sites diagnostiques est réalisée par un séquençage direct du produit d'amplification par PCR dans la région choisie. Ceci permet de confirmer ou non si les clades identifiés par 'clustering' sont bien des gènes autosomaux (avec présence de double pics aux sites diagnostiques sur les individus hétérozygotes), et ne mélangent pas des gènes paralogues (ou tous les individus sont hétérozygotes) ou ne constituent pas des lignées alléliques (ou tous les individus sont homozygotes).

Alvinella pompejana		
	3' Forward	ATCGTGTTACGTCATGGGTGGCCTTG
	3' Reverse	CTCAGTGAAATGAAGCAGGTGAGTTATG
	5' Forward	ATGACGTATTCTGTAGTTGTGACGCTGGTC
	5' Reverse	ATCCGGTAAGATCGTCGTAAATGGCTCC
Genotypage		
	P1_Forward	ACATCTACAGATTGGTGCTATCGAC
	P2_Forward	CTACAGATTGGTGCAGCCGAC
	P3_Forward	CATCTACAGATTGGTGCTGTGGAT
	P4_Forward	AACAGATTGGTGCTGTCGCC
	P5_Forward	TTTACATAGATTGGTGTTTCTTCTCTGAG
	P1_Reverse	GTTGAGGTGGCCAGCTGC
	P2-Reverse	GTTGGGGTGGCCAACTGC
	P3_Reverse	ATGTTGGGGTGATCAGCTGC
	P4_Reverse	GATGTTGAGGTGGCCAGCTAT
	P5_Reverse	GTTTCATGAAATGTGGCAGATG
Alvinella caudata		
	5' Forward	GTTACGTATTCTGTAGTCACGACGCTG
	5' Reverse	GGTAAGATCGTCGTAAATGGCTCC
	3' Forward	GTCGTGTTACCTGATGGGTGGC
	3' Reverse	AATATGCCAAAACAGGCGAATTACG

Tableau 1. Amorces utilisées pour séquencer le précurseur protéique chez les deux espèces d'Alvinellidae et pour effectuer le génotypage sur un fragment du gène de l'alvinellacine chez l'espèce *A. pompejana*.

Des amorces locus-spécifiques ont donc été utilisées dans une région diagnostique du gène (en région 5') pour des locus supposés paralogues avec un TM élevé de 60°C et une longueur d'amplicon inférieur à 350pb pour éviter tout problème de recombinaison artéfactuelle. Toutes les amorces utilisées lors de cette étude sont récapitulés dans le Tableau 1.

3.5. Etape supplémentaire

Une étape supplémentaire de caractérisation des transcrits à partir de l'extraction d'ARN à partir de plusieurs tissus *d'A. pompejana* a été effectuée pour vérifier que tous les gènes paralogues sont transcrits. Cette étape et les résultats obtenus sont présentés en Annexe 1.

4. Méthodes d'analyses

4.1. Estimateurs de diversité génétique et tests statistiques de neutralité

L'estimation des indices de diversité et les tests de neutralité qui leur sont associés ont été réalisés grâce au logiciel DNAsp v5.0 (Librado and Rozas, 2009). Nous avons plus particulièrement calculé le nombre de sites polymorphes S, la diversité nucléotidique θ_π, la diversité nucléotidique à partir de S à l'aide du theta de Watterson (θ_w) et la diversité haplotypique Hd. Ces estimateurs ont été comparés par le test de Tajima (D) et le test de Fu & Li (F) pour évaluer si l'accumulation des mutations dans le polymorphisme d'un gène s'effectue de façon neutre.

- Le nombre de sites polymorphes S dépend à la fois de la longueur de la séquence et de la taille de l'échantillon (nombre de séquences comparées) ce qui rend difficile sa comparaison entre locus et échantillons et représente seulement le nombre de sites ségrégeant dans un échantillon représentatif d'allèles trouvés dans une population.
- L'estimateur S (ou Eta) permet d'estimer le théta de Watterson (θ_W) en rapportant le nombre de sites variables S dans l'échantillon à la longueur de l'alignement (nombre de sites polymorphes normalisés par la taille de l'échantillonnage (Watterson, 1975). Cet estimateur est égal à $4Ne.\mu$ (Ne la taille efficace de l'espèce et μ le taux de mutation spécifique à chaque gène) sous l'hypothèse d'une accumulation neutre des mutations dans le polymorphisme.
- La diversité nucléotidique π représente le nombre moyen de différences nucléotidique par site et par paire de séquences (Tajima, 1983). Cet estimateur noté π est lui aussi équivalent à $4Ne.\mu$ sous l'hypothèse d'une évolution neutre mais est beaucoup moins sensible à un effet sélectif (balayage) et/ou démographique passé.

Les tests statistiques de Tajima (Tajima, 1989) et Fu& Li (Fu and Li, 1993) ont été utilisés dans le but de détecter un écart à la neutralité sous l'hypothèse d'équilibre mutation dérive. Le modèle d'accumulation neutre des mutations est basé sur le modèle de Wright-Fisher où il existe un état d'équilibre entre les mutations apparaissant dans la population à chaque génération et l'effet de dérive génétique qui élimine certaines mutations aléatoirement à chaque génération.

L'estimateur de Tajima, D, compare deux mesures de diversité citées précédemment : θπ et θw. En effet, comme ces 2 indices estiment le même paramètre $4N_e.\mu$, Tajima (1989) a proposé un premier test de neutralité basé sur la différence des estimateurs θw et θπ sachant que θw est plus sensible à l'accumulation de variants rares (ou singletons). En cas d'évolution neutre, la différence pondérée par la variance (θπ-θw)/σ2(θπ-θw) ou D de Tajima est égal à zéro. Dans le cas d'une population en expansion et/ou d'un balayage sélectif (i.e. fixation d'un allèle avantageux), cette valeur est fortement négative. Dans le cas d'une population subissant un goulot d'étranglement et/ou un effet de sélection balancée (maintien de plusieurs lignées alléliques), cette valeur est positive.

Fu & Li (1993) ont ensuite proposé une amélioration du D de Tajima. Ce test est en effet plus sensible pour détecter les changements démographiques passés en estimant la diversité nucléotidique en ne prenant en compte que la catégorie des mutations récemment apparues (singletons), estimé par Eta (η_e).

4.2. Description de la diversité génétique le long du gène

Pour les deux espèces soeurs *A. pompejana et A. caudata*, la diversité nucléotidique π a été calculée tout au long du gène (exons et introns) en estimant celle-ci dans une fenêtre mobile (50pb, pas de 10pb) à l'aide du logiciel DNAsp en regardant plus spécifiquement

- la diversité nucléotidique moyenne intra-clade (moyenne des diversités estimées pour chaque locus)
- et la diversité nucléotidique globale (diversité estimée sur l'ensemble du jeu de données) qui prend principalement en compte la divergence des séquences entre les clades.

4.3. Réseau et arbre de coalescence.

Après correction du jeu de données, nous avons effectué une reconstruction des relations phylogénétiques des différents allèles recapturés au niveau de la région 3' sur 4 et 8 individus de chaque espèce (*A. caudata et A. pompejana* respectivement) et de tous les allèles recapturés au niveau de la région 5' chez l'espèce *A. pompejana*. Le choix de prendre la région 5' ou 3' a été dicté par le nombre d'allèles amplifiés et notre capacité à discriminer les différents clades au niveau de l'alvinellacine. En effet, des difficultés d'amplification de la

région 5' ont été rencontrées chez l'espèce *A. caudata*, ce qui a entraîné un jeu de données plus conséquent pour la région 3'. Chez l'espèce *A. pompejana*, les différents clades d'allèles étaient plus facilement distinguables au niveau de l'intron 1 et des parties introniques de façon générale au niveau de la région 5'.

Pour l'identification des paralogues chez les deux espèces en région 3', l'arbre a été réalisé sous MEGA 6.0 en Maximum de vraisemblance et avec le modèle de substitution K2P (Kimura, 1980) pour représenter les relations phylogénétiques entre les différents allèles des deux espèces sœurs. A l'aide du logiciel MEGA 6.0, les régions exoniques de l'alignement ont été définies dans un premier temps puis les différents clades –analyse visuelle- ont également été indiqué dans le but de calculer la divergence nette par groupe de séquences (divergence inter-clade intra-espèce). Dans le cas de l'espèce *A. pompejana*, une analyse a été effectuée avec le logiciel SplitsTree qui permet mieux de rendre compte la recombinaison inter-génique dans les relations inter-alléliques en utilisant l'algorithme NeighborNet (Huson, 1998).

4.4. Choix du modèle de substitution par jModelTest pour la construction de l'arbre phylogénétique à partir de la région 5' pour l'espèce A.pompejana

A partir de l'alignement des individus les plus recapturés pour la région 5', une sélection du meilleur modèle de substitution a été effectuée grâce au logiciel jModelTest 2.1.7 (Darriba et al., 2012) à l'aide des deux critères d'ajustement AIC et BIC et en utilisant le hLRT (hierachical Likelihood Ratio Test) pour construire la phylogénie des allèles. Une fois le modèle choisi, deux arbres récapitulatifs des relations phylogénétiques entre les différents clades ont été inférés à l'aide des logiciel PhyML et MEGA 6.0 (pour comparaison) dans le but de résumer au mieux les relations phylogénétiques qui existent entre les allèles ainsi que les évènements de recombinaisons naturelles entre paralogues.

4.5. Recherche de sélection positive le long du précurseur protéique

Cette partie s'intéresse uniquement aux régions exoniques du gène en mesurant l'intensité de la sélection qui s'exerce en certains points (domaines) du gène en calculant le ratio (d_N/d_S ou Ka/Ks) à partir du taux de mutations non-synonymes par site non-synonyme (d_N) et du taux de mutations synonymes par site synonyme (d_S). Une valeur supérieure à 1 indique l'action de la sélection positive (excès local de mutations NS souvent en déséquilibre de

liaison) alors qu'une valeur inférieure à 1 indique l'action de la sélection purifiante (élimination des mutations non-synonymes délétères).

Les analyses

- Le long du gène, les ratios ont été calculés à la fois entre paralogues (d_N/d_S) et au niveau du polymorphisme de chaque paralogue ($\pi a/\pi s$) toujours en utilisant le logiciel DNAsp 5.0 avec une fenêtre mobile (largeur 50pb avec un pas d'échantillonnage toutes les 10pb).

- Dans le but de détecter la trace d'une sélection positive dans un domaine particulier du précurseur protéique, une séquence consensus (multi-individus) de chaque paralogue a été construite pour chaque domaine. Ces séquences ont ensuite été utilisées par paire pour calculer un d_N/d_S par domaine en utilisant le module yn00 du logiciel PaML (Yang, 2007).

4.6. Cartographie des remplacements en acides aminés au sein des domaines BRICHOS et prorégion.

Les séquences consensus obtenues pour les régions du BRICHOS et de la prorégion ont été alignées à l'aide du logiciel GENEIOUS en excluant toutes les autres régions de l'alvinellacine chez l'espèce *A. pompejana* en prenant une séquence d'*A. caudata* en groupe externe (outgroup) afin d'orienter les mutations.

Une sélection du modèle de substitution le plus adéquat a été réalisé pour construire la phylogénie des allèles du BRICHOS à l'aide du logiciel jModelTest. L'arbre en Maximum Likelihood généré à l'aide du logiciel MEGA 6.0 a ensuite été utilisé pour faire une reconstruction des séquences ancestrales des 2 domaines pris séparément afin de cartographier les changements en acides aminés entre paralogues au cours de l'évolution du gène en utilisant le module aaML du logiciel PAML 4.0 (Yang 2010). Les probabilités bayésiennes pour chaque modification en acide aminés ont également été données. Finalement, pour chaque position polymorphe chez *A. pompejana*, la nature de l'acide aminé correspondant chez l'espèce sœur *A. caudata* a été utilisée pour obtenir une indication quant à l'état ancestral ou dérivé de la mutation observée.

4.7. Recherche de sélection positive au sein des domaines Prorégion et BRICHOS

L'analyse CodeML (sous PAML) a pour but de détecter d'éventuelles lignées ou branches de l'arbre sous sélection positive (modèles de branches) ou catégoriser l'intensité de la sélection sur chaque codon au sein d'un gène (modèles de sites) et permet de tester l'hypothèse d'une évolution non neutre de certaines lignées en comparant les valeurs de vraisemblance de différents modèles de substitution de codons (Yang, 2007). Cette analyse permet en effet de détecter l'action de pression de sélection (sélection positive) sur certains codons en comparant ce type de modèle à celui d'un relâchement des pressions de sélection où tous les codons sont alors libres d'évoluer de façon « presque neutre ». Dans le cas des évènements de duplication, la théorie prédit qu'une des copies peut alors évoluer sans contrainte voire vers une nouvelle fonction puisque l'autre copie assure la fonction initiale de la protéine. Nous proposons de tester l'hypothèse de néofonctionalisation des domaines BRICHOS et PROREGION via la comparaison de différents modèles emboîtés qui seront ajustés à nos données de séquences sous la contrainte d'une topologie de gènes paralogues préalablement définie (user tree). Ainsi, l'une des premières étapes de cette analyse consiste à fournir un arbre de référence ('user tree') dont la topologie correspond à l'histoire présumée du gène.

Pour le domaine BRICHOS, l'alignement obtenu après détermination du meilleur modèle de substitution avec jModelTest pour l'analyse aamL a également été utilisé pour générer un arbre de référence utilisé dans l'analyse CodeML (paramétrage : seqtype =1 : codons ; CodonFreq= F3x4; model =0, un seul d_N/d_S par branche; NSsites = modèle d'évolution testé) and ncatG = 10, Rateancestral=1). Cette analyse permet également de reconstituer les séquences ancestrales à chaque nœud avec leurs probabilités bayésiennes d'occurrence.

Pour la prorégion, l'analyse et les paramétrages sont les mêmes sauf que le modèle de substitution (sélectionné par jModeltest) utilisé pour construire l'arbre de référence est différent de celui utilisé pour le domaine BRICHOS.

Les modèles de sites testés sont présentés dans le Tableau 2 :

- Un premier test compare le modèle 'presque neutre' appelé M_1 pour lequel les codons sont classés selon deux catégories : les codons neutres (avec un $d_N/d_S=1$) et les codons sous sélection purifiante ($d_N/d_S<<1$) à un modèle de sélection M_2 pour

lequel une troisième catégorie est ajoutée : celle des codons sous sélection positive ($d_N/d_S>1$).

- Un deuxième test compare 2 autres modèles emboîtés appelés M_7 et M_8, analogues aux modèles précédents mais pour lesquels la variable ω (d_N/d_S) n'est pas catégorisée mais se distribue selon une loi beta. En effet, Le modèle M_7 fait varier ω selon une loi beta (p,q) discrétisée K fois (K étant fixé à 10) et, le modèle M_8 ou ω varie selon une loi beta discrétisée K fois avec cependant une classe supplémentaire dans laquelle ω > 1 (autorise la sélection positive).

	Paramètres
M_1 (neutre)	p_0 (p1=1-p_0), $\omega_0<1$, $\omega_1=1$
M_2 (sélection)	p_0, p_1, p_2=1-p_0-p_1), $\omega_0<1$, $\omega_1=1$, $\omega_2>1$
M_7 (beta)	p,q
M_8 (beta & ω)	p_0, p_1=1-p_0, p, q, ωs>1

Tableau 2. Récapitulatif des paramètres caractérisant les différents modèles implémentés sous PAML pour détecter l'action de la sélection positive

Les modèles emboîtés sont ensuite comparés statistiquement grâce à un 'Likelihood Ratio test' (LRT) et le plus vraisemblable est choisi si sa valeur de vraisemblance (LnL) est significativement la moins élevée. La comparaison de ces valeurs d'ajustement à l'arbre permet ensuite de déterminer lequel des deux modèles s'ajuste le mieux aux données de séquences. Le test de vraisemblance (LRT) est un test de χ2 (avec $\chi2= 2(LnL(M_1) - LnL(M_2))$ pour comparer les deux modèles emboîtés choisis avec un degré de liberté équivalent à la différence entre les nombres de paramètres estimés dans les 2 modèles.

5. Induction du gène de la préproalvinellacine en réponse à un stress abiotique

Une quantification par PCR en temps réel (qPCR) de l'expression relative du gène codant la préproalvinellacine et d'un gène inductible HSP70 du polychète *A. pompejana* a été effectuée à partir de lots d'individus expérimentés à différentes pressions hydrostatiques et différentes températures à la pression du fond. Un premier lot d'individus dépressurisés a été comparé à des individus gardés à la pression de leur habitat (250 bars). L'expression des gènes a été quantifiée sur des animaux dépressurisés à l'issue de leur remontée en surface

et comparés à ceux d'individus re-pressurisés pendant 12 heures à 250 bars dans des enceintes hyperbares de type IPOCAMP. Une deuxième expérience à différentes températures a également été effectuée sur des individus remontés et gardés sous pression avec un prototype de récolte sous pression. Cet équipement appelé BALIST pour « *Biology of Alvinella, Isobaric Sampling and Transfer* » a donc permis de récupérer des animaux sur le fond, de les remonter et les transférer sous pression à bord dans un aquarium expérimental pour réaliser trois chocs thermiques de 2 heures à 20, 42, 54°C suivi d'une récupération des animaux à 20°C pendant 2 nouvelles heures. Ces expérimentations ont été réalisées lors de la campagne Mescal 2012 et l'amplification en temps réel réalisée par la suite à terre après l'extraction de l'ARN total des spécimens étudiés. Les procédures d'extractions et d'amplifications réalisées à bord par l'équipe sont indiquées dans l'article de Ravaux et al., 2013.

6. Analyse différenciation génétique nord/sud de l'EPR

Après assignation des allèles aux différents gènes paralogues avec les techniques présentées précédemment, un inventaire des allèles observés chez les deux paralogues les mieux recapturés entre individus de l'espèce *A. pompejana* (paralogues 1 et 4) a été effectué en retournant sur l'ensemble des séquences obtenues par la technique marquage-recapture de Bierne et al. (2007). Une séquence consensus par allèle a été définie puis un allèle par individu a été tiré au hasard pour obtenir un alignement d'allèles pour chaque paralogue en prenant soin d'éliminer de nouveau les recombinants artéfactuels au sein de chaque individu. Une analyse génétique de la différenciation des individus a ensuite été entreprise en fonction de leur provenance géographique (N=nord ; S=sud).

Dans le but de détecter l'effet de la barrière équatoriale aux flux de gènes décrite par Plouviez et al. 2009 et Jang et al. 2016 sur la dispersion des allèles du gène de la préproalvinellacine, un arbre phylogénétique et un réseau d'haplotypes ont été réalisés sur les paralogues 1 et 4 et le niveau de différenciation génétique entre les populations sud et nord a été testé via des indices de fixation tels le F_{st} ou le Phi_{st} sur la base des fréquences des différents allèles au sein des 2 localités.

Ainsi, à partir des alignements finaux des séquences alléliques (un allèle par individu), la généalogie des allèles observés au sein et entre populations a été visualisée sous MEGA6

(Tamura et al., 2007) sous la forme d'un arbre selon la technique du Minimum Evolution en mode 'Complete deletion' (modèle K2P, 100 bootstraps). Bien que le nombre d'allèles soit faible, un réseau d'haplotypes a été également construit. L'origine géographique de chacun des haplotypes est visualisée grâce à un code couleur sur le réseau qui permet donc de mieux comprendre le rôle de la géographie sur la structure des populations. Ces réseaux ont été effectués en Median joining à l'aide du logiciel PopArt (Leigh and Bryant, 2015)

Les indices de différenciation génétique (Phi_{st} et F_{st}) ont été calculées pour chaque paralogue entre les 2 populations via le logiciel Arlequin version 3.0 (Excoffier et al., 2005). Phi_{st} utilise chaque site polymorphe d'une séquence comme un locus et intègre donc un signal de différenciation sur une histoire évolutive plus longue plus en lien avec l'histoire tectonique de la dorsale. L'index F_{st} est estimé sur la base des fréquences haplotypiques et est plus sensible à la dérive d'échantillonnage avec de petits effectifs. Il permet néanmoins de détecter plus efficacement des effets d'isolement lorsqu'ils sont récents dans la mesure ou les fréquences des haplotypes sont correctement estimées. Un test sur la significativité de l'écart de ces indices vis-à-vis de zéro (absence de différenciation) est réalisé sur 1000 permutations des allèles entre populations par le même logiciel. Un test exact de différenciation est également effectué (H_0= non différenciation entre populations sous hypothèse de panmixie) sur la base de la distribution des fréquences des haplotypes.

Dans le but de déterminer si les populations ont été séparées depuis longtemps, la divergence nette entre les deux populations nord et sud (Dxy, proportion moyenne de différences nucléotidiques entre populations) a été calculé clade par clade entre paires de séquences de populations différentes.

RESULTATS

1. Histoire phylogénétique de l'alvinellacine dans la famille des Alvinellidae

Afin de retracer l'histoire évolutive du gène de la prépropalvinellacine à l'échelle de la famille des Alvinellidae, la séquence codante de la préproalvinellacine d'*A. pompejana* a été blastée par tBLASTx sur les transcriptomes assemblés sous Trinity à partir de données RNAseq d'annélides terebellomorphes dont 7 espèces d'Alvinellidae. A cela s'est ajouté les séquences déjà connus du prépropeptide de l'arénicine et de la capitellacine qui appartiennent à la même famille de PAM à BRICHOS comme 'outgroups'. La recherche de séquences homologues a permis de trouver plusieurs séquences codant pour ce prépropeptide chez toutes les espèces testées avec une très forte diversité de la séquence primaire du peptide mature entre les différentes espèces analysées. Les parties conservées des séquences traduites obtenues (prorégion + BRICHOS) ont été alignées à l'aide du logiciel Geneious et un arbre phylogénétique a été effectué à l'aide du logiciel MEGA6 en utilisant la méthode de Maximum Likelihood et la matrice JTT de substitution d'acides aminés. Cet arbre est présenté dans la Figure 7 en faisant une distinction entre les espèces vivant en milieu hydrothermal (rouge, orange), celles vivant en milieu polaire (bleu) et celles vivant en intertidal tempéré (marron). Il apparaît que pour les espèces hydrothermales, les séquences du précurseur protéique se regroupent en deux clades avec une valeur de bootstrap robuste (>92) qui correspondent aussi à la présence d'un PAM plutôt long ou plutôt court. Dans le clade contenant le PAM alvinellacine *sensu stricto* (entouré en noir), les deux espèces sympatriques *Alvinella pompejana* et *Alvinella caudata* se regroupent ensemble par rapport aux espèces *Paralvinella*. De plus, dans chaque clade, les relations phylogénétiques entre espèces d'Alvinellidae sont conformes aux attendues de la classification des différentes espèces établie par (Desbruyeres and Laubier, 1991). Notamment on observe un regroupement des espèces sœurs *P. sulfincola* et *P. fijiensis* qui, comme *A. pompejana* et *A. caudata*, sont morphologiquement ressemblantes.

Du point de vue de la structure primaire, la forme de l'alvinellacine la plus courte (20-22 aa) est généralement composée de 4 cystéines qui forment des ponts disulfures et lui confère une structure en épingle. La seconde forme d'alvinellacine est beaucoup plus longue (36-38 aa) et est plus riche en acides aminés aromatiques et en cystéines laissant présager des structures secondaires/tertiaires plus complexes avec une augmentation du nombre de

ponts disulfures, suggérant un mode d'action différent dans la lutte antimicrobienne. Les séquences trouvées chez l'amphitritidae antarctique sont quant à elles dépourvues de cystéines et pourraient avoir perdu leur activité antimicrobienne. Ces résultats suggèrent l'existence d'une duplication ancestrale s'étant produite très tôt dans l'histoire des annélides terebellomorphes et qui n'a pas été retrouvé dans le génome de *Capitella*.

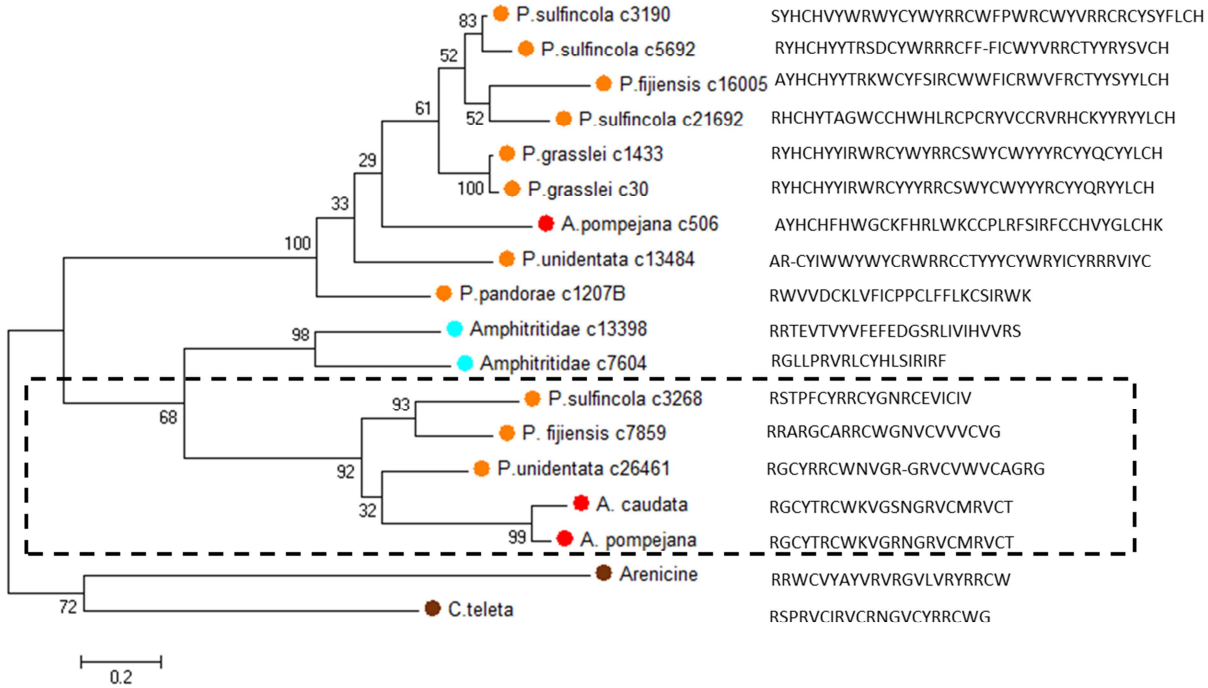


Figure 7. Analyse phylogénétique des relations entre les différents précurseurs protéiques retrouvés par tBLASTx à partir de la séquence de la préproalvinellacine, *sensu stricto*, sur les transcriptomes d'autres espèces terebellomorphes. La pastille de couleur correspond au type d'habitat dans lequel l'espèce est retrouve (rouge : chaud ; bleu : froid, marron : espèces outgroup côtières).

2. Amplification du gène et recapture des allèles

L'amplification de la partie 5' fournit un produit d'amplification ayant une taille approximative de 1300pb (comprenant un premier intron relativement long de 449pb) alors que le produit d'amplification de la région 3' est beaucoup plus court, soit à peu près 700pb (1949pb au final). Ceci est une moyenne puisque chaque amplification est composée de plusieurs allèles ayant une taille légèrement différente. En effet, la taille des différents gènes varie légèrement (une centaine de pb) et cette différence est principalement due à des délétions dans les régions introniques. Aucune délétion n'est décrite dans les régions

exoniques sauf dans le cas du paralogue 2 qui montre deux délétions dans la séquence codante : une de 10 codons et une autre de 22 codons dans cette région 5' sans jamais changer le cadre de lecture (pas de décalage en région 3').

Pour tester si le nombre d'allèles détectés correspond à plusieurs copies d'un même gène qui aurait été dupliqué en tandem des tests d'amplification par PCR du gène ont été effectués chez l'espèce *A. caudata* avec des temps d'élongation plus long (2min30). La Figure 8 montre les résultats de ces amplifications sur gel d'agarose et révèle l'existence d'au moins 3 copies du gène avec des longueurs attendues de produits de PCR de 1300 (région 5'), 3300 (gène entier de 2000pb + region 5') et 5300 bp (etc…).

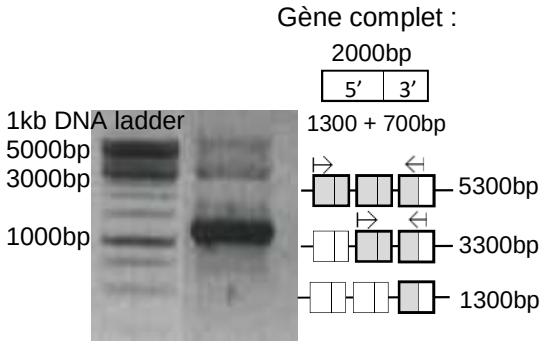


Figure 8. Amplification du gène de la preproalvinellacine (2kb) avec les amorces 5' (longueur : 1300pb) et un temps d'élongation de 2min30 chez *A. caudata* pour mettre en évidence une duplication en tandem.

3. Diversification génique au sein du genre *Alvinella*

Le jeu de données final a permis de réaliser un arbre des relations entre les différents allèles de 4 et 8 individus chez les deux espèces soeurs du genre *Alvinella* à partir de la région 3' du gène (le nombre d'allèles recapturés pour l'espèce *A. caudata* étant plus important en 3' avec un arbre plus résolutif).

Cette phylogénie d'allèles est présentée dans la Figure 9. Le nombre d'allèles par individu excède largement 2 pour les deux espèces et la topologie de cet arbre reflète la présence d'une monophylie réciproque entre les 2 espèces. Les allèles au sein de chaque espèce sont plus proches entre eux qu'avec les allèles de l'autre espèce. Ceci suggère l'existence de duplications qui ont eu lieu indépendamment au sein de chaque espèce après spéciation et

non d'une ou plusieurs duplications antérieures à la séparation des 2 espèces comme proposé par la théorie dans la partie gauche de la Figure 9. Le Tableau 3 présente les divergences nettes obtenues entre les 2 espèces (dans les régions exoniques uniquement, la partie intronique n'étant pas alignable) calculées par groupe de séquences sur les différents clades identifiés. Ces divergences ont permis de mettre en évidence qu'au niveau intra-espèce la séparation des paralogues est beaucoup plus récente que la séparation des 2 espèces (d'un facteur 10) et que la diversification par duplication a sans doute eu lieu encore plus récemment dans le cas de *caudata* (valeurs de divergence AP=0,13 vs AC=0,007).

La comparaison des diversités nucléotidiques $\theta\pi$ le long du gène de la préproalvinellacine entre les 2 espèces est présentée dans la Figure 10. Chez *A. pompejana*, il existe un très fort pic de diversité nucléotidique au sein de la région 5' localisé au niveau de l'intron 1, cet intron étant très divergent (voire non-alignable) entre les différents paralogues trouvés chez cette espèce. Cette divergence non attendue pourrait être le résultat d'une insertion d'un fragment exogène non codant d'ADN au niveau d'une des différentes copies du gène. Chez *A. caudata*, il existe aussi une augmentation locale de la diversité nucléotidique mais celle-ci est préférentiellement localisée au niveau du dernier intron dans la région 3' du gène. De la même façon, cet intron plus divergent correspond plus particulièrement à l'une des copies du gène. Une des raisons pouvant expliquer les différences de localisation des pics de diversité pourrait être que les amorces utilisées - qui ont été dessinées pour amplifier le gène chez l'espèce *A. pompejana* - n'ont pas permises d'amplifier tous les allèles de la région 5' pour l'espèce sœur *A. caudata* contrairement à la région 3'. Cette région, déjà assez diversifiée chez *A. pompejana* pourrait être trop divergente par rapport à son espèce sœur pour avoir été amplifié (problème d'amorces trop spécifiques ?) ce qui expliquerait le nombre plus faible de séquences obtenues au final dans la région en 5' chez *A. caudata* (ainsi que les difficultés d'amplification lors des manipulations au laboratoire).

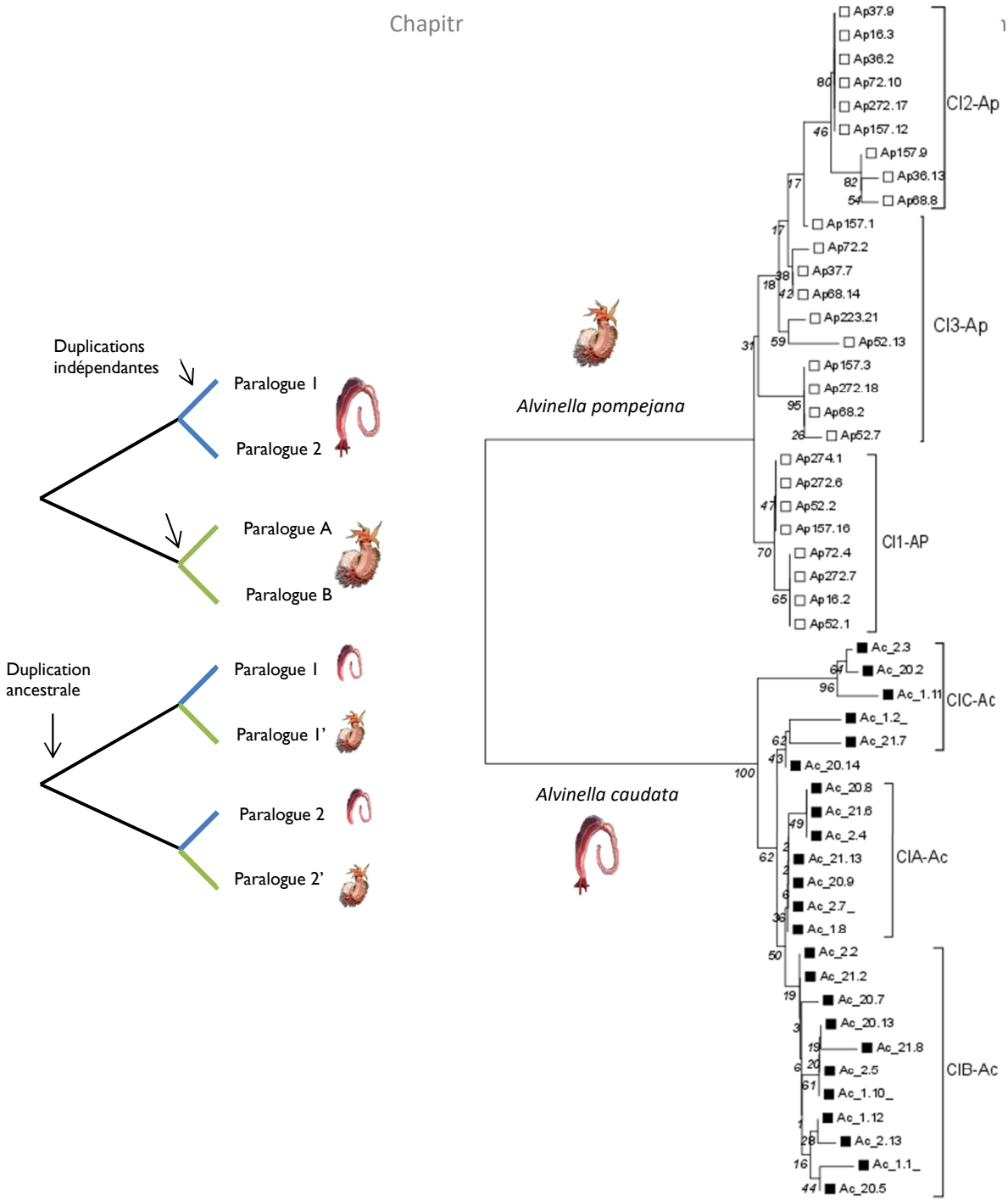


Figure 9. A droite : Diversification génique de la préproalvinellacine (avec monophylie réciproque des espèces) chez les deux espèces sœurs *A. caudata* (carré noir) et *A. pompejana* (carré blanc). A gauche : attendues théoriques d'une duplication avant et après spéciation.

		Ap			Ac		
		Clade 1	Clade 2	Clade 3	Clade A	Clade B	Clade C
Ap	Clade 1						
	Clade 2	**0.013**					
	Clade 3	**0.014**	**0.011**				
Ac	Clade A	0.114	0.124	0.118			
	Clade B	0.119	0.129	0.119	**0.005**		
	Clade C	0.115	0.124	0.115	**0.01**	**0.008**	

Tableau 3. Divergences nettes entre clades de séquences au sein et entre les deux espèces *A. pompejana* (Ap) et *A. caudata* (Ac). Les divergences ont été calculées sur les régions exoniques uniquement et les clades ont été définis visuellement grâce à l'arbre en Figure 9.

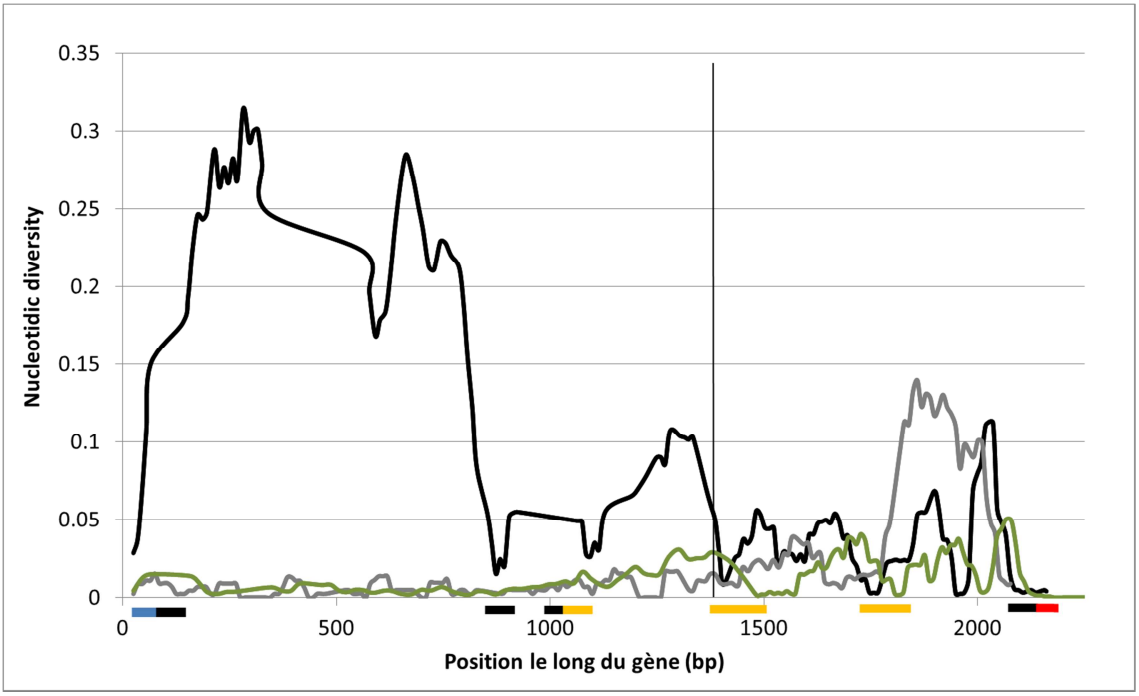


Figure 10. Distribution de la diversité nucléotidique globale le long du gène (en pb) chez les deux espèces (*A. pompejana* en noir et *A. caudata* en gris). La diversité moyenne par paralogue a également été estimée pour l'espèce *A. pompejana* (en vert). La position des différents domaines du prépropetide (régions exoniques) est également fournie selon un code couleur. En bleu : peptide signal, en noir: prorégion, en jaune : BRICHOS et en rouge : le PAM mature.

Au niveau de la région 3' la mieux recapturée pour l'espèce *A. caudata*, les deux espèces montrent des niveaux de diversité équivalents (π=0,025) mais moindres dans les régions exoniques par rapport aux introniques (Figure 10). Il est intéressant de noter que la forte diversité nucléotidique observée en région 5' chez l'espèce *A. pompejana* disparaît quand la diversité est moyennée par paralogue : ce fort pic de diversité intronique est donc lié à de la divergence inter-paralogue au niveau de ce premier intron (avec un seul paralogue très divergent dans cette région du gène). De plus, il apparaît que pour les deux espèces, la région codant pour le PAM montre une diversité nucléotidique proche de zéro.

La suite de ce chapitre s'intéressera plus particulièrement à mieux comprendre les causes de cette diversité génétique élevée au niveau du précurseur protéique de l'alvinellacine chez l'espèce *A. pompejana* pour laquelle un effort de recapture important a été effectué tant au niveau du nombre d'allèles séquencés qu'au niveau du nombre d'individus échantillonnés.

4. Identification et caractérisation de la famille multigénique chez *A. pompejana*.

4.1. Filtration des données brutes

Chez *A. pompejana*, avec un effort de recapture de 3, un total de 1152*2 (F&R) soit 2304 séquences a été obtenu en séquençage des clones pour les deux régions du gène amplifiées séparément. Après séquençage, 900 séquences (F & R) correspondant à 450 allèles ont été réellement exploitables pour la région 5' et 724 (362 allèles) pour la région 3'. Chez l'espèce *A. caudata*, 78 séquences sur les 128 : 20 en région 5' et 58 en région 3' ont pu être analysées. Toutes ces séquences ont ensuite fait l'objet d'un filtrage pour constituer un jeu de données de qualité sans séquence chimérique entre allèles d'un même individu.

A la suite de ce filtrage réalisé sur les 10 individus les mieux recapturés (expliqué en M&M), tous les allèles considérés comme recombinants entre les allèles d'un même individu ont été éliminés du jeu de données. Ceux qui correspondent à des recombinants avec une séquence inconnue ont été provisoirement gardés et alignés avec les allèles sans recombinaison à l'échelle de la population pour détecter si des séquences avec les mêmes points de recombinaison étaient présentes chez d'autres individus et, donc constituer de vrais recombinants naturels (peu de probabilités que la recombinaison affecte 2 allèles aux

mêmes endroits chez deux individus différents). Les séquences inconnues sont vérifiées car des recombinants pourraient provenir d'un évènement de recombinaison apparu tôt dans l'histoire évolutive de la famille et avoir acquis leurs propres mutations au cours de l'évolution – ou plus simplement juste liés à un biais d'échantillonnage.

La dernière étape a consisté à récupérer dans les séquences d'individus moins recapturés des allèles non recombinés déjà présents dans le jeu de données (allèles validés) pour les ajouter au jeu de données filtré.

Au final, 30% de recombinants artéfactuels ont été trouvés pour le jeu de données en région 3' et seulement 6% dans le jeu de données de la région 5'. Ceci pourrait être dû au fait que les différentes copies du gène sont plus divergentes dans la région 5' avec la présence d'un premier intron très long et fortement différencié ce qui n'est pas du tout retrouvé sur un grand nombre de copies du gène. La partie 3' est quant à elle plus stable avec des régions exoniques plus nombreuses et une divergence plus faible entre les différentes copies du gène par rapport à la région 5'.

Pour illustrer l'effort de filtrage, la Figure 11 montre le type de jeu de données brut pour un individu très bien recapturé (individu 274 en région 5') avec le programme Seaview (Gouy et al., 2009) et la méthode BioNJ sous le modèle d'évolution Jukes-Cantor avec omission des indels. L'arbre réalisé à partir d'un alignement brut de toutes les séquences obtenues pour l'individu 274 et la recherche de recombinants intra-individuels montre toute la complexité des séquences trouvées et la présence de plusieurs gènes paralogues pour le codage de la préproalvinellacine. Sur les 27 séquences trouvées, il apparait que 8 séquences (274.8; 274.2; 274.4; 274.13; 27416; 274.11; 274.20; 274.10, dont 3 présentent des motifs «unknown» non recapturés au sein des allèles de l'individu) sont des recombinants artéfactuels entre les différents allèles présents. Après avoir enlevé les séquences recombinantes au sein de chaque individu, un nombre de 8 à 10 allèles semble caractériser chaque individu, suggérant la présence d'au moins 5 gènes dupliqués.

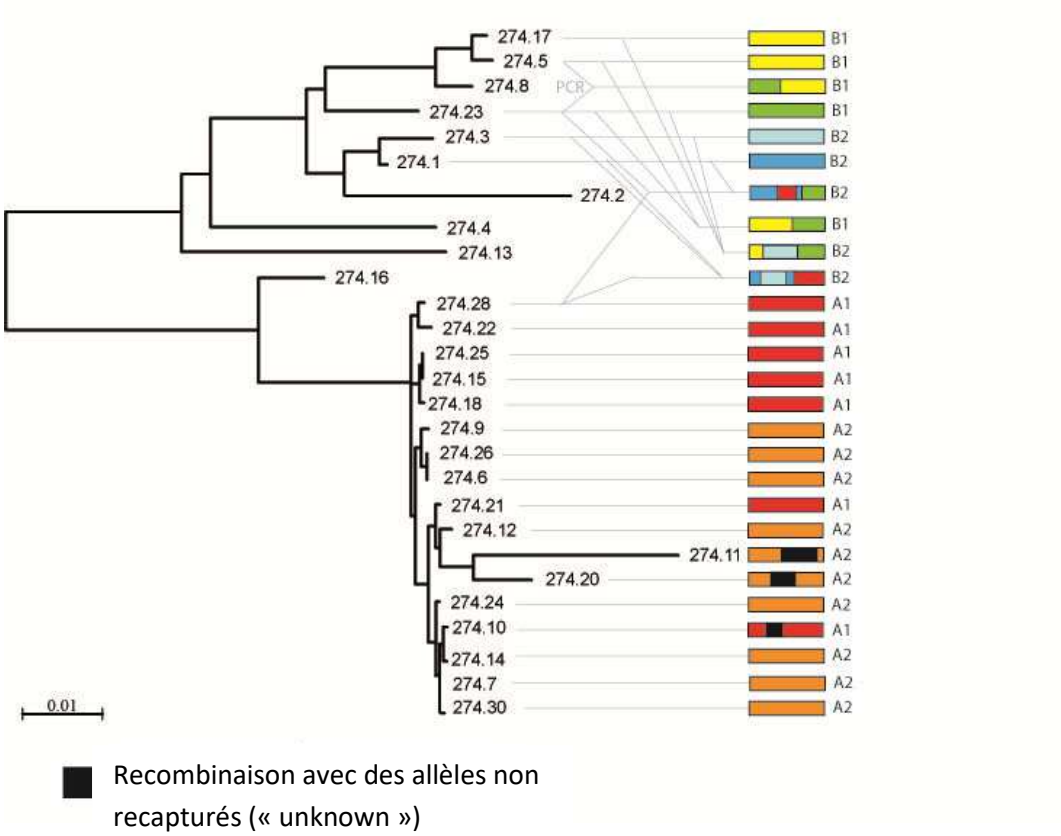


Figure 11. Arbre des relations phylogénétiques entre les allèles recapturés de l'individu 274. Les différentes couleurs désignent les différents clades retrouvés visuellement sur l'alignement.

Mode opératoire pour l'analyse et la suppression des mutations artéfactuelles.

Pour les 10 individus les plus recapturés, pour chaque clade lorsque le nombre de séquences recapturées est au moins supérieur à 3 dans un individu donné, un décompte du nombre de singletons est réalisé. Ceci permet a permis d'estimer un taux d'erreur de la polymérase lors de l'amplification, et d'estimer le nombre de singletons à éliminer dans le jeu de données global des séquences retenues après filtrage des recombinants artéfactuels. L'élimination des singletons surnuméraires est alors faite de manière aléatoire pour ne pas biaiser les estimateurs de diversité et l'estimation des fréquences alléliques.

Après suppression des séquences recapturées plusieurs fois pour un individu donné après élimination des singletons et suppression des recombinants artéfactuels, l'alignement final de la région 5' ne contient finalement que 74 séquences et celui de la région 3' que 36 séquences. Au final, après le travail de nettoyage des séquences, 10% des mutations de type singleton étaient artéfactuelles.

4.2. Caractérisation de la famille multigénique codant le précurseur protéique preproalvinellacine

Dans un premier temps, une analyse plus détaillée de la diversification génique du préproalvinellacine a donc été réalisée pour l'espèce *A. pompejana* afin de déterminer précisément les relations entre les différents allèles.

Ceci a été réalisé au sein de la région 5' qui est visuellement plus polymorphe que la région 3' notamment au sein de l'intron 1. La reconstruction des relations phylogénétiques entre les différents allèles recapturés dans les différentes populations d'*A. pompejana* a été effectué à l'aide du logiciel SplitsTree4 (Figure 12) et permis de reconstruire dans un second temps d'identifier les différentes lignées paralogues et leurs recombinants naturels à partir d'un arbre ML obtenu avec le logiciel MEGA6 (Figure 13).

Choix du modèle de substitution

A l'aide du logiciel jModelTest, le meilleur modèle de substitution a dans un premier été choisi. Selon les indices AIC et BIC, les modèles GTR+G et TPM3uf+G ressortent comme étant plus adaptés aux données. En réalisant un LRT qui compare les valeurs de vraisemblance avec un degré de liberté égal à la différence des paramètres des deux modèles emboîtés, une p-value est obtenue pour vérifier qu'un modèle avec plus de paramètres reste suffisamment robuste pour décrire les données. Ainsi, il apparaît que le modèle GTR+G+I et le GTR/TPM3uf+G sont compatibles ne montrent pas de différences significatives au niveau de la topologie des arbres produits par PhyML. il apparaît que chacun de ces modèles donne au final la même topologie qui a été retenue pour reconstruire les différentes lignées paralogues et leurs recombinants (cf. Tables S2 and S3, Fig. S2 de l'article pour les arbres avec les modèles sélectionnés par jModelTest).

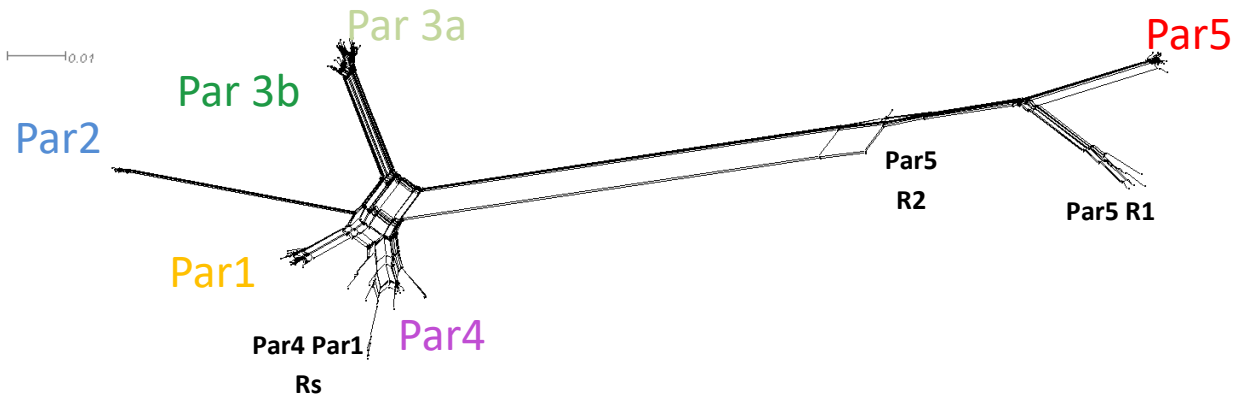


Figure 12. Réseau (NeighborNet) réalisé sous Splitstree permettant de visualiser la famille multigénique qui code pour le précurseur protéique en région 5'.

Au final, 12, 6, 13, 9, 14 et 20 séquences (correspondant aux séquences les mieux recapturées) ont été gardées pour décrire la diversité des paralogues 1, 2, 3a, 3b, 4 et 5 dans la région 5' respectivement. Le paralogue 4 est le gène le plus court avec une délétion majeure au niveau du premier intron. Le paralogue 5 est le plus long et contient au niveau du premier intron une séquence avec une région microsatellite qui lui est propre et non alignable avec l'intron 1 des autres paralogues (insertion d'une séquence exogène dû à un crossing over inégal ?).

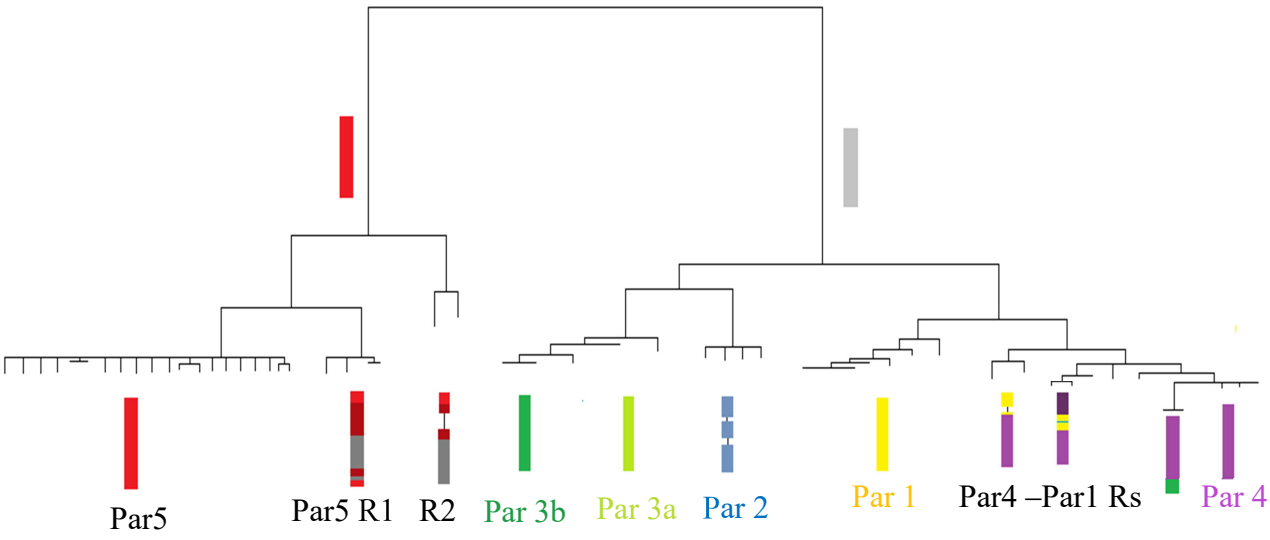


Figure 13. Reconstruction schématique de l'histoire des différents évènements de duplication et de recombinaison (R) qui ont eu lieu au cours de l'histoire du gène. En couleur rouge, verte, bleue, jaune et violette, les paralogues 1, 2, 3a et 3b, 4 et 5 (Par). En couleur grise et rouge et violet plus foncée, les régions recombinantes avec leurs propres mutations. Les bars correspondent aux indels (par exemple le paralogue 2 avec ses deux délétions dans cette région).

En effet, les régions introniques des paralogues varient beaucoup en terme de longueur de séquence d'un paralogue à l'autre, allant de 1114bp (par4) à 1293bp (par1) avec des longueurs d'intron de 798bp (par4) à 978bp (par1) supérieures aux régions exoniques de 214 à 349bp (Figure 13 et Tableau 4). Les paralogues 3a et 3b sont les moins divergents et le paralogue 2 est celui qui montre deux délétions au niveau des régions codantes (10 acides aminés et 22 acides aminés en région 5').

Puisque toutes les lignées sont présentes au sein d'un même individu, il a été possible de mettre en évidence que le gène de la préproalvinellacine est codée par une famille multigénique d'au moins 6 gènes paralogues, certains allèles (5 séquences trouvées et validées) étant des recombinants naturels inter-géniques notamment entre les paralogues 4 et 5 (trouvés chez au moins deux individus avec les mêmes points de recombinaison).

Les paralogues présumés par l'analyse phylogénétique ont été validés par une étape de génotypage en utilisant des amorces spécifiques de chaque lignée trouvée. Le génotypage par séquençage direct a été mené sur une vingtaine d'individus pour chaque locus présumé, ce qui a permis de confirmer la relation clade/locus en analysant la ségrégation Mendélienne aux mutations diagnostiques du locus testé (présence ou non de double pics) en comparant les résultats obtenus par le séquençage direct au jeu de données initial. Ainsi, tous les clades d'allèles trouvés correspondent bien à des gènes paralogues, c'est-à-dire présentent bien des individus homos et hétérozygotes (doubles pics) aux mutations diagnostiques génotypées lors du séquençage direct desdits clades. Par exemple, dans notre première analyse, une des hypothèses était que les paralogues 3a et 3b ne soient que des lignées alléliques d'un même locus. L'étape de génotypage a permis de montrer que 100% des individus montraient plus de deux allèles au paralogue 3 indiquant qu'il s'agissait bien là d'une duplication.

La caractérisation des recombinants a permis de mettre aussi en évidence que 3 des 5 recombinants naturels ayant une fréquence suffisante dans la population pour être recapturés dans au moins 2 individus sont apparus depuis suffisamment longtemps pour avoir leur propres mutations (couleurs plus foncées dans la Figure 13). C'est le cas de l'allèle Par5-R1 qui montre 30 mutations sur 881 sites (avant le point de recombinaison) et 17 sur 498 sites (après le point de recombinaison). Ceci donne une moyenne de 0,034 mutation/site. Ces résultats suggèrent que les évènements de recombinaison entre paralogues sont apparus tôt dans l'histoire évolutive de cette famille de gènes et n'ont pas été contre-sélectionnés. Ces évènements de recombinaison sont décrits uniquement dans les régions introniques et ne changent donc pas l'ordre des exons.

5. Diversité génétique et tests de neutralité associés aux différents paralogues

Le nombre de sites polymorphes est supérieur dans les régions introniques par comparaison aux régions exoniques, ceci aboutissant à des diversités nucléotidiques globales variant de 0,0026 (par2) à 0,0121 (par4) et des diversités nucléotidiques dans les exons variant de 0,0024 (par5) à 0,0112 (par4) (Tableau 4). Par comparaison chez la même espèce, des diversités nucléotidiques du même ordre ont été retrouvés par Plouviez et al. (2010) pour les gènes nucléaires de la Globine X et de la PGM (autour de 0,004 et 0,005 respectivement) avec de fortes diversités haplotypiques (minimum 0,725, surtout supérieures à 0,86).

Les diversités haplotypiques pour la préproalvinellacine sont elles aussi élevées puisque supérieures à 97% et les tests pour détecter un écart à la neutralité dans l'accumulation des mutations polymorphes ne sont pas significatifs sauf dans le cas du paralogue 5 pour lequel un excès significatif de variants rares (singletons) est détecté au seuil de 5% ($p<0.05$).

	N	L	S(ex)	Nsites(ex)	π(ex)	Nsites (ex+int)	S (ex+int)	π(ex+int)	θW	Hd	Tajima	Fu&Li F	Fu & Li D
par1	12	1293	10	315	0.0078	978	25	0.0048	0.0064	0.985	-1.145	-1.218	-1.365
par2	6	1124	2	214	0.0029	910	8	0.0026	0.0031	1	-1.072	-1.063	-1.145
par3a	13	1242	8	315	0.0058	927	30	0.0059	0.0078	1	-1.047	-0.998	-1.157
par3b	9	1239	8	316	0.0055	923	22	0.0055	0.0065	1	-0.804	-0.674	-0.791
par4	14	1114	12	316	0.0112	798	54	0.0121	0.0158	0.978	-1.039	-0.617	-0.842
par5	25	1270	8	349	0.0024	921	37	0.0032	0.0077	0.983	-2.207**	-3.360**	-3.518**

Tableau 4. Diversité génétique et tests statistiques visant à détecter un écart à la neutralité associés pour les 6 paralogues qui codent pour le précurseur protéique de l'alvinellacine d'*A.pompejana*. N : Nombre de séquences; L : Longueur du paralogue. S : Nombre de sites polymorphes dans les regions exoniques (ex) ou sur le gène entier (ex+int), Nsites : taille effective en exon du gène, π : diversité nucléotidique, θW : théta de Watterson, Hd : diversité haplotypique, Tajima : test de Tajima's D, Fu&Li F : test F de Fu and Li's F test, Fu&Li D : test D de Fu and Li's D test. Niveau de significativité ** $p<0.05$.

6. Recherche de sélection positive au sein des différents domaines du préproalvinellacine

6.1. Divergence et d_N/d_S entre domaines

Une des questions posées par une diversification des gènes après duplication est de savoir si celle-ci a donné lieu à une néo-fonctionalisation des duplicats suite à l'adaptation d'une espèce à un changement environnemental (acquisition d'une symbiose par exemple). Une des méthodes les plus utilisées pour détecter de la sélection positive sur une région codante est le calcul des taux de substitutions synonymes et non synonymes. Ainsi, le ratio d_N/d_S a été calculé dans un premier temps pour chaque domaine du prépropeptide entre les différents paralogues (Tableau 5).

Dans cette partie, nous utiliserons une nomenclature différente des allèles entre la région 5' et la région 3' puisqu'il n'a pas été possible de relier avec confiance/exactitude les allèles des deux régions à partir de la zone de recouvrement, notamment en raison des évènements de recombinaison assez nombreux pour brouiller le signal. Les deux régions ont donc été traitées indépendamment (en duplicat), ce qui permet de comparer les deux régions et la congruence des données entre régions.

Il apparaît que seules les comparaisons avec le paralogue le plus divergent (par5 en région 5' et parE en région 3') montrent un d_N/d_S supérieur à 1 dans la prorégion. Quelques ratios sont également supérieur à 1 au niveau de la comparaison par1/par5 (1,58) pour le peptide signal et dans la comparaison parB/parE au niveau du domaine BRICHOS. Les autres ratios sont élevés mais toujours inférieurs à 1 (0,21 à 0,86) sauf pour la comparaison par3a/par3b dans la prorégion et la comparaison parB/parD dans le domaine BRICHOS. Ceci suggère que la plupart des paralogues ont évolué d'une façon presque neutre, avec dans quelques cas, la possibilité d'une innovation adaptative pour l'un des domaines.

Au sein du peptide signal, 4 des paralogues présentent la même signature en acides aminés (par2, par3a, par3b par4). De la même façon, on constate que le peptide antimicrobien *sensu stricto* est monomorphe (M) sans aucun variant détecté entre et au sein des paralogues.

Region	Domaine	Paralogue	Paralogue					
			par5	par1	par2	par3a	par3b	par4
		par5						
		par1	1,5842					
		par2	0,3786	0,8005				
	SP	par3a	0,3597	0,7612	0			
		par3b	0,3597	0,7612	0	0		
5'		par4	0,7772	0,7612	0	0	0	
			par5	par1	par2	par3a	par3b	par4
		par5						
		par1	1,5667					
		par2	1,9158	0,858				
	PR	par3a	1,3084	0,2798	0,3757			
		par3b	1,0855	0,3998	0,4868	1,4101		
		par4	0,9976	0,327	0,492	0,252	0,5152	
			parE	parA	parB	parC	parD	
		parE						
		parA	0,5741					
	BRICHOS	parB	1,8245	0,214				
		parC	0,5629	0,4216	0,8065			
3'		parD	0,8603	0,8649	1,8177	0,2782		
			parE	parA	parB	parC	parD	
		parE						
		parA	M					
	PAM	parB	M	M				
		parC	M	M	M			
		parD	M	M	M	M		

Tableau 5. Comparaison des ratios de d_N/d_S entre paires de paralogues (par) pour chaque domaine à partir de séquences consensus pour chaque paralogue. SP : peptide signal ; PR : prorégion, BRICHOS et PAM.

6.2. Evolution du d_N/d_S le long du gène

Les estimateurs du ratio des divergences non-synonyme et synonyme, à savoir d_N/d_S entre paralogues et du ratio des polymorphismes non-synonyme et synonyme, $\pi a/\pi s$ au sein des paralogues ont ensuite été calculés le long du gène dans le but de localiser des potentielles zones ou le gène serait sous sélection positive. Ces données sont présentées dans la Figure 14.

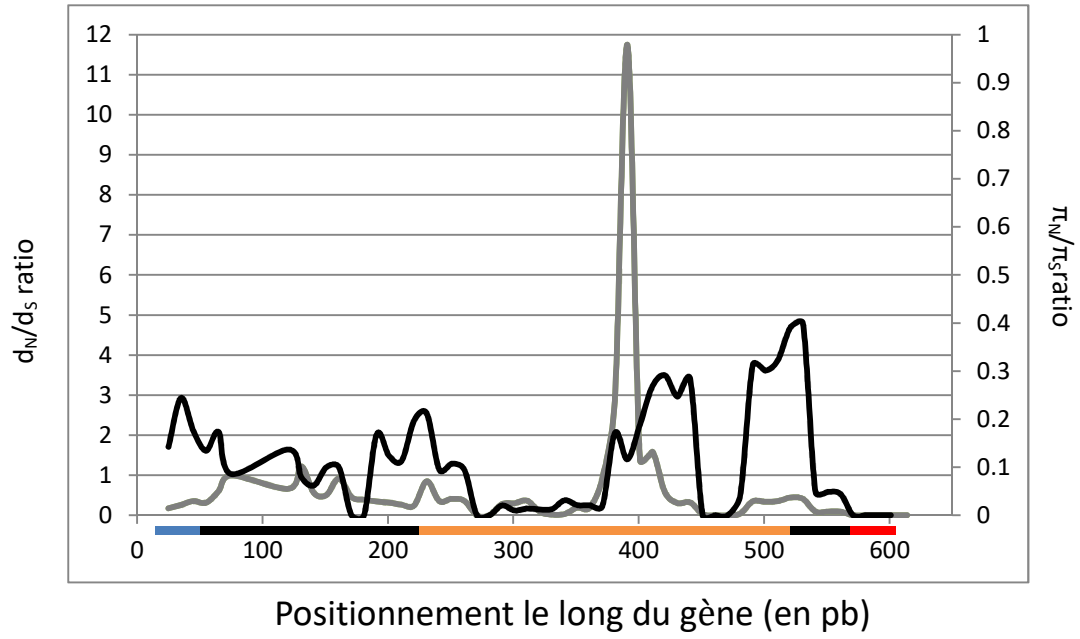


Figure 14. Valeurs de d_N/d_S (gris) et de $\pi a/\pi s$ (noir) le long de l'ADNc avec la position des différentes régions. En bleu : peptide signal, en noir : prorégion ; en jaune : BRICHOS et en rouge : PAM.

Le d_N/d_S augmente jusqu'à 12 au sein du domaine BRICHOS sur une petite portion de quelques dizaines de codons alors que le reste des exons montre des ratios oscillant entre 0 et 1. Un autre étude (résultats non montrés) en changeant le pas d'échantillonnage (20 nucléotides) de la fenêtre a été réalisée et permet de diminuer ce pic qui reste toujours largement supérieur à 1 (autour de 5) et d'augmenter légèrement les pics trouvés dans la prorégion (autour de 150pb). Le $\pi a/\pi s$ moyenné entre paralogues montre, quant-à-lui, plusieurs pics de faible amplitude n'excédant pas 0,4. Ces résultats suggèrent que la plupart de la sélection positive agit au niveau de la divergence entre paralogues, et plus spécifiquement dans le domaine BRICHOS. En amont de ce domaine les d_N/d_S et $\pi a/\pi s$

présentent des valeurs faibles (proche de 0). La région du PAM ne montre pas de signal en termes de Ka/Ks et $\pi a/\pi s$ dû à l'absence de polymorphisme génétique sur les 22 résidus du peptide.

6.3. Reconstruction ancestrale et cartographie des remplacements d'acides aminés entre paralogues du prépropeptide

6.3.1. Le domaine BRICHOS

Les différents variants du BRICHOS obtenus chez *A. pompejana* et chez *A. caudata* sont présentés sous la forme d'un alignement dans la Figure 15 pour le domaine BRICHOS et le PAM. Cet alignement montre qu'il n'y a aucun remplacement d'acide aminé dans la région du PAM chez *A. pompejana*, ce qui n'est pas le cas chez *A. caudata* ou l'on observe un polymorphisme relativement équilibré K/T entre individus sur l'un des sites du PAM. A l'inverse, le domaine BRICHOS présente de nombreuses mutations non synonymes au sein et entre les paralogues chez les 2 espèces. Les remplacements en a.a. sont particulièrement agrégés dans une région restreinte du domaine (entre le 50^{e} et le 60^{e} acide aminé du domaine, correspondant au pic de d_N/d_S précédemment localisé). De plus, il apparaît que les deux espèces divergent dans leur séquence primaire du PAM par un seul remplacement (S/N) fixé en position 119. En position 54 du domaine BRICHOS, un remplacement non synonyme semble être partagé par les deux espèces (D/G) et pourrait constituer une réminiscence de polymorphisme ancestral ou une convergence.

Choix du modèle jModeltest pour le domaine BRICHOS

L'alignement des 36 séquences du domaine BRICHOS a été analysé sous jModelTest pour déterminer quel modèle de substitution était le plus adapté aux données à l'aide des valeurs AIC/BIC et l'analyse hLRT. Cette analyse a permis de mettre en évidence que le modèle GTR+I+G fait partie des modèles adaptés aux données (avec une topologie similaire de l'arbre obtenu par le modèle K80+I sélectionné par jModeltest ; cf. supplementary data de l'article de Papot et al., 2017). Ainsi, pour aller plus loin dans la compréhension de l'histoire des mutations qui ont modelé ce domaine BRICHOS après duplication, chaque mutation non synonyme a été cartographiée grâce une reconstruction des séquences ancestrales à l'aide du programme aaML de PaML en prenant comme référence l'arbre obtenu à partir de la région 3' avec le modèle GTR+I+G (Figure 16).

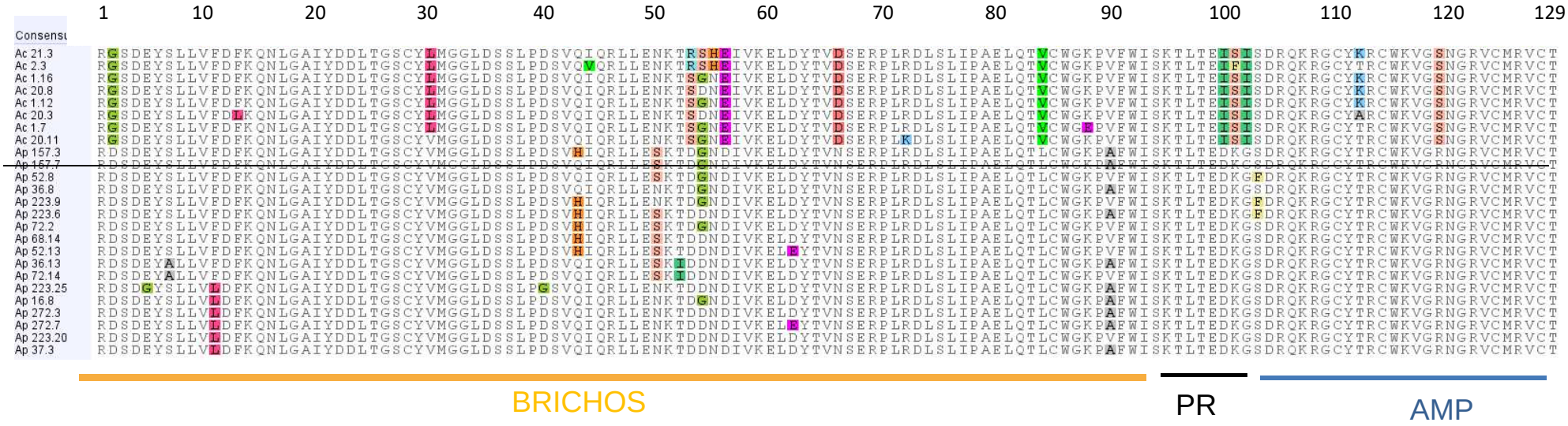


Figure 15. Alignement protéique des domaines BRICHOS et PAM des deux espèces (*A. caudata* – en haut du trait noir- et *A. pompejana* – en bas du trait noir). Les couleurs soulignent chaque changement en acide aminé : mutation non synonyme.

Sept remplacements ont été positionnés et orientés à l'aide des séquences de l'espèce sœur *A. caudata* (Figure 16). Ainsi la séquence ancestrale DQNTDDV a pu être déduite et toutes les modifications en acides aminés aux nœuds ancestraux montrent des probabilités robustes (0,999) sauf pour la mutation N129S (0,564) trouvée plusieurs fois dans les différentes lignées. Ceci indique que ce polymorphisme pourrait prédater la diversification du domaine BRICHOS ou s'être propagé entre paralogues par recombinaison. Les remplacements figurés en rouge représentent les sites diagnostiques d'un paralogue donné alors que les autres mutations se retrouvent partagées entre les différentes copies du gène. Plus spécifiquement, les mutations N129S (N50S en comptant à partir du premier codon du domaine BRICHOS), T131I (T52I) et D133G (D54G : polymorphisme retrouvé chez *A. caudata*) sont spécifiques d'un clade alors que la plupart des mutations portées par les branches terminales sont partagées par les 3 clades principaux (probabilités bayésiennes d'occurrence de 1). Par exemple, les mutations V169A (V90A) et Q121H (Q43H) apparaissent comme étant distribuées de façon aléatoire entre paralogues/clades ou au sein d'une lignée/clade respectivement (et pourrait être le fruit de la recombinaison inter-génique).

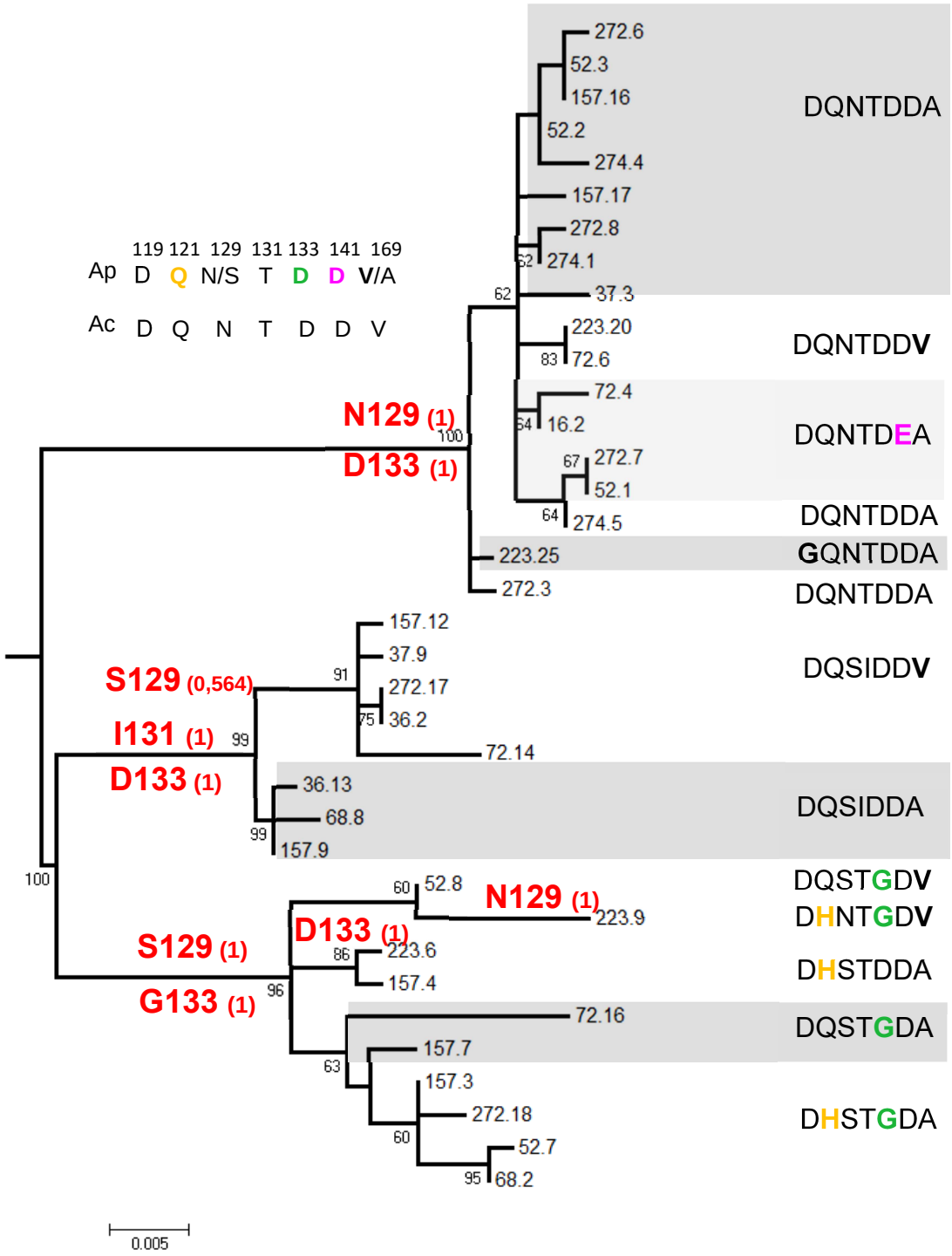


Figure 16. Cartographie des mutations du domaine BRICHOS chez *A. pompejana* (Ap) et orientation des mutations par rapport à l'espèce *A. caudata* (Ac). Les valeurs de bootstrap à chaque nœud ont également été calculées. Les chiffres avant le point correspondent à l'appartenance aux différents individus (223.9 : individu 223 clone 9). Entre parenthèses : à chaque remplacement à un nœud ancestral a également été attribué sa probabilité d'occurrence bayésienne.

6.3.2. Le domaine prorégion

Le même type d'analyse a été réalisé pour la prorégion à partir du jeu de données obtenu en région 5'. De la même façon, une reconstruction des séquences ancestrales à chaque nœud a été menée avec le logiciel CodeML dans le but de déterminer comment se distribuent les différentes mutations non synonymes dans l'arbre. L'arbre de reconstruction des états ancestraux et l'alignement correspondant sont respectivement présentés dans les Figure 17 et 18.

La Figure 17 montre les différents variants de la prorégion obtenus chez *A. pompejana* et chez *A. caudata* et suggère qu'un 'hot spot' de diversité non-synonyme existe aussi au début de la prorégion, ce qui est concordant avec l'évolution du ratio d_N/d_S le long du gène. Les mutations sont en effet particulièrement agrégées dans une région restreinte avant le 38^{e} acide aminé du domaine.

Cette Figure 17 permet également d'orienter les différentes mutations selon l'état ancestral trouvé pour les 2 espèces (en Figure 18) bien que cette partie du gène soit beaucoup plus variable (insert d'acides aminés chez *caudata*) et divergente entre les espèces. Ces données permettent de mettre en évidence que les mutations polymorphes non synonymes sont principalement diagnostiques des différents paralogues. Par exemple les mutations non synonymes (I22, D33, Y38, I60) caractéristiques du paralogue 5 l'opposent à tous les autres paralogues. Ces sites sont également impliqués dans la divergence trouvée au paralogue 2 (T22, Q26, V31, S38, T49, D76). Ceci pourrait suggérer une néo-fonctionnalisation de ces paralogues notamment si certains de ces sites sont trouvés ensuite comme étant sous sélection positive.

De la même manière que pour le domaine BRICHOS, certaines mutations sont cependant partagées entre paralogues comme A60T ou encore Q68E, ce qui pourrait renforcer l'hypothèse d'évènements de recombinaison inter-génique qui seraient plus fréquents entre les paralogues les moins divergents (1 vs 3a, 3b vs 4 et 3a vs 3b).

La prochaine étape s'intéressera donc, avec le même logiciel à étudier si certains codons sont significativement sous sélection positive dans les 2 domaines analysés.

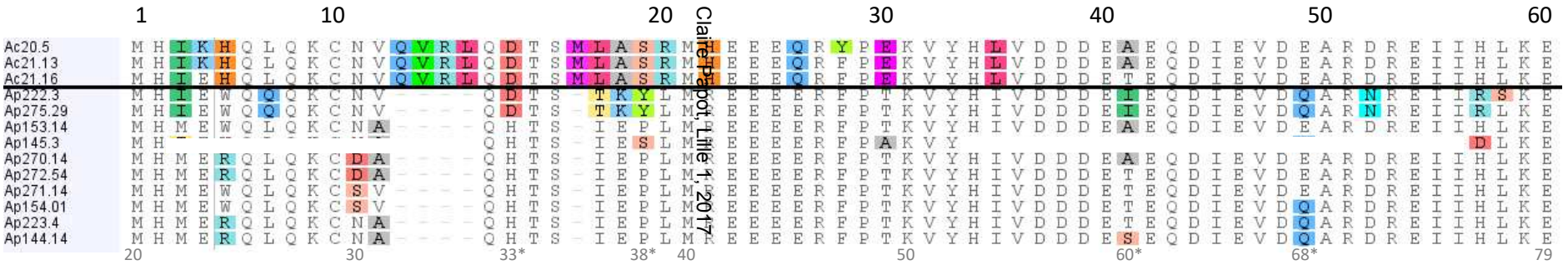


Figure 17. Alignement des domaines prorégion pour les deux espèces (*A. caudata* (Ac) et *A. pompejana* (Ap)) La numérotation des acides aminés en haut de l'alignement fait référence au premier codon de la proregion et, en bas de l'alignement à leur position par rapport à la méthionine initiatrice du gène. Dans le décompte des deux, les acides aminés supplémentaires (QVLR et M) pour l'espèce *A. caudata* ne sont pas comptabilisés.

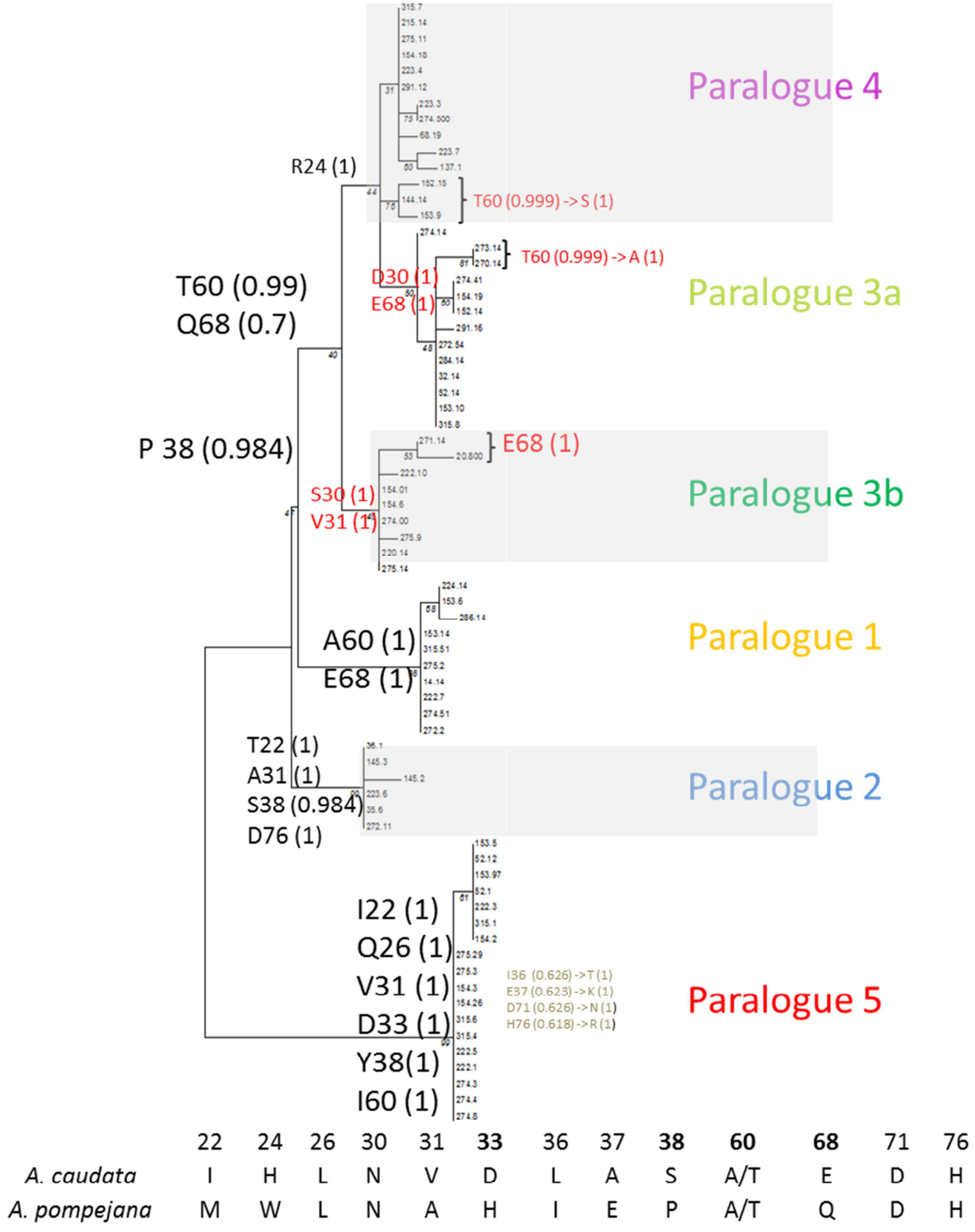


Figure 18. Reconstruction des séquences ancestrales de la prorégion à chaque nœud de l'arbre des paralogues de la préproalvinellacine ainsi que les probabilités bayésiennes associées calculées à l'aide du logiciel codeML (Yang, 2007).

6.4. Recherche de codons sous sélection positive au sein des domaines BRICHOS et Prorégion

6.4.1. Le domaine BRICHOS

Choix du modèle de substitution pour l'analyse codeML

Pour cette analyse, l'arbre de référence utilisé correspond à celui précédemment obtenu avec le modèle GTR+I+G sur l'ensemble des 36 séquences du domaine BRICHOS. Cependant, une deuxième analyse a été effectuée pour l'article de Papot et al., 2017 (en Annexe 2) avec des séquences consensus de chaque clade (paralogue) avec le modèle K80+I qui avait été retenu par jModelTest. Cette analyse présentée dans l'article ne change pas les résultats ni la discussion associée. Elle a été effectuée avec un nombre plus restreint de séquences pour éviter d'introduire des données de polymorphisme dans l'analyse de la divergence inter-paralogues. Cette dernière analyse ajoute seulement de deux mutations supplémentaires dans le groupe des mutations sous sélection positive (P38Y et Q68E ($p>0.96$, BEB avec M8) pour le domaine BRICHOS.

6.4.2. Le domaine Prorégion

Choix du modèle de substitution pour l'analyse codeML

Encore une fois, le meilleur modèle de substitution pour construire l'arbre de référence a été choisi à l'aide du logiciel jModelTest, et c'est le modèle Tim2ef+I+G qui a été retenu bien que les valeurs de AIC et BIC donnaient des modèles différents (K80+I et TPM2+I+G respectivement : cf. Annexe 2). Ce choix a été effectué après vérification que les topologies obtenues pour ces 3 modèles restaient identiques et que les p-values du hLRT n'étaient pas significativement différentes entre les modèles.

Recherche de codons sous sélection positive dans les 2 domaines

Les Tableaux 6 et 7 présentent les résultats des estimations des paramètres et des valeurs de vraisemblance d'adéquation des données aux six modèles testés sur le remplacement des codons dans la prorégion et le domaine BRICHOS.

Les valeurs de LRT montrent des résultats significatifs pour les trois comparaisons des modèles emboîtés (modèles de sélection vs modèles presque neutre) pour la prorégion alors que seule la comparaison M_0 vs M_3 montre un LRT significatif pour le domaine BRICHOS

(Tableau 7). La comparaison entre les modèles M_0 et M_3 indiquent que le d_N/d_S est distribué de façon hétérogène entre les codons mais ne fournit pas d'information sur l'action éventuelle de la sélection positive sur le domaine tant que les deux autres comparaisons ne sont pas significatives. Il apparaît grâce au Tableau 6 qu'une fraction non négligeable des codons pourrait être localement sous sélection diversifiante au sein du domaine BRICHOS. En effet, la recherche de sites sous sélection positive par l'approche bayésienne NEB et BEB présentés dans le Tableau 6 suggère que plusieurs sites impliqués dans la divergence inter-paralogues pourrait être effectivement sous sélection positive malgré une évolution presque neutre du domaine (i.e. nombreux sites approchant un d_N/d_S de 1).

Les autres comparaisons M1a vs M2a et M7 vs M8 permettent en effet d'affiner l'analyse et de conclure qu'une portion non négligeable des sites apparaît comme étant sous sélection diversifiante (19%, ω=3,99; Tableau 6). Sept sites sous sélection positive ont été identifiés pour le domaine BRICHOS en utilisant la méthode NEB (dont les mutations N129S, I131T, D133G) mais aucun de ces sites n'a été validé de façon significative avec la méthode BEB (plus robuste qui teste la valeur du d_N/d_S du codon contre un 'background' neutre).

Au niveau de la prorégion, jusqu'à 12 sites ont été classés dans la catégorie 'sous sélection positive' bien que seulement 4 sites présentent des probabilités bayésiennes supérieures à 0,95 avec la méthode NEB et deux avec la méthode BEB (T60IAS*; Q68E*). Ces résultats sont similaires à ceux obtenus l'analyse sur les séquences consensus et le modèle K80+I qui trouve également deux sites sous sélection positive avec un BEB supérieur à 0,95 bien qu'un des deux sites ne soit pas le même (P38Y* and Q68E* cf. Papot et al. 2017). Ainsi, ces résultats sembleraient indiquer l'action d'une sélection diversifiante au sein des deux domaines sur au moins trois sites impliqués dans la différenciation des locus paralogues. Il est intéressant de noter que les mutations P38Y et I60TAS opposent le paralogue 5 aux autres paralogues et que la mutation Q68E retrouvée dans toutes les analyses est la seule à être impliquée dans de la rétention de polymorphisme ancestral et/ou de la convergence par recombinaison.

BRICHOS	Model					
Parameter	M0	M3 (Discrete)	M1 (Neutral)	M2 (Selection)	M7 (beta)	M8 (beta +w)
l	-443,76	-437,09	-439,14	-437,09	-439,17	-437,09
Parameters estimates	ω= 0,612	ω0=0, p0=0,79, ω1=2,788 p1=0,18, ω2=2,79 p2=0,03	ω0= 0, p0=0,67, (ω1=1) p1=0,33	ω0=0, p0=0,79, ω1=1 p1=0, ω2=2,78 p2=0,21	p=0,005, q=0,0117	p0=0,78, p=0,005, p1=0,21, q= 2,74, ω=2,78
Sites with dN/dS>1 (NEB analysis)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all **)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all: **)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all:**)
Sites with dN/dS>1 (BEB analysis)	n.a.	n.a.	n.a.	Q121H; N129S; D133G; D141E; V169A (not significant)	n.a.	Q121H; N129S; D133G; D141E; V169A (not significant)
PROREGION	Model					
Parameter	M0	M3 (Discrete)	M1 (Neutral)	M2 (Selection)	M7 (beta)	M8 (beta +w)
l	-593,27	-585,85	-591,82	-585,88	-592,88	-585,88
Parameters estimates	ω=1,24	ω0=0,49, p0=0,76, ω1=3,33 p1=0,14, ω2=7,32 p2=0,1	ω0= 0, p0=0,29, (ω1=1) p1=0,71	ω0=0,45, p0=0,61, ω1=1 p1=0,21, ω2=5,96 p2=0,18	p=0,012, q=0,005	p0=0,81, q=5,23, p1=0,19, q= 3,99, ω=5,85
Sites with dN/dS>1 (NEB analysis)	n.a.	N22I; W24R; L26Q; N30S; A31V; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S	n.a.	W24R; L26Q; N30S; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S	n.a.	N22I; W24R; L26Q; N30S; A31V; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S
Sites with dN/dS>1 (BEB analysis)	n.a.	n.a.	n.a.	W24R; L26Q; N30S; H33D; P38YS;T60IAS*; Q68E*; H76RD	n.a.	W24R; L26Q; N30S; H33D; P38YS;T60IAS*; Q68E*; H76RD; L78S

Tableau 6. Paramètres estimés pour les modèles de sélection de codons implémentés dans PaML avec le programme codemL et les valeurs de vraisemblance de ces modèles pour les deux régions BRICHOS et la prorégion. Ces modèles autorisent (M2,M3,M8) ou non (M0,M1,M7) l'action d'une sélection positive sur les codons.

BRICHOS DOMAIN				
LRT	Models	**M0 versus M3**	**M1 versus M2**	**M7 versus M8**
	2Δl	6,66	4,1	4,16
	df	2	2	2
	pvalue	**0,0357**	0,1287	0,1249
PROREGION				
LRT	Models	**M0 versus M3**	**M1 versus M2**	**M7 versus M8**
	2Δl	14,84	11,88	14
	df	2	2	2
	pvalue	**0,0005**	**0,0026**	**0,0009**

Tableau 7. Tests de comparaison de modèles (LRT : 2ΔLnl) entre modèles emboîtés (M1 vs M2 ; M0 vs M3 et M7 vs M8).

7. Induction de la préproalvinellacine sous différents stress.

La préproalvinellacine et la chaperone moléculaire HSP70 présentent le même patron d'expression à savoir l'induction du gène lorsque l'animal est soumis à des températures non optimales (20 et 54°C) ou lorsque les animaux sont dépressurisés (Figure 19). Ces données suggèrent donc que le précurseur protéique du PAM est inductible par un stress environnemental et que cette induction pourrait être associé au domaine BRICHOS qui, tout comme les HSP70s, est documenté comme inductible et agissant comme un chaperon moléculaire capable de corriger le repliement de la molécule associée (Sánchez-Pulido et al., 2002; Willander et al., 2011).

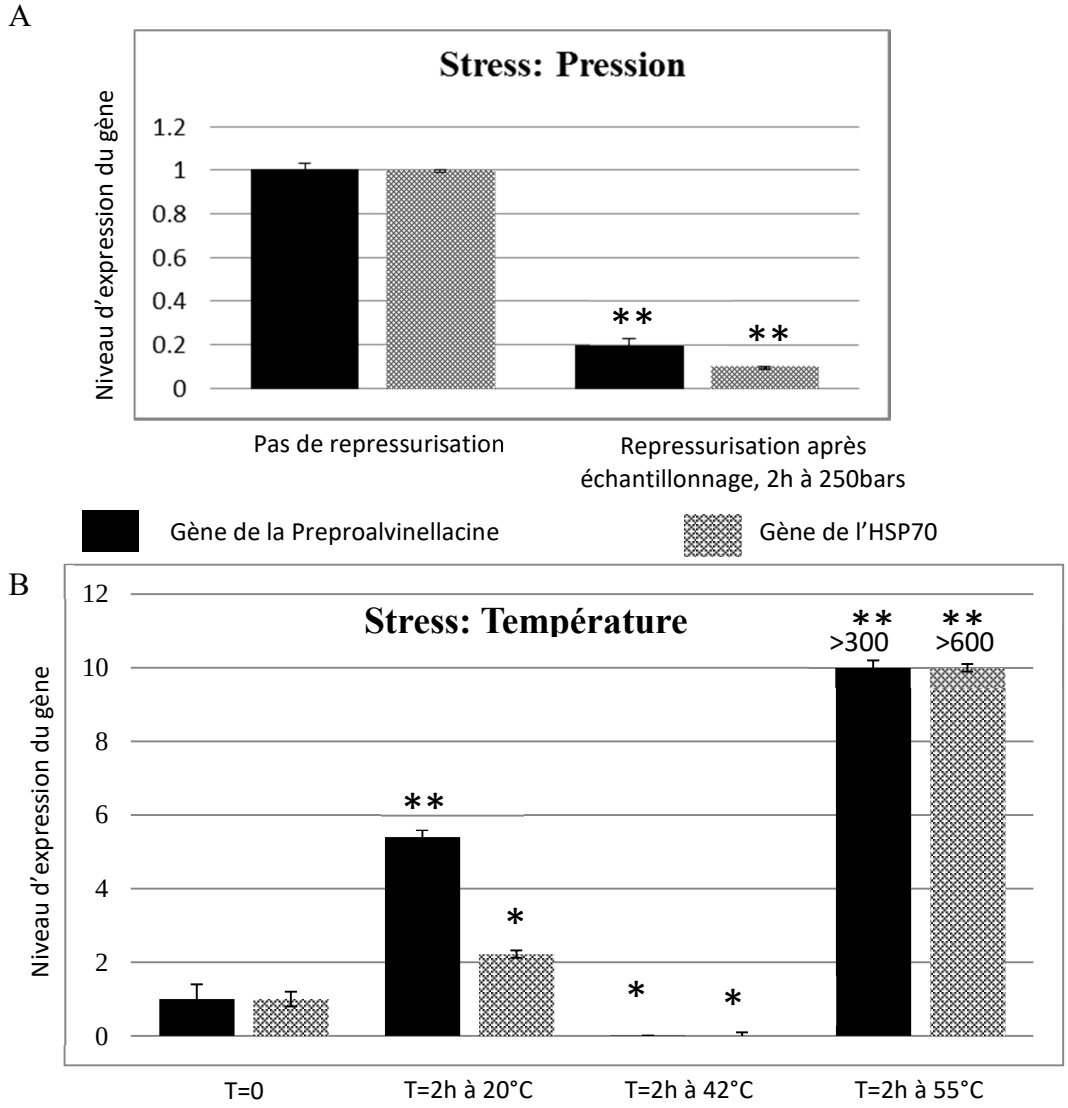


Figure 19. Niveaux d'expression des gènes codant la préproalvinellacine et l'HSP70 en réponse à deux stress : température (B) et pression hydrostatique (A). La condition 2h à 42°C ainsi que la repressurisation des animaux à 250 bars sont considérées comme conditions optimales puisque se rapprochant de celles dans lesquelles évolue l'annélide.

8. Différenciation des populations nord/sud au gène de la préproalvinellacine

Une étude de la différenciation génétique des populations nord et sud de l'EPR a été effectuée sur la base des allèles recapturées pour 6 individus de 9°50N et 10 individus de 18°25S au niveau du paralogue 1 et, des allèles de 10 individus de 9°50N et 7 individus de 18°25S pour le paralogue 4 (Figure 20 et Tableau 8). Ces tests de différenciation génétique ont été effectués en échantillonnant qu'un seul allèle par individu et sont donc fortement conditionnés par le faible effectif des échantillons produits.

Bien que s'appuyant sur un faible effectif d'allèles, les arbres de coalescence (Figure 20) montrent une tendance différente pour les deux locus paralogues analysés. Alors que certains allèles du nord (N58, N14, N152, N146, N153) se retrouvent ensemble et s'opposent à un clade plus spécifiquement formé par les allèles sud au niveau du paralogue 4, tous les allèles nord et sud se mélangent dans l'arbre du paralogue 1 avec de faibles valeurs de bootstrap qui ne supportent pas l'existence d'une dichotomie nord/sud. Au niveau des réseaux d'haplotypes, il apparaît ainsi que pour le paralogue 4, un haplotype principal regroupant la plupart des allèles du sud s'oppose à plusieurs haplotypes divergents au nord alors qu'aucune structure géographique n'est distinguable entre les allèles nord et sud pour le paralogue 1.

En ce qui concerne les tests de différenciation génétique (Tableau 8), les valeurs de F_{st} et Phi_{st} sont concordantes avec une différenciation génétique significative entre le nord et le sud pour le paralogue 4 (Phi_{St} = F_{st} = 0,18 avec des p-value de 0,027 et 0,03 respectivement). Les tests exacts quant à eux ne montrent pas de valeurs significatives quant à l'hypothèse nulle de non-différenciation. En ce qui concerne le paralogue 1, les p-values pour les deux tests sont non significatives avec des valeurs de Phi_{st} et F_{st} égales à zéro (non différenciation génétique entre populations). Le test exact de non différenciation génétique entre populations fournit en conséquence des probabilités proches du seuil de 5% (0,058 et 0,052 respectivement).

		Phi_{ST}		F_{ST}		D_{xy}	Exact test p value	
	Nb ind (N/S)	index	p-value	index	p-value		SNP	Haplotype
Paralogue 1	6N/10S	0	0,55	0	0,59	0,0039	**0,058**	**0,052**
Paralogue 4	10N/7S	0,18	**0,027**	0,18	**0,03**	0,0086	1	1

Tableau 8. Indices de différenciation (Phi_{st} et F_{st}) avec leur probabilité de s'écarter de zéro par un test de permutation des allèles entre populations et divergence nette entre populations (nord vs sud) pour les deux paralogues les plus recapturés (Paralogue 1 et Paralogue 4). Les tests exacts de non-différenciation génétique (H_0) entre populations associés sont également produits.

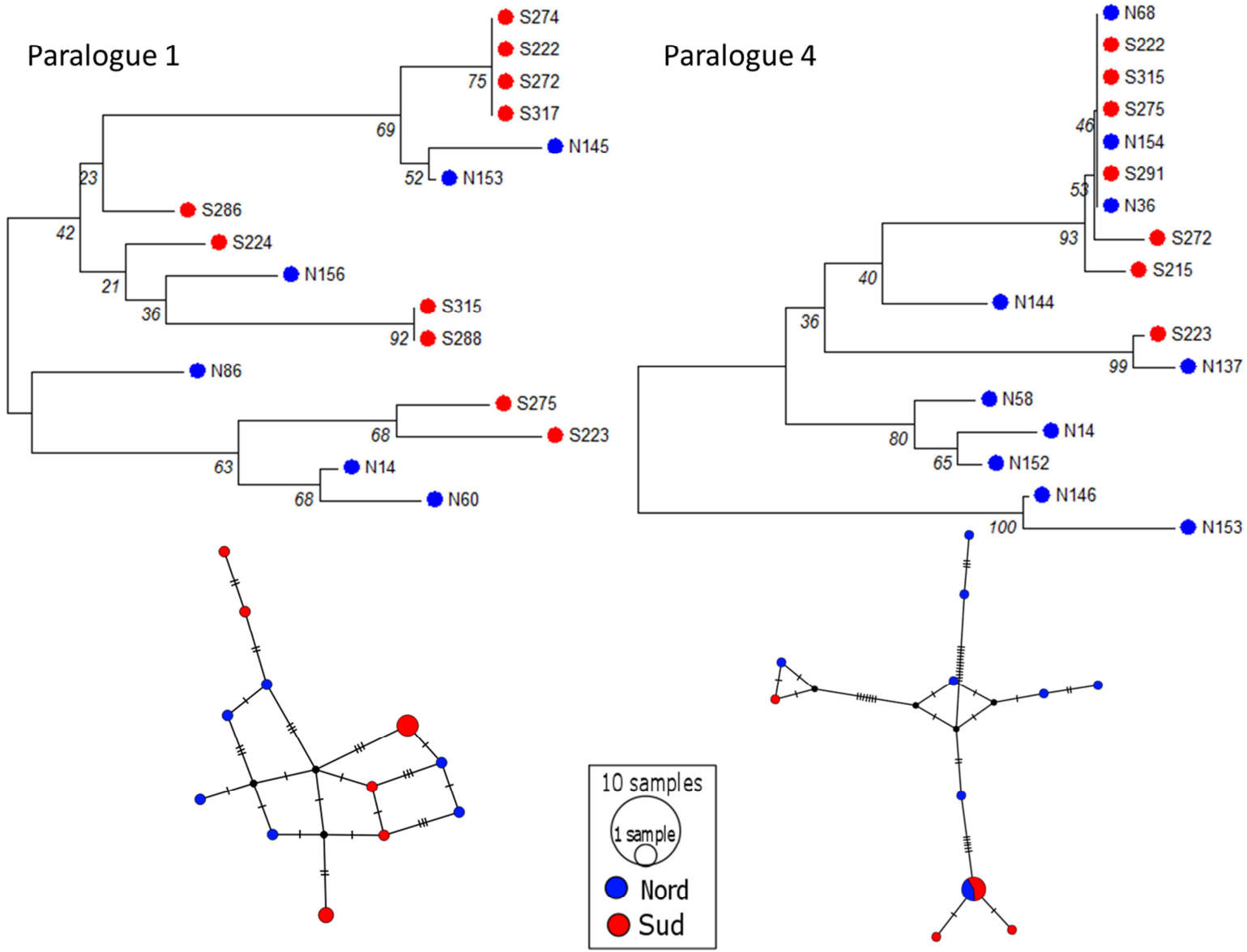


Figure 20. Analyse de la différenciation des populations nord/sud pour deux paralogues (1 et 4) : relations phylogénétiques entre les allèles des individus nord (en bleu) et sud (en rouge) avec le réseau d'haplotypes correspondant.

DISCUSSION

L'évolution moléculaire de la famille multigénique codant pour un précurseur protéique de peptide antimicrobien spécifiquement impliqué dans l'immunité externe des vers a été étudiée chez l'annélide hydrothermal *Alvinella pompejana* et son espèce sœur *A. caudata*. Ce PAM de 22 acides aminés a été décrit comme permettant de contrôler une ectosymbiose obligatoire en sélectionnant la microflore associée à *A. pompejana* (Tasiemski et al., 2014). Le précurseur protéique de ce PAM étant constitué de plusieurs sous-régions (peptide signal, prorégion, domaine BRICHOS et PAM), la diversification de ce gène et l'action de la sélection sur celle-ci a été étudiée au sein de ces différentes sous-unité fonctionnelles dans le contexte écologique propre aux deux espèces, à savoir des conditions extrêmement fluctuantes du point de vue du régime thermo-chimique puisque ce ver est décrit comme l'une des espèces les plus thermotolérantes de notre planète (Cary et al., 1998; Chevaldonné et al., 2000; Piccino et al., 2004; Ravaux et al., 2013).

Histoire phylogénétique de la molécule : un gène ancien

L'histoire de ce gène apparaît comme étant relativement ancienne et partagée entre de nombreuses espèces d'annélides appartenant aux Alvinellidae, Terebellidae, Capitellidae et Arenicolidae et donc plus spécifiquement associée aux espèces appartenant au Terebellomorpha (Rousset et al., 2007). Une duplication ancestrale du gène semble avoir eu lieu avant la radiation des Alvinellidae puisque deux clades de prépropeptide présentant les mêmes regroupements d'espèces ont été mis en évidence, l'un des clades semblant plutôt spécifique de la famille des Alvinellidae. Ces 2 lignées seraient donc présentes depuis au moins 60 millions d'années puisque plusieurs études s'accordent à dater la radiation évolutive de cette famille après l'extinction massive de la faune profonde à la fin du Crétacé (Jacobs and Lindberg, 1998; Little and Vrijenhoek, 2003 D. Jollivet, comm. pers.). Cette analyse a révélé l'existence d'un deuxième PAM différent plus long (32-36 a.a.) et plus riche en cystéines et éléments aromatiques que l'alvinellacine. La séquence du PAM putatif serait dès lors plus riche en ponts disulfures qui pourraient lui conférer une stabilité plus grande vis-à-vis de la température et la pression hydrostatique. Chez l'espèce polaire *Amphitritides*, les PAMs putatif trouvés sont dépourvus de cystéines ce qui pourrait traduire un relâchement des pressions de sélection lié à un environnement stable et froid depuis des

millions d'années (nécessité d'aller vers des molécules plus flexibles) mais qui pourraient dès lors être dépourvu de nombreuses lignées pathogènes. Cette évolution pourrait même avoir abouti à une néo-fonctionnalisation du peptide comme cela a pu être décrit chez les poissons nototheniidae qui présentent des protéines antigel faisant suite à une duplication du gène codant à la base pour la « sialic acid synthase » (Deng et al., 2010).

En ce qui concerne plus spécifiquement l'alvinellacine *sensu stricto*, on observe une grande variabilité de la structure primaire du peptide entre les différentes espèces même si celle des domaines de la prorégion et du BRICHOS est relativement conservée pour ces mêmes espèces. Cela contraste fortement avec les résultats trouvés au sein et entre les deux espèces d'*Alvinella*, chez lesquelles le PAM ne montre pas (ou presque pas) de différences/variation au niveau de sa structure aminée (contrairement aux domaines BRICHOS/prorégion qui, eux sont extrêmement variables et divergent entre les 2 espèces). Cette opposition entre l'évolution à long terme de la molécule et celle observée à l'échelle de l'espèce suggère que l'habitat et le mode de vie des espèces puissent jouer un rôle déterminant dans l'évolution complexe de ce système PAM-chaperon, le taux de divergence n'étant pas synchrone ou proportionnel entre les différents domaines du prépropeptide. La suite s'intéressera donc à décrire la diversité génétique du prépropeptide au sein de ces deux espèces sœurs pour mieux comprendre le rôle de la sélection positive dans cette évolution.

Une diversification génique récente de la préproalvinellacine possiblement liée à l'acquisition d'une épibiose trophique

La monophylie réciproque des différents gènes dupliqués de la préproalvinellacine retrouvée entre les deux espèces sœurs d'*Alvinella* semble indiquer qu'une diversification par duplication du gène s'est effectuée indépendamment au sein de chaque espèce après la spéciation. Si l'on pose comme hypothèse que l'augmentation en nombre de copies et leur diversification constitue une première conséquence de l'acquisition de l'épibiose des 2 vers, il semblerait alors que cette symbiose puisse avoir été acquise de façon indépendante après la séparation des 2 espèces. L'évolution non-homogène de la diversité le long du gène des 2 espèces (premier intron pour *A. pompejana*, dernier intron pour *A. caudata*) supporte une histoire séparée de la diversification de ce gène de l'immunité chez les 2 espèces. De plus,

cette famille de gènes montre des niveaux de diversités nucléotidiques élevés (près de 0,005 dans les régions exoniques) et similaires pour les 2 espèces. Surtout, les niveaux de divergence nettes interclades au sein des deux espèces sœurs sont du même ordre de grandeur ce qui semblerait indiquer que l'acquisition de la symbiose aurait pu se faire à peu près au même moment. Des duplications indépendantes ont été rapportées dans le cas des beta-défensines murines après la divergence entre rat et souris (il y a presque 40 millions d'années) (Maxwell et al., 2003).

Les deux espèces *A. pompejana* et *A. caudata* représentent des espèces ayant une longue histoire écologique commune au sein desquelles les duplications sont apparues très tôt après la spéciation (environ 30Ma) avec l'acquisition de l'épisymbiose. Les évènements de duplication sont souvent considérés comme une réponse à des changements environnementaux puisqu'ils représentent un moyen de créer de nouvelles fonctions sans modifier la fonction première du gène dupliqué et donc sans coût en terme de sélection différentielle (Kondrashov, 2012; Kondrashov et al., 2002). Du point de vue immunitaire, cette étude permettrait donc de faire l'hypothèse d'un avantage adaptatif des duplications géniques quant à l'acquisition/mise en place/maintien d'une association épibiotique considérée comme l'un des moyens les plus sûrs pour vivre dans un écosystème basé sur la chimiosynthèse comme celui des sources hydrothermales (Bright and Giere, 2005). Les deux espèces auraient alors subi une évolution parallèle pour faire face au même environnement microbien lors de la colonisation de la partie la plus hostile des cheminées hydrothermales il y a plusieurs dizaines de millions d'années (Tunnicliffe et al., 1998).

Les données sur les fossiles d'espèces hydrothermales dans les ophiolites supportent l'existence de vers annelés tubicoles sur les cheminées hydrothermales dès la fin du Dévonien (Hymon et al., 1984; Little et al., 2004). Ces auteurs font l'hypothèse que ces organismes, décrits dans les montagnes de Californie (« Franciscan complex ») et datant du Jurassique (150 Ma), étaient déjà au contact avec des bactéries chimioautotrophes puisque les conditions environnementales (notamment la présence d'H_2S et de bactéries filamenteuses) décrites dans ces roches sont similaires à celles documentées pour les sources hydrothermales actuelles (Little et al., 2004). Les données géochimiques suggèrent d'ailleurs que les microbes chimiosynthétiques auraient pu exister dans les océans pendant les périodes protérozoïques, mais la première preuve microscopique concernant des filaments microbiens associés à des gisements de sulfures les dateraient de l'Ordovicien (il y

a 480 Ma) au nord-est de l'Australie (Duhig et al., 1992). Ainsi, il semblerait que l'environnement microbien chimiosynthétique ait été présent et relativement stable depuis l'explosion précambrienne et peut être même que les ancêtres des Alvinellidae aient pu déjà s'associer avec des bactéries symbiotiques puis perdre les espèces porteuses de cet avantage lors des extinctions massives de la fin du Crétacé. La diversification récente de la préproalvinellacine suggère quant à elle que l'espèce ancestrale aux deux espèces sœurs d'*Alvinella,* tout comme les *Paralvinella* ne devait pas être symbiotique et a dû s'adapter par la suite pour supporter/maintenir une interaction durable avec les communautés microbiennes des cheminées.

Une évolution par duplications en tandem avec de la recombinaison inter-génique

Le rôle des duplications dans la diversification des effecteurs immunitaires (dont les gènes codant pour les PAMs) a été largement documentée (Lynn et al., 2004b; Semple et al., 2003; Tennessen and Blouin, 2007). Des duplications géniques fréquentes de ces gènes permettent une accumulation de variants au sein du génome (sans coût en termes de sélection diversifiante) permettant aux molécules d'être alors libre d'évoluer plus ou moins indépendamment/rapidement. Ceci peut à ce titre représenter un processus adaptatif dans la course à l'armement entre l'hôte et ses pathogènes (« gene for gene ») (Tennessen, 2005) et/ou représenter un avantage adaptatif vis-à-vis d'éventuelles modifications des paramètres biotiques et/ou abiotiques du milieu dans lesquels les espèces évoluent (« matching alleles »).

Le précurseur protéique de l'alvinellacine répond à cette stratégie en étant codé par une famille multigénique d'au moins 6 gènes chez *A. pompejana* alors que le gène n'a été retrouvé qu'à une seule copie dans le génome de *Capitella teleta* (JGI). Ce gène se serait dupliqué, au moins partiellement en tandem comme le montrent les résultats d'amplification du gène chez *A. caudata*. La reconstruction phylogénétique des allèles du précurseur protéique en région 5' montre qu'un des gènes : le paralogue 5 (rouge, retrouvé indifféremment chez des individus nord et sud) est nettement plus divergent par rapport aux 5 autres. Cette divergence explique une partie non négligeable de la diversité nucléotidique globale observée dans la première partie du gène. En effet, dans cette région, le premier

intron du paralogue 5 ne s'aligne pas avec ses homologues séquencés ce qui pourrait laisser penser à une origine exogène de l'intron (transposon ou recombinaison ?). De plus, les diversités nucléotidiques et haplotypiques ainsi que les valeurs des tests de Tajima, Fu & Li qui sont significativement négatives soulignent un excès de variants rares indiquant soit l'action d'une sélection positive (fixation d'un allèle avantageux par balayage sélectif) ou un effet démographique tels qu'un goulot d'étranglement suivi d'une phase d'expansion. Puisque tous les locus ne sont pas touchés de la même façon, un effet démographique peut être exclu à la faveur d'un balayage sélectif. En effet, les autres locus montrent des niveaux de diversités nucléotidique et haplotypique ainsi que des tests d'écart à la neutralité qui sont semblables à d'autres gènes nucléaires déjà étudiés pour l'espèce. En effet, par comparaison, les diversités nucléotidiques observées par Plouviez et al. (2010) pour les gènes nucléaires de la Globine X et de la PGM sont autour de 0,004 et 0,005 respectivement avec de fortes diversités haplotypiques (minimum 0,725, surtout supérieures à 0,86). Ces diversités relativement élevées s'expliquent sans doute par l'existence de queues d'introgression de part et d'autre de la barrière équatoriale qui séparent les populations nord et sud. Le fait que le paralogue 5 soit peu diversifié et insensible à la barrière géographique est donc un élément supplémentaire en faveur d'un balayage sélectif.

De la même façon, Schmitt et al. (2010) ont étudié le polymorphisme des défensines de l'huître *Crassostrea gigas* et ont trouvé qu'un des gènes, *Cg-defh1*, montrait un polymorphisme très faible contrairement aux autres défensines de la même espèce. Ces auteurs ont émis l'hypothèse que plus que l'action d'une sélection purifiante, il pourrait s'agir là de l'action d'un balayage sélectif récent qui aurait réduit la diversité génétique du gène. Une autre hypothèse invoquée pourrait également résider dans le fait que ce gène ne soit apparu par duplication que très récemment et n'aurait alors pas eu le temps de se diversifier. Dans le cas du paralogue 5, cette dernière hypothèse n'est pas convaincante puisque ce gène est aussi le gène le plus divergent entre les différents paralogues trouvés. La première hypothèse est également renforcée par le fait que le paralogue 5 est aussi celui le plus retrouvé comme source dans les recombinants naturels gardés par la sélection naturelle, ces recombinants devant être en fréquence suffisamment élevée pour avoir été échantillonnés et porteurs pour certains de leurs propres mutations (apparition ancienne de ces recombinants dans les populations).

Impact de la recombinaison dans l'histoire évolutive de la famille multigénique

Plusieurs recombinants inter-géniques naturels ont en effet pu être identifiés entre les paralogues de la preproalvinellacine et ce surtout dans la région 5' du gène codant le précurseur protéique sans perte d'exon ni apparition de codon stop. En effet, les points de recombinaison sont uniquement observés au sein des régions introniques et ne modifient pas l'ordre des exons ni leur longueur (si tel avait été le cas, ils auraient été contre-sélectionnés). De plus ces recombinants montrent leurs propres mutations (0,05 mutation/site en moyenne) et impliquent des recombinaisons entre les paralogues les plus divergents (paralogue5 et l'ancêtre commun des autres paralogues). Ces résultats indiquent une histoire de recombinaison qui est apparue tôt dans l'histoire de cette famille multigénique. De façon générale, la recombinaison inter-paralogues est souvent possible pendant la première phase de l'évolution séparée des duplicats (Fawcett and Innan, 2011) et représente un mécanisme évolutif majeur qui crée de la diversité génétique bien que leur contribution relative entre gènes et organismes varie grandement (Awadalla, 2003; Worobey and Holmes, 1999). Des évènements de recombinaison entre gènes dupliqués ont pu être retrouvé chez des PAMs (défensines) d'invertébrés (Schmitt et al., 2010) et comme agissant de façon intra-génique sur les défensines de moules (Boon et al., 2009). Concernant d'autres effecteurs de l'immunité, les gènes codant pour certains gènes du CMH montrent également des cas de recombinaison qui contribueraient à la diversité allélique de ces gènes (avec l'action de la sélection balancée fréquence dépendante) (Wutzler et al., 2012). Par exemple chez les Rhacophoridae, 18.5% des allèles des gènes de classe I du CMH (issus d'évènements de duplication) sont générés par recombinaison inter-locus (Zhao et al., 2013). Ces évènements de recombinaison sont décrits comme permettant d'entretenir la complexité du système immunitaire au niveau de l'assemblage des régions exoniques et jouent un rôle crucial pour façonner la diversité de 27 allèles de class I du CMH. Ainsi, dans notre cas, il semblerait que la recombinaison joue aussi un rôle important dans l'histoire évolutive de cette famille multigénique même si ces évènements ne changent pas l'architecture génétique du précurseur protéique en ne modifiant pas l'ordre des exons (recombinaison uniquement gardée dans les régions introniques). Ces évènements de recombinaison pourraient être responsables cependant de la propagation de certaines mutations partagées entre duplicats, et que l'on retrouve dans la cartographie des

mutations de la prorégion (A60T, Q68E) et du domaine BRICHOS. Le plus probable serait en effet que ces mutations soient transférées entre paralogues par des évènements de recombinaison (appelé «*shared polymorphism*») qu'il s'agisse de conversion génique entre gènes dupliqués en tandem ou bien par 'crossing-over inégal' (Fawcett and Innan, 2011), et gardées par la sélection naturelle, soit pour leur valeur intrinsèque, soit pour augmenter la diversité allélique de l'espèce vis-à-vis d'un mécanisme de reconnaissance et/ou de stabilité thermique. Une autre hypothèse pour ce polymorphisme partagé pourrait cependant être une rétention de polymorphisme ancestral par sélection balancée notamment pour certaines mutations qui sont également partagées par l'espèce sœur *A. caudata* (par exemple le A60T dans la prorégion) et qui pourrait expliquer en partie les 'hot spots' de mutations non-synonymes situées de part et d'autre de ce site.

Analyse de la différenciation génétique le long de la dorsale est pacifique (EPR)

Malgré le faible nombre d'allèles recapturés dans les populations d'*Alvinellas*, nous avons également cherché à savoir si certains allèles des différents gènes de la préproalvinellacine étaient capables de franchir la barrière aux flux de gènes trouvée à l'équateur entre les populations de l'EPR sud et nord. Cette barrière semble jouer un rôle important dans la différenciation géographique de nombreuses espèces (Plouviez et al., 2009) et notamment *A. pompejana* (Hurtado et al., 2004; Plouviez et al., 2010; Jang et al., 2016). Une absence de différenciation génétique de part et d'autre de cette barrière pourrait alors traduire des pressions de sélection favorisant le maintien de ces lignées à des fréquences optimales en l'absence de migration. L'analyse de différenciation nord/sud des populations indique que cette barrière n'a pas le même effet selon le paralogue de la préproalvinellacine considéré. Alors que pour le paralogue 1, aucune différenciation génétique entre les populations nord et sud n'est mise en évidence, les populations apparaissent bien différenciées au paralogue 4. En effet, de façon opposée au paralogue 1, les indices de fixation Phi_{st} et F_{st} sont significativement différents de zéro pour le paralogue 4 avec des tests exacts de non différenciation non significatifs. Ceci est corroboré par la forme des arbres et réseaux pour lesquels le paralogue 1 montre clairement un mélange d'allèles entre le nord et le sud avec une valeur de Dxy plus faible pour ce paralogue. Le paralogue 4 quant à lui possède des allèles diagnostiques divergents sur plusieurs mutations entre le nord et le sud même si les données ne montrent pas de monophilies réciproques entre les 2 populations. Ceci

indiquerait que les allèles codant pour la préproalvinellacine introgressent beaucoup plus entre les 2 entités génétiques (nord et sud) dans le cas du paralogue 1 que dans le cas du paralogue 4 avec une réduction plus franche de la diversité génétique au sud qu'au nord.

La phylogéographie comparée de plusieurs espèces hydrothermales a montré la correspondance de plusieurs barrières génétiques entre les populations nord et sud de la dorsale est Pacifique (EPR) (Hurtado et al., 2004, Plouviez et al., 2009). La séparation de ces peuplements s'est effectuée il y a au moins 1-2 millions d'années par l'apparition d'une série de failles transformantes (failles transformantes Discovery/gofar le long de l'EPR) à l'équateur renforcée par l'effet d'un gyre profond perpendiculaire à l'axe de la dorsale, et constitue une barrière aux flux de gènes chez de nombreuses espèces hydrothermales. Chez *A. pompejana* cela se traduit à la fois par une divergence de 1% sur le gène mitochondrial mtCOI (quasi-monophylie réciproque entre les individus du nord et sud de l'EPR) et une différenciation géographique de part et d'autre de la barrière équatoriale sur plusieurs locus nucléaires dont celui de la phosphoglucomutase (PGM) (Plouviez et al., 2010).

Dans cette étude, il apparaît que des allèles puissent être capables de traverser la barrière comme dans le cas plus spécifique du paralogue 1 mais également au niveau du paralogue 4 ou l'haplotype majoritaire au sud a été retrouvé chez un individu du nord. Ceci pourrait donc corroboré les résultats de migration/introgression entre les 2 populations sans que cette étude puisse statuer sur une orientation des flux de gènes. L'utilisation du logiciel IMa par Plouviez et al. (2010) a permis de révéler l'existence d'une migration (légèrement) asymétrique pour certains marqueurs nucléaires à travers la barrière avec une zone de transition qui serait une zone de contact secondaire (i.e. existence d'individus introgressés/hybrides pour le gène de la PGM) aux alentours de 7°S. Notre étude supporte donc l'hypothèse d'une barrière semi perméable aux flux de gènes avec une introgression d'allèles plus ou moins prononcée selon le paralogue de la préproalvinellacine testé.

Les données observées sur le paralogue 4 sont assez proches de celles déjà trouvées pour le gène *mt Cox1* puisque les deux études menées par Jang et al., 2016 et Plouviez et al., 2009 s'accordent sur un haplotype majoritaire au sud avec une couronne d'haplotypes dérivés récents (forme en étoile), alors que la population nord présente des haplotypes assez divergents les uns des autres. Ce réseau a une forme en étoile diagnostique d'un événement

d'expansion. Les études de Plouviez et al. (2009, 2010) suggèrent une expansion géographique dans le sud de l'EPR au cours des 500 000 dernières années qui être expliquée par des évènements d'extinctions/recolonisation plus fréquents dûs à une activité tectonique plus intense dans l'EPR sud. Cette expansion au sud a été confirmée par l'observation d'un réseau en étoile et des tests de Tajima et de Fu & Li significativement négatifs pour toutes espèces étudiées. A la lumière de des précédents résultats, il semblerait donc que le paralogue 4 présente bien la signature phylogéographique attendue par l'histoire évolutive de l'espèce alors que celle du paralogue 1 est atypique et pourrait supporter l'idée d'une introgression adaptative.

Chez les moules intertidales du genre *Mytilus* qui forment une zone hybride en mosaïque, des études portant sur les flux géniques des gènes codant pour deux peptides antimicrobiens (mytiline B et défensine MGD2) ont permis de mettre en évidence les échanges inter-génomiques (i.e. échanges d'allèles entre génomes) liés à l'adaptation locale dans un cas complexe d'hybridation d'espèces après remise en contact (Boon et al., 2009). En effet, cette étude a montré une absence de différenciation génétique pour ces 2 PAMs entre les espèces *edulis* et *galloprovincialis*, séparées depuis un million d'année et remises en contact il y a 25 000 ans, alors que celles-ci sont fortement différenciées au niveau de leur « background » nucléaire. A partir de ces données, les auteurs ont conclu à une introgression (supposée adaptative) secondaire des peptides antimicrobiens analysés, les 2 espèces présentant un avantage sélectif à partager leurs « pools » alléliques pour assurer leur défense vis-à-vis des agents pathogènes. Dans notre cas, il semblerait que l'alvinellacine qu'un flux important de gènes persiste au niveau de l'alvinellacine malgré l'évènement de spéciation en allopatrie d'*A. pompejana* nord et *A. pompejana* sud. Ceci pourrait entre autre être notamment dû au fort potentiel dispersif de l'espèce qui présente des œufs riches en réserves vitellines avec un développement lécitotrophe pouvant être suspendu dans les eaux abyssales froides. Ces caractéristiques lui permettent d'être présent dans la colonne d'eau sur de longues durées et être parmi les premières espèces colonisatrices sur les cheminées hydrothermales nouvelles (Pradillon et al., 2005).

Un peptide antimicrobien sous forte sélection purifiante

Chez *A. pompejana*, le peptide antimicrobien *sensu stricto* est strictement monomorphe pour les 6 paralogues mis en évidence et ne présente aucune divergence fixée entre paralogues : aucune diversité génétique n'est mise en évidence dans cette région du gène sur l'ensemble du jeu de données, même entre individus de localités géographiques différentes séparées par une barrière semi-perméable aux flux de gène. De plus, les séquences des PAM des deux espèces d'*Alvinella* divergent uniquement d'un seul acide aminé bien que la spéciation entre les deux espèces date de plusieurs millions d'années (23% de divergence sur le mitochondrial). Cette absence de polymorphisme suggère donc l'action d'une sélection purifiante extrêmement forte sur le PAM qui pourrait être liée à l'environnement contraignant dans lesquelles les deux espèces ont évoluées. Dans le cas des deux espèces de moules côtières *M. galloprovinciallis* et *M. edulis*, il a été montré que ces deux espèces, qui ont divergés depuis 1 million d'années et sont entrées secondairement en contact il aurait environ 25 000 ans, ont désormais un pool d'allèles communs pour les deux défensines analysées avec une divergence entre variants issus de l'isolement initial nettement supérieure à la divergence de l'alvinellacine trouvée entre *A. caudata* et *A. pompejana*. Ceci semble suggérer que, sauf si un flux de gènes persiste encore entre les deux espèces, le PAM a été fortement contraint (paramètres biotiques et/ou abiotiques) de façon indépendante chez les deux espèces pour, au final, ne montrer qu'un faible taux de divergence entre elles. Comme il a été montré qu'au moins un phylotype dominant d'epsilon protéobactérie était partagé entre les épibiotes d'*A. pompejana* et d'*A. caudata* (Cary et al., 1997), il est très probable que la pression de sélection exercée par l'épibiose soit la même entre les 2 espèces. Ces deux espèces d'*Alvinella* vivent en effet en syntopie sur les parois de cheminées hydrothermales de la dorsale Est Pacifique ce qui suggère fortement que l'interaction hôte-symbionte n'aurait que peu évoluée depuis la mise en place de l'épibiose. La forte spécialisation de la microflore hydrothermale pour un environnement très particulier qui n'a que peu changé au cours du temps et ce, malgré les très fortes variations du fluide hydrothermal (en température, en flux..) aurait pu geler les interactions hôte/pathogènes. L'action de la sélection purifiante sur l'évolution de plusieurs orthologues du PAM *Defb103* (défensines) a pu être mis en évidence chez les primates dont la fonction de défense a été fixée très tôt dans l'évolution des mammifères (Crovella et al., 2005). Dans

notre cas, cela pourrait donc signifier que la fonction du PAM a également été fixée très tôt dans l'évolution au sein d'espèces qui se sont adaptées il y a très longtemps au milieu hydrothermal et à son microbiome qui resterait alors stable dans le temps et l'espace (Danovaro et al., 2014). Grzymski et al. (2008) suggèrent sur la nature des épibiotes d'*A. pompejana* que « *The success of Epsilonproteobacteria as episymbionts in hydrothermal vent ecosystems is a product of adaptive capabilities, broad metabolic capacity, strain variance, and virulent traits in common with pathogens*". Ces propriétés leur permetteraint de prospérer dans cet environnement aux régimes thermique et chimique changeants depuis des centaines de millions d'années. Cette diversité métabolique des épibiontes constitue donc un avantage adaptatif indéniable pour les 2 espèces d'*Alvinella* qu'il a été/est nécessaire de préserver. Dès lors, la modalité d'évolution de l'interaction durable serait plutôt de préserver un état d'équilibre ou toute variation serait bannie plutôt qu'une course à l'armement qui par un chevauchement spatio-temporel de balayages successifs (ou sélection balancée fréquence dépendante) contribue plutôt à augmenter la diversité génétique qu'à la restreindre.

Dynamique évolutive d'un précurseur protéique paradoxalement très polymorphe

Alors que le PAM *sensu stricto* ne présente aucun variant au sein de l'espèce *A. pompejana*, son précurseur lui apparaît extrêmement diversifié du point de vue de sa structure primaire. Cette différence entre la diversité génétique trouvée au niveau du PAM et de son précurseur peut apparaître paradoxale et traduit vraisemblablement des régimes sélectifs différents entre domaines. Cette étude avait pour but de déterminer si l'action de la sélection positive peut avoir eu un rôle dans l'évolution du précurseur protéique et si celle-ci a affecté plus spécifiquement l'un des domaines du précurseur protéique.

Le ratio du nombre de mutations non-synonymes par site non-synonyme sur le nombre de mutations synonymes par site synonyme permet de renseigner sur l'action de la sélection positive qui agit sur le gène puisque celle-ci se traduit par un excès de mutations non-synonymes à certains codons (mutations non-synonymes peu ou pas contre sélectionnées). Les données obtenues le long du précurseur protéique montre que ce ratio évolue vers des valeurs approchant 1 (évolution presque neutre) mais rarement supérieures à 1 au niveau des différents domaines précédant le PAM. Ceci semblerait indiquer que

l'évolution des duplicats serait due à un relâchement des pressions de sélection surtout depuis qu'une étude récente (Fontanillas et al., 2017) a permis de mettre en évidence que les protéines des espèces d'*Alvinella* évoluaient à travers une forte pression de sélection purifiante avec un d_N/d_S n'excédant jamais 0,025.

Dans notre cas, cette évolution affecte plus spécifiquement les domaines de la prorégion et le domaine BRICHOS avec des valeurs de 0,83 et 0,84 respectivement. Cependant, le domaine BRICHOS montre un pic de d_N/d_S (associé à la divergence inter-paralogue) sur une zone limitée à une vingtaine de codons. La cartographie des mutations du BRICHOS sur l'arbre des paralogues montre que cette sélection positive a plutôt permis la diversification des différents duplicats puisque les mutations non-synonymes trouvées dans cette partie du BRICHOS sont celles retrouvées dans les branches internes de l'arbre. Ainsi, les mutations 129N, 131T et 133D sont diagnostiques des différents clades alors que les autres mutations du domaine BRICHOS sont partagées entre les différents clades (cf. V169 - «*shared polymorphism*») ou diagnostiques d'un seul gène paralogue. Ceci est également vrai pour la prorégion dont les mutations non synonymes significativement détectées comme étant sous sélection positive, sont concentrées au début du domaine et constituent également des mutations diagnostiques des différents paralogues dans l'arbre de reconstruction des séquences ancestrales même si des mutations partagées sont également retrouvées pour ce domaine.

Ce type de signature est en général souvent associé à de la sélection balancée autour d'un ou quelques remplacements d'acides aminés clefs impliqués dans le maintien d'un polymorphisme (Hudson et al., 1987; Kreitman and Hudson, 1991). Cependant, à l'inverse d'un gène mono-locus où la sélection maintient plusieurs lignées alléliques au cours du temps (Papot et al., 2016), cette sélection semble ici avoir lieu entre différentes copies d'un même gène. Comme présenté en introduction, dans le cas des gènes de l'immunité, la sélection balancée surtout de type fréquence-dépendante est répandu et se produit lorsque l'aptitude d'un génotype dépend de sa fréquence dans la population et/ou est le reflet d'une sélection naturelle qui fluctue dans le temps et l'espace en réponse à des décalages dans la diversité des pathogènes. Ce type de sélection est également largement rapportée par exemple pour certains gènes du CMH chez de nombreux organismes (Aguilar et al., 2004; Spurgin and Richardson, 2010; Zhao et al., 2013; Zeng et al., 2016; Hughes and Yeager, 1998;

Ota et al., 2000) ou elle est responsable du maintien du polymorphisme. Bien que cette sélection soit largement documentée pour les gènes de l'immunité adaptative notamment à cause de la co-évolution hôte-pathogène, l'action de la sélection positive n'a pu être démontrée significativement que sur l'évolution des familles multigénique codant pour les peptides antimicrobiens chez les vertébrés (Hollox and Armour, 2008; Hughes and Yeager, 1997b; Tennessen, 2005). Chez les invertébrés, ce n'est que très récemment que des hypothèses de polymorphisme trans-spécifique ont commencées à être avancées grâce à l'observation de cas de convergence évolutive entre espèces proches chez de nombreux PAMs (Unckless and Lazzaro, 2016; Unckless et al., 2016). Pour la préproalvinellacine, cette sélection semblerait jouer un rôle dans la création et le maintien de la diversité des domaines BRICHOS et de la prorégion (en plus du cas de polymorphisme trans-spécifique) sans qu'il y ai de coévolution avec le peptide antimicrobien puisque celui-ci est sous sélection purifiante.

Cette vision d'une sélection balancée entre plusieurs copies d'un même gène n'est néanmoins pas supporté par l'analyse CodeML des différents variants et par les tests de MacDonald & Kreitman (utilisant la double information des sites non-synonymes fixés et polymorphes). Ces méthodes ne permettent pas en effet de mettre en évidence de manière significative l'action de la sélection positive sur les codons du domaine BRICHOS mais supportent celle d'une évolution positive pour la prorégion avec 3 sites impliqués dans la divergence des paralogues sous sélection positive. L'analyse codeML détecte bien 9 sites sous sélection positive dans le domaine BRICHOS avec la méthode NEB mais pas avec la méthode BEB, méthode plus robuste qui élimine les faux positifs lorsqu'une molécule évolue de façon presque neutre. Il est néanmoins très probable qu'une partie de la prorégion co-évolue de manière positive avec le domaine BRICHOS. Globalement, le manque de significativité des tests pourrait être dû aux évènements de recombinaison récurrents qui ont pu propager certaines mutations 'adaptatives' entre paralogues et, que l'action de la sélection balancée pourrait s'effectuer non pas sur les allèles d'un paralogue mais entre paralogues.

Cette évolution presque neutre du précurseur s'oppose de toute façon à la forte sélection purifiante sur le PAM lui-même. Ainsi, contrairement à ce qui aurait pu être attendu et documenté dans la littérature, toute la variation génétique est ici regroupée dans le

précurseur protéique : au sein des domaines BRICHOS et prorégion. Le peptide signal montre quant-à-lui des valeurs de d_N/d_S faibles proches de zéro pour certaines comparaisons (valeurs maximales pour les comparaisons avec le paralogue 5). Globalement, cette région pourrait donc se retrouver sous sélection purifiante pour les paralogues 1-2-3a-3b-4, comme le PAM et lié au mode d'action de celui-ci (sécrétion vers le milieu extérieur). Ainsi les variants du domaine BRICHOS et de la prorégion pourraient être maintenus par l'action d'une sélection diversifiante qui agit sur une famille multigénique complexe composée de 6 gènes dupliqués en tandem dans laquelle les différentes copies s'influencent par recombinaison/conversion génique (d'où la difficulté de trancher sur l'action de la sélection balancée).

En conclusion, il s'agirait ici d'un maintien du polymorphisme des différentes copies tout en recombinant pour former de nouveaux allèles et propager les mutations favorables et augmenter ainsi la diversité au sein de chaque paralogue en s'approchant donc d'un mécanisme sélection balancée fréquence-dépendante. Une telle évolution est assez proche du complexe majeur d'histocompatibilité ou, recombinaison, sélection balancée et duplication génique sont les moteurs évolutifs qui contribuent à la diversité des gènes qui constituent ce système complexe de l'immunité adaptative (Zhao et al., 2013).

Le domaine BRICHOS et la dynamique de repliement du PAM

Quand la chaîne polypeptidique est synthétisée, les régions hydrophobes sont exposées et peuvent interagir directement avec d'autres régions hydrophobes du compartiment cellulaire, ce qui peut entraîner un problème de conformation de la protéine (Radford, 2000) ou encore la formation d'agrégats protéiques (e.g. plaques amyloïdes). Les chaperons moléculaires représentent un des moyens pour que la conformation des protéines soit optimale (Ellis and Van der Vies, 1991). Ces molécules sont définies comme : « *a large and diverse group of proteins that share the property of assisting the noncovalent folding and unfolding and the assembly and disassembly of other macromolecular structures, but are not permanent components of these structures when they are performing their normal biological functions*" (Ellis, 2006). Le domaine BRICHOS est un chaperon moléculaire qui empêche l'agrégation des peptides qui lui sont associés en feuillet beta, cette fonction ayant été mise en évidence à de nombreuses reprises (Peng et al., 2010; Sánchez-Pulido et al., 2002;

Willander et al., 2011). En général, les chaperons moléculaires sont inductibles suite à des conditions de stress (température et pression) comme cela a été montré chez les protéines HSP70 qui font partie des chaperons moléculaires les plus étudiés (Feder and Hofmann, 1999). Il apparaît que le transcrit codant le précurseur protéique de l'alvinellacine est fortement surexprimé (presque 350 fois la condition 'contrôle') lors d'un choc thermique (2h à 55°C) et lors d'une décompression des animaux de la même façon que les HSP70. Ces données supportent l'hypothèse d'une forte induction de la préproalvinellacine en conditions de stress en réponse à un choc entraînant la déstabilisation voire la dénaturation de nombreuses protéines. De plus, puisqu'il a été montré dans le cas de *Bri2* que ce domaine est clivé par une furine et l'ADAM10 en milieu extracellulaire, le BRICHOS pourrait avoir un rôle non négligeable dans la reconfiguration de l'alvinellacine après sa sécrétion dans l'environnement (tube) d'*Alvinella*. L'environnement hydrothermal se caractérise en effet par un pH acide, de fortes concentrations en métaux (Cu, Zn, Fe, Cd…) et de fortes températures qui varient fortement au cours du temps. Ces conditions sont documentées comme favorisant l'agrégation des peptides en feuillets beta (Miura et al., 2000). Ainsi, le BRICHOS pourrait assurer dans un premier temps un repliement correct du PAM dans le milieu intracellulaire puis continuer à exercer ce rôle dans le milieu extracellulaire après le clivage du PAM secrété. Le polymorphisme du BRICHOS refléterait ainsi une réponse à l'environnement dans lequel l'organisme a évolué surtout depuis qu'il a été montré que l'adaptation des organismes aux fortes températures inclus entre autre la protection des protéines au sein de l'environnement cellulaire par des molécules chaperones (Baross and Holden, 1996). Les différents variants du domaine BRICHOS pourraient donc contribuer à la stabilisation du peptide antimicrobien sur une large gamme de conditions environnementales dans le but de maintenir sa conformation en épingle à cheveux indispensable à son activité. Ceci permettrait *in fine* de maintenir une immunité externe efficace et le contrôle efficace de l'épibiose obligatoire en conditions hypervariables. Habituellement, l'évolution rapide du compartiment microbien (comme ce qui est généralement documenté dans les relations hôte/parasites) entraîne une coévolution rapide du système immunitaire de l'hôte, et la mise en place d'une sélection balancée sur les allèles/copies d'un gène qui contribuerait au maintien du polymorphisme des PAMs (Rolff and Schmid-Hempel, 2016; Unckless et al., 2016). Dans le cas particulier de la préproalvinellacine dans un environnement caractérisé par une instabilité constante, ce

phénomène s'observe sur des régions du précurseur autres que le PAM qui demeure monomorphe. De plus, bien que nos analyses de détection de l'action de la sélection positive sur l'évolution des paralogues ne soit pas totalement conclusive pour le domaine BRICHOS, les quelques codons trouvés sous sélection positive de manière significative dans la prorégion et possiblement en déséquilibre de liaison avec les sites du BRICHOS accréditent l'hypothèse d'une diversification adaptative des paralogues avec notamment la présence d'un polymorphisme trans-spécifique pour certains sites.

CONCLUSION

Cette étude a permis de mettre en évidence un comportement évolutif loin des attendus notamment car les duplications décrites pour les gènes codants des précurseurs protéiques de PAM et la sélection positive détectée, entre autre, sur la région du PAM est caractéristique d'un processus adaptatif dans la course à l'armement entre hôte et ses pathogènes (Tennessen, 2005) et/ou représente un avantage adaptatif vis-à-vis d'éventuelles modifications des paramètres (a)biotiques du milieu dans lesquels les espèces évoluent. En effet, une forte diversification non synonyme des paralogues – et non neutre – au sein des deux domaines du précurseur (BRICHOS et Prorégion) contraste fortement avec l'absence totale de diversité génétique dans le peptide antimicrobien. De plus, le faite que le peptide antimicrobien de chacune des espèces sœurs ne divergent que d'un seul acide aminé pourrait être un indice concernant l'intérêt évolutif à garder la fonction fixée tôt dans l'évolution de ce peptide antimicrobien via l'action de la sélection purifiante. Ceci serait donc le signe de l'impact d'un environnement particulièrement sélectif avec des communautés microbiennes possédant une diversité métabolique telle que ces espèces dépendantes de leur symbionte ont un avantage certain à la mise en place d'une interaction durable avec ces partenaires. De plus, ce peptide antimicrobien est secrété par le tégument de l'animal et se retrouve directement en contact avec les symbiotes dans l'environnement hydrothermal (Tasiemski et al., 2014). Du point de vue fonctionnel, l'intérêt de diversifier de façon significative la proregion et le domaine BRICHOS (une chaperonne) pourrait signifier que les différents paralogues veilleraient à ce que PAM soit toujours bien conformé qu'importe les paramètres hypervariables du milieu hydrothermal (impact des pressions abiotiques sur l'évolution d'un gène codant un peptide antimicrobien.

Chapitre 3. Diversité génétique et histoire évolutive du gène codant le précurseur protéique de la preprocapitellacine chez l'annélide côtier *Capitella spp.*

INTRODUCTION

Au sein des communautés benthiques, les polychètes côtiers jouent un rôle majeur notamment en terme de recyclage de la matière organique, en retravaillant les sédiments (bioturbation) et en constituant des espèces proies pour les oiseaux et les poissons (Dhainaut and Scaps, 2001). Ils sont souvent dominants à la fois en termes de biodiversité mais aussi en abondance et, leur flexibilité trophique ainsi que la diversité de leurs traits d'histoires de vie leur permettent de s'adapter aux conditions locales des habitats des plus perturbés (Giangrande et al., 2005; Hutchings, 1998). Ces polychètes, souvent associés à des habitats riches en matière organique et plus ou moins anoxiques, se retrouvent en effet en contact avec une large variété de polluants nocifs dont une partie provient des activités humaines (DelValls et al., 2002). Historiquement, de nombreuses études toxicologiques ont permis d'étudier l'effet d'une grande diversité de composants toxiques (et/ou polluants) sur les organismes marins, et notamment les polychètes. Certains métaux (bien que de nombreux éléments jouent un rôle important sur les fonctions cellulaires des organismes, e.g. Fe, Cu, Se, Mn) sont documentés comme étant particulièrement toxiques pour les organismes marins et les polluants chimiques sont la cause de mortalités chez les invertébrés marins (Southward and Southward, 1978). Cependant, chez les mollusques, et les organismes filtreurs en particulier, de fortes concentrations en métaux lourds (Cu, Cd, Zn, Hg) ont été mesurées dans les tissus, entraînant chez ces organismes la mise en place de mécanismes rapides de détoxication (revu par Viarengo & Nott 1993). Par exemple, l'utilisation des métallothionéines, qui participent aux processus homéostasiques sur les concentrations des métaux physiologiques (Zn, Cu…), interviennent aussi dans la détoxication des métaux non essentiels (comme Cd ou Hg). Bien qu'aucune métallothionéine

n'ait été décrite chez les polychètes, il a été montré que l'exposition au cadmium mettait en jeu deux types de protéines chez *Hediste diversicolor* : une protéine de haut poids moléculaire appelée MPI (>67kDa) et une protéine de plus petit poids moléculaire : la MPII (environ 14kDa) (Demuynck & Dhainaut-Courtois 1993; Demuyncke *et* al. 1993). Cette dernière a été montrée comme jouant un rôle de détoxication dans le métabolisme d'*Hediste* en se liant aux métaux (Demuynck and Dhainaut-Courtois, 1994). Les invertébrés marins peuvent également stocker les polluants sous des formes moins toxiques : Ag sera par exemple stocké dans une forme biologiquement inactive dans le tissu conjonctif et les lysosomes des néphridies chez *Sabella pavonina* puis excrété à l'extérieur de l'animal (Koechlin and Grasset, 1988). Dans le cas de l'hydrogène sulfuré présent à faible concentration dans l'environnement, chez les annélides, l'oxydation par les mitochondries de ce composé en thiosulfate, moins toxique, constitue également une adaptation à la vie dans les milieux hypoxiques (Völkel and Grieshaber, 1996). A ces stratégies de détoxication, les organismes marins (dont les annélides) inféodés aux sédiments riches en agents toxiques, ont également développés une immunodéfense qui leur est spécifique. Chez les annélides, les réponses humorale et cellulaire sont altérées par des conditions considérées comme toxiques chez la plupart des organismes. En effet, les défenses immunitaires (à la fois cellulaire et humorale donc), étudiées principalement chez les oligochètes en conditions toxiques, mettent en évidence une altération de l'immunocompétence chez les animaux exposés (Dhainaut & Scaps 2001; Cuvillier-Hot *et* al. 2014 pour revues). Cependant, contrairement à la réponse cellulaire, qui est majoritairement inhibée, la réponse humorale est activée chez certains annélides. Par exemple, cette dernière est stimulée par les xénobiotiques puisque la présence de PCB (polychlorobiphényles) va augmenter l'activité du lysozyme chez *Lumbricus terrestris, Eisenia fetida et Eisenia hortensis* (Ville et al., 1995) alors que leur présence et celle de métaux lourds (Hg, Cd, Zn) conduit à l'inhibition de la phagocytose chez ces mêmes espèces (Fugère et al., 1996; Ville et al., 1995). Récemment, une étude réalisée sur le polychète *Hediste diversicolor* a permis de mettre en évidence l'impact de la pollution sur le statut immunitaire de ces annélides polychètes (Cuvillier-Hot et al., 2017). Cette étude montre, entre autre, chez des populations prélevées *in situ* une surexpression du gène codant un précurseur de peptide antimicrobien (preprohédistine) couplée à une activité antibactérienne plus forte chez les animaux récoltés dans des habitats pollués par rapport à ceux des zones non polluées. Ces auteurs en concluent qu'une

augmentation de l'immunité humorale de ce ver constitue une réponse spécifique aux sédiments pollués aux phtalates et éléments métalliques. En plus des mécanismes de défense aux composés toxiques des environnements, l'activation d'une immunité humorale peut constituer un moyen d'adaptation supplémentaire de ces organismes dans le cas de la mise en place de symbioses avec des bactéries chimioautotrophes et/ou détoxifiantes qui participera à la survie des organismes dans ces environnements. Ceci peut avoir un rôle protecteur notamment via des fonctions secondaires de détoxication des métaux lourds ou de toxines ou encore dans la production de molécules de défense vis-à-vis d'autres micro-organismes (Bright and Giere, 2005). Plusieurs études ont permis de montrer que des organismes marins vivants dans des environnements pollués peuvent être protégés des métaux toxiques via l'association avec des procaryotes qu'ils soient endosymbiotiques ou non. Par exemple, les insectes de la famille des chironomidés sont décrits comme étant particulièrement résistants à la pollution métallique et organique (Groenendijk et al., 2002) en raison d'une association avec une communauté stable de bactéries endosymbiotiques qui, en dégradant les métaux, leur permettraient de protéger ces insectes (Senderovich and Halpern, 2013). Chez l'annélide polychète *Capitella teleta*, une espèce soeur de l'espèce *C. capitata*, il a été également montré une plus forte survie et un taux de croissance plus élevé chez les individus vivant dans des environnements riches en sulfures en se nourrissant de bactéries chimiosynthétiques (Tsutsumi et al., 2001). Ces auteurs ont également mis en évidence que la production de matière organique issue de la chimiosynthèse bactérienne pourrait également constituer un facteur important qui contrôlera la distribution et la dynamique des populations de plusieurs espèces de *Capitella* notamment *C. iatapiuna*.

Ecologie et developpement d'une symbiose facultative chez *Capitella spp.*

Capitella capitata est un taxon communément retrouvé dans les sédiments sablo-vaseux réduits riches en matière organique et est décrite comme une espèce au comportement opportuniste (Grassle and Grassle, 1976). Cette espèce a la capacité de proliférer dans des environnements enrichis en matières organiques et constitue une espèce pionnière lorsqu'une perturbation du milieu marin est rapportée (Giangrande et al., 2005). Cette espèce est donc largement retrouvée dans les environnements pollués/perturbés (Bellan et al., 1988) ou elle est capable de rapidement former des populations denses. Cet annélide fouisseur dépositivore non sélectif est présent dans les premiers centimètres des sédiments

envasés et/ou eutrophisés infralittoraux d'Atlantique et de la Manche mais a été également décrit jusqu'en zone arctique (Blake et al., 2009) en passant par le Golfe du Mexique (Hilliard et al., 2016) et la mer du Japon (Tomioka et al., 2016; Tsutsumi and Kikuchi, 1984). Ces espèces sont également associées à une large variété d'habitats réduits : des sédiments jusqu'aux sources hydrothermales ou encore les carcasses de baleines échouées (Gamenick et al., 1998; Silva et al., 2016, 2017; Thiermann et al., 1997). Le genre *Capitella* représente un complexe d'espèces cryptiques qui présente une distribution mondiale. Les espèces morphologiques dont définies comme étant 'cosmopolites' et ne présentent que très peu de traits distinctifs d'un point de vue morphologique (stase morphologique). En effet, historiquement, l'espèce *C. capitata* (Fabricius 1780) a été décrite comme ubiquiste et supposée être la seule espèce de *Capitella* trouvée dans l'environnement vaseux côtier. Cette donnée a été contredite par une étude génétique réalisée sur sept populations localisées près de Woods Hole et Gloucester (Massachussets, USA) à l'aide d'allozymes (Grassle, 1980; Grassle and Grassle, 1976). Cette étude a révélé l'existence d'au moins 6 espèces cryptiques dans les échantillons. Ces espèces étaient semblables morphologiquement mais présentaient des traits d'histoire de vie et des modes de reproduction différents. Au moins 50 espèces (dont 19 nominales) sont désormais recensées en fonction de leurs caractéristiques génétiques, physiologiques, reproductives et développementales (Méndez et al., 2000). Il apparaît que la caractérisation des espèces s'effectue principalement partir du type développemental (direct ou indirect : larves planctotrophique ou lécithotrophique), du nombre de larves par ponte, de la taille des larves et de la durée des différentes étapes du développement, et non sur la base de critères morphologiques (Méndez, 2006; Méndez et al., 2000).

Ces espèces sont caractérisées par un court cycle de vie avec un potentiel de croissance élevé, ce qui leur permet de rapidement coloniser les habitats même si le consortium de *Capitella spp.* impliqué dans cette colonisation puisse être dû à une succession d'espèces ayant des caractéristiques de reproduction variables. La durée du cycle de vie de *Capitella* en laboratoire peut par exemple varier de 1-2 mois à 9 mois selon les espèces (Lardicci and Ceccherelli, 1994; Méndez et al., 1997; Tsutsumi et al., 1990). La longévité potentielle des espèces peut également varier de 5 mois au sein d'une baie estuarienne de Méditerranée à au plus 2 ans pour une population d'Angleterre (Warren, 1976). La maturité sexuelle est

atteinte après une période de temps de 1 à 4 mois (MÉNDEZ et al., 2000; Tsutsumi and Kikuchi, 1984). Il semblerait de plus que la saison de reproduction soit continue puisque les pontes s'effectuent de manière asynchrone tout au long de l'année (Méndez et al., 1997; Warren, 1976). Les espèces du genre *Capitella* sont des organismes itéropares qui se reproduisent de façon sexuée et il existe un dimorphisme sexuel entre mâles, femelles et hermaphrodites (Blake et al., 2009; Seaver, 2016). Les ovaires sont visibles par transparence chez les femelles matures et les mâles possèdent des crochets reproducteurs comme ceux décrits pour d'autres annélides polychètes telles qu'*Alvinella pompejana* (Desbruyeres et al., 1998; Jouin-Toulmond et al., 1997; Pradillon and Gaill, 2003). Les embryons et les étapes de développement précoce sont incubés au sein d'un tube ouvert des deux côtés construit par les femelles autour de leur propre corps. Au final, les études ont jusqu'à maintenant permis d'identifier, entre autre, des différences spécifiques en termes de mode de reproduction, types de larves et de gamètes, vitesses de développement et de dispersion, taille des adultes (Silva et al., 2017).

Le complexe *Capitella spp*. en Manche

La répartition des annélides au sein de la région Hauts-de-France a été établie d'après les suivis effectués dans le cadre du programme VERMER (financement FRB, région Nord Pas de Calais). Si potentiellement *Capitella spp.* peut se retrouver sur une grande partie du littoral de la Manche Orientale (région Hauts-de-France), elle est tellement inféodée aux sites très enrichis en matière organique qu'elle ne représente une biomasse non négligeable que dans des zones très restreintes : enceintes portuaires. De plus, sa présence varie selon la saison du fait de son opportunisme et de son cycle de vie relativement court (rapport VERMER 2017). Au niveau des densités moyennes le long des côtes des Hauts de France, cette espèce est rare (<1ind/m²) sauf au sein du Port de Boulogne-sur-Mer. Dans ce rapport VERMER, Dunkerque n'est pas répertorié mais les données de terrain montrent sa présence au sein du port de Dunkerque. Sur les zones portuaires bretonnes tel que le port de Roscoff, les *Capitella* sont présentes dans la plupart des sédiments envasés (comme le fond de l'aber de Roscoff : Laber) qui est une zone ouverte relativement peu exposée aux activités anthropiques mais se révèlent particulièrement nombreux dans des zones polluées par des dérivés de plastique (phtalates), des Eléments Traces Métalliques (ETM) mais surtout par des

dérivés de peintures avec le tributylétain (TBT), comme rapportés dans le port de pêche de Roscoff.

Dans ces zones fortement anthropisées, une proportion significative (environ 20% mais dépendant de la saison) de *Capitella* sont en association avec un tapis de microorganismes au niveau tégumentaire. Cette association pourrait traduire la présence d'une épibiose détoxifiante qui aurait un effet bénéfique sur le ver ou à une infection microbienne contractée à la suite d'une fragilisation des vers. Ceci implique dans les deux cas une modification du système immunitaire de l'espèce qui serait alors devenue soit tolérante à la mise en place d'une symbiose microbienne, soit sensible au développement d'un pathogène qu'il ne parviendrait plus à éliminer. Ces micro-organismes forment des structures filamenteuses très particulières correspondant à la bactérie géante sulfo-oxydante *Thiomargarita nelsonii* (Schulz, 2006) à gram négatif de la famille des γ-protéobactéries (Figure 1) dont l'identification a été réalisé par Sébastien Duperron (UMR 7138, UPMC, Paris 6). Cette bactérie immobile anaérobie facultative a par ailleurs été décrite sous forme libre ou associée au byssus d'une moule des sources hydrothermales (Salman et al., 2011) et possède des vacuoles contenant des nitrates essentiels à l'oxydation des sulfures. Cette association est originale puisqu'elle semble transitoire et facultative : elle n'est rencontrée que dans des zones enrichies en sulfures. Cependant, le fait que cette association soit perdue lorsque le milieu est appauvri sans que l'hôte ne soit *a priori* lésé (20 % des *Capitella spp.* associées à la bactérie sulfo-oxydante *in natura*) suggère que l'association est coûteuse pour l'annélide dans les environnements non pollués.

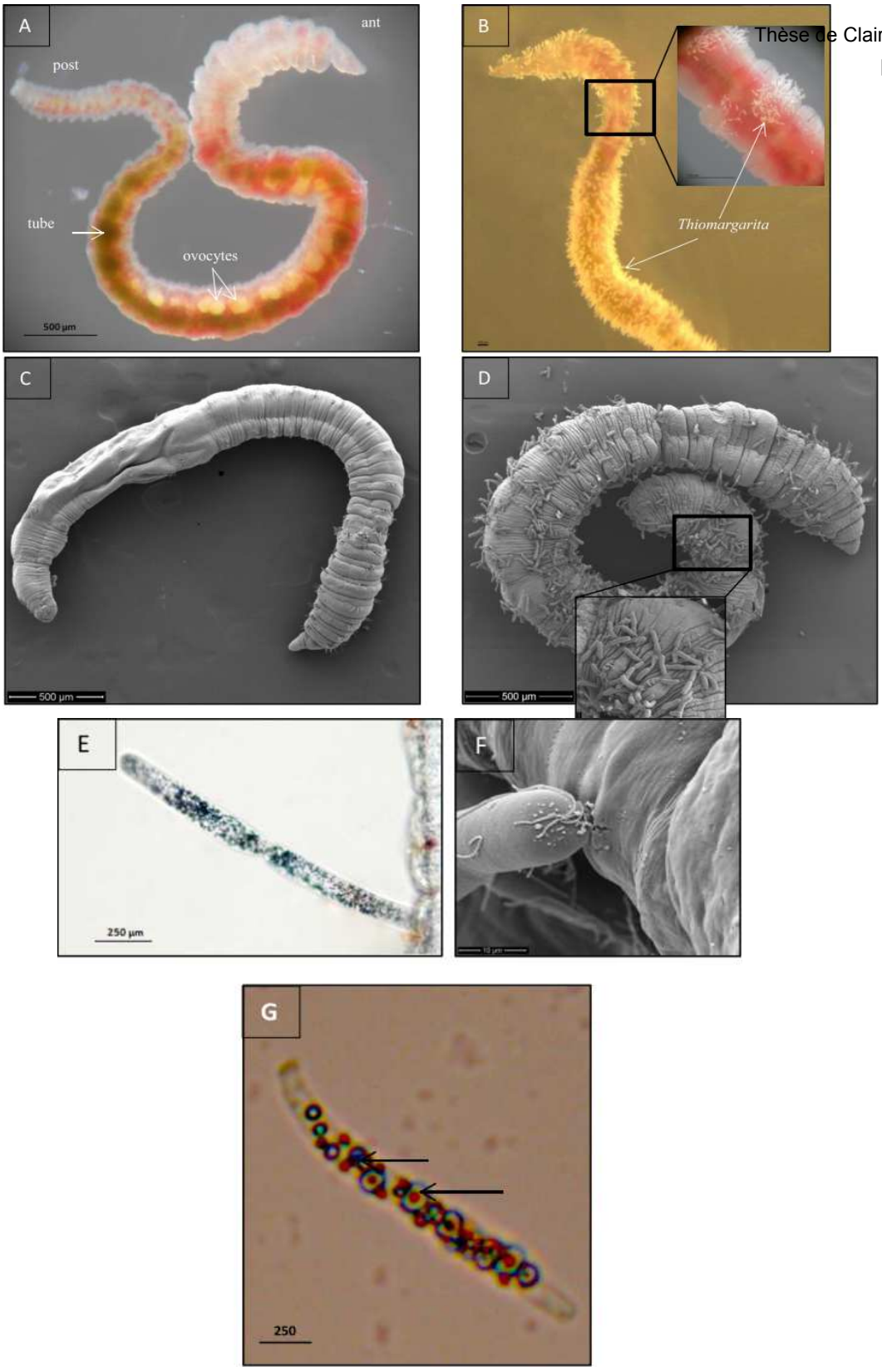


Figure 1. Photos de *Capitella spp.* sous loupe binoculaire (A) seule ou (B) associée à la bactérie sulfo-oxydante *Thiomargarita nelsonii*. Extrémités antérieure (ant) et postérieure (post). Images de microscopie électronique à balayage d'une *Capitella* (C) sans association et (D) en association avec *Thiomargarita* sur un individu récolté dans le port de Roscoff. (E) visualisation de *Thiomargarita* ancrée à la surface du tégument de *Capitella* en microscopie photonique d'une coupe en semi fine et (F) en microscopie électronique à balayage. (G) image en microscopie photonique de la bactérie *Thiomargarita* libre maintenue au laboratoire. Les vésicules (flèches) accumulent les sulfures (issu du stage M2 de Lolita Roisin, 2013).

Rôle du peptide anti-microbien capitellacine dans l'association avec la bactérie sulfo-oxydante.

Chez *Capitella spp.* à partir de données RNAseq (effectué au laboratoire), l'orthologue du peptide antimicrobien étudié au chapitre 2 avait été caractérisé en amont de cette thèse et cette molécule de 23 acides aminés ne montre aucune homologie de séquence avec l'alvinellacine bien que ces deux peptides antimicrobiens appartiennent à la même famille : les peptides 'arenicine-like'.Ils montrent quand même certaines homologies : de structure en se repliant tous les 2 en épingle à cheveux avec 2 ponts disulfures et d'architecture puisque ces deux PAM sont clivés d'un précurseur protéique comprenant un peptide signal, une proregion et un domaine BRICHOS en amont de la région du peptide antimicrobien en tant que tel.

Comme dans le cas d'*Alvinella pompejana*, le rôle du PAM dans l'immunité externe du vers et le contrôle de l'association entre *Capitella* et *Thiomargarita* a été investigué au laboratoire. Des données d'immunohistochimie *in toto* ont permis d'étudier la localisation tégumentaire de la capitellacine chez des *Capitella spp*. associées ou non à la bactérie *Thiomargarita* via l'utilisation d'anticorps anti-capitellacine en microscopie confocale. Ceci a permis de mettre en évidence une accumulation du PAM dans les cellules glandulaires de l'épithélium uniquement chez les animaux associés aux bactéries (C. Boidin-Wichlacz, A. Tasiemski, comm. pers., données non publiées). De plus, la sécrétion du peptide antimicrobien dans le milieu extérieur (puisque retrouvé dans le mucus) par les animaux associés à des bactéries a également été mise en évidence (Figure 2).

Le groupe a de plus récemment mis en évidence que ce PAM possède, à forte concentration, une activité antimicrobienne vis-à-vis de *Thiomargarita* et à plus faible concentration, une activité anti *Vibrio alginolyticus* et anti-*Bacillus*, deux bactéries pathogènes trouvées dans l'environnement du vers. Cela montre que le PAM participe à la défense antimicrobienne mais aussi pourrait aider au contrôle de l'ectosymbiose du ver.

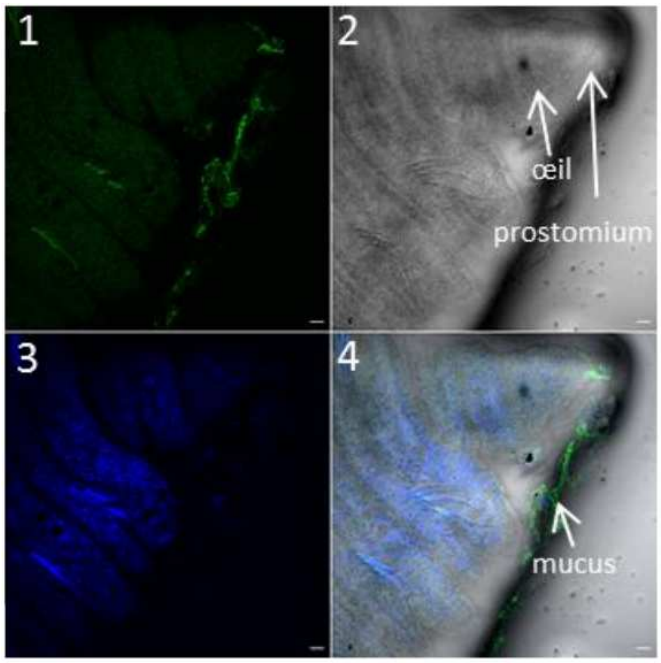


Figure 2. Images de microscopie confocale montrant la localisation de la capitellacine par immunohistochimie *in toto* chez une *Capitella spp.* associée à des *Thiomargarita*. (1) fluorescence FITC, (2) en lumière visible, (3) marquage nucléaire des bactéries au DAPI et (4) superposition des 3 images. Le marquage en couleur verte (FITC) montre les zones d'accumulation de la capitellacine. (Stage de Lolita Roisin, M2)

OBJECTIFS

Ce troisième chapitre permet d'effectuer un parallèle évolutif entre 2 modes d'association symbiotique en étudiant un deuxième peptide anti-microbien de la même famille (i.e. famille à BRICHOS) au sein d'une complexe d'espèces d'annélides côtières relativement éloignées d'*A. pompejana* présentant une association facultative avec des micro-organismes, potentiellement liée aux récentes pollutions environnementales d'origine anthropique.

Dans un premier temps, comme le genre *Capitella* constitue un complexe d'espèces cryptiques (Grassle and Grassle, 1976), il s'agira tout d'abord de clarifier la position taxonomique de nos échantillons de *Capitella* prélevés le long des côtes françaises de la Manche et de l'Atlantique Nord et évaluer leur degré de structuration génétique à l'échelle de la Manche afin de pouvoir mieux évaluer le rôle de la géographie et de l'environnement sur la distribution de ces polychètes.

Dans un deuxième temps, il s'agira de retracer l'histoire de la diversification génétique du PAM à BRICHOS dans la/les espèce(s) de *Capitella spp.* qui pourrait avoir un rôle dans la régulation de la prolifération de *Thiomargarita*.

Dus à des trajectoires évolutives distinctes, les espèces cryptiques peuvent posséder des adaptations uniques et doivent être considérées comme des unités évolutives séparées (OTU). En effet, dans de nombreux cas, ces espèces ont des préférences d'habitat (salinité, profondeur, température, type de substrat : Derycke et al., 2016; Knowlton, 1993) ou encore des phénologies différentes (Scriven et al., 2016) permettant de suggérer qu'elles occupent des niches écologiques différentes et donc des évolutions particulières dans lesquelles les interactions interspécifiques (de type symbiotique, antagoniste, ou parasite) peuvent jouer un rôle non-négligeable. Par exemple, chez des espèces cryptiques du nématode *Litoditis marina*, il a été montré que le microbiome avait un impact sur la distribution des espèces et notamment leur degré de sympatrie (Derycke et al., 2016). En effet, le microbiome diffère entre espèces, ce qui a un impact sur la physiologie de l'hôte influençant ses interactions écologiques et sa distribution.

Ainsi quatre questions principales peuvent être dégagées :

- l'espèce Manche/Atlantique forme-t-elle un complexe d'espèces cryptiques ? et si oui
  - i) quel est le statut phylogénétique de ces espèces par rapport au complexe mondial des *C. capitata* et *C. teleta ?*
  - ii) peut-on faire des hypothèses phylogéographiques concordantes avec les données préexistantes sur les refuges glaciaires dans cette région ?
  - iii) quel est le niveau de diversification du ou des locus codant la capitellacine et existe-t-il des échanges inter-génomes de ce PAM selon la géographie et l'environnement rencontré ?
  - iv) l'évolution du ou des gène(s) de la preprocapitellacine est-elle liée à des processus sélectifs récents, notamment à travers l'acquisition d'une symbiose facultative ?

Matériel et méthodes

1. Echantillonnage des *Capitella spp.*

Les animaux ont été prélevés à marée basse par tamisage des premiers centimètres de sédiments prélevés le long des côtes de la Manche pour lesquelles des populations de *Capitella* sp avaient déjà été référencées. De préférence, des zones riches en matières organiques présentant des petits trous à la surface du sédiment typique de galeries d'annélides ont été particulièrement choisies. Ces localités sont présentées sur la Figure 3 et le Tableau 1. Pour chaque localité, des bèches cube de sédiments ont été tamisées à des distances n'excédant pas quelques mètres pour pouvoir augmenter le nombre de *Capitella* récoltés. Une fois au laboratoire, la faune des sédiments a ensuite été triée sous loupe binoculaire et chaque *Capitella* a été préservée dans un eppendorf contenant de l'éthanol 90° en notant son caractère « parasité » ou « non parasité ».

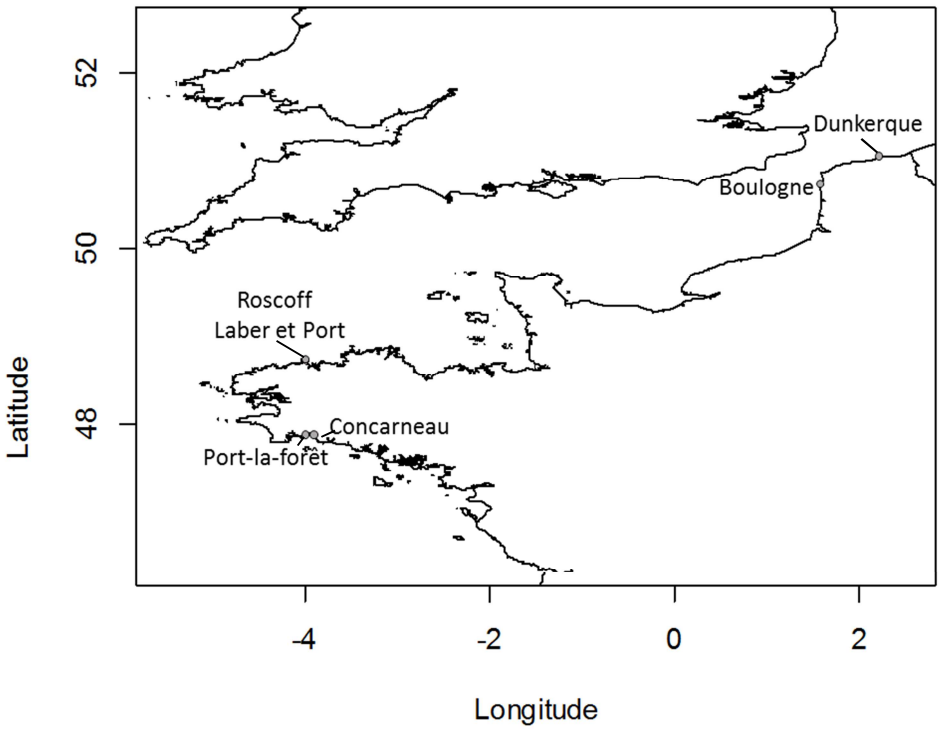


Figure 3. Position des localités d'échantillonnage des *Capitella spp.* le long des côtes de la Manche et de l'Atlantique nord.

L'échantillonnage des *Capitella* spp. a été effectué sur 5 sites principaux : zones portuaires de Dunkerque, Boulogne, Roscoff, Port-la-Forêt et de Concarneau ainsi qu'une zone non polluée dite du Laber à Roscoff. Le nombre d'individus séquencés (ainsi que leur statut parasité-non parasité) pour les deux gènes COI et Capitellacine sont récapitulés dans le Tableau 1.

COI		**Capitellacine**	
Roscoff- Port	30	Roscoff- Port	8
Roscoff- Port Parasité	13	Roscoff- Port Parasité	2
Roscoff- Laber	17	Roscoff- Laber	2
Roscoff - Laber parasité	6	Roscoff - Laber parasité	1
Port-la-Forêt	14	Port-la-Forêt	11
Boulogne	38	Boulogne	4
Dunkerque	36	Dunkerque	4

Tableau 1. Tableau récapitulatif des effectifs séquencés pour COI et Capitellacine (nombre d'individus par localité).

2. Gènes analysés

Notre étude s'intéresse à l'évolution moléculaire de la capitellacine mais nécessite l'utilisation d'un marqueur cytoplasmique non recombinant : le gène codant pour la cytochrome c oxydase 1 (*Cox-1*) pour identifier nos différentes espèces du genre *Capitella*, ce dernier constituant un complexe d'espèces cryptiques déjà connu (Grassle et Grassle 1976). Cette identification moléculaire des espèces fait cependant l'hypothèse qu'il n'y a pas eu de balayage sélectif trans-spécifique du génome mitochondrial.

L'intérêt de cette étude à 2 locus sera également de savoir quelle est la part des processus sélectifs liés à l'habitat/microbiome et de l'histoire démographique passée des différentes espèces sur la structure génétique actuelle du gène de la capitellacine. Il est admis qu'un effet démographique affecte l'ensemble du génome et que les processus sélectifs ont un effet plus local et agissent de façon indépendante sur les locus. A partir de l'éclairage mitochondrial, il s'agira plus particulièrement d'étudier les mécanismes évolutifs ayant donné lieu au polymorphisme moléculaire de la capitellacine et notamment d'identifier si ces espèces sont capables de d'hybrider et donner lieu à des introgressions d'allèles entre espèces proches, ces introgressions pouvant elles-même être adaptatives.

2.1. Cox-1

Le gène mitochondrial *Cox-1* code pour la sous-unité I de la cytochrome c oxydase. Cette enzyme est un élément indispensable de la chaîne respiratoire de la mitochondrie qui produit une grande partie de l'énergie (sous forme d'ATP) nécessaire à la cellule pour assurer le cycle de Krebs et la glycolyse. Du fait de son taux d'évolution rapide, l'absence de recombinaison et de sa facilité à être amplifié chez la plupart des taxons à l'aide d'amorces « universelles » (Folmer et al. 1994), ce gène a été choisi chez les animaux comme un barcode moléculaire idéal permettant l'identification rapide et relativement sûre des espèces en tirant partie de la signature spécifique de substitutions propres à chaque espèce (Hebert et al., 2003). Ce gène est en effet très résolutif à l'échelle de l'espèce ou du genre (Berry, 2006) mais son évolution n'est souvent pas neutre et liée à l'évolution du génome mitochondrial qui peut être différent de celui de l'espèce, l'héritabilité mitochondriale étant en général maternel sauf dans les cas particulier de double transmission uniparentale ou la lignée mâle est préservée dans les gonades des individus mâle (comme chez *Mytilus* : Fisher and Skibinski, 1990). Chez les invertébrés, ce gène ne constitue pas un marqueur idéal d'identification des espèces et doit être considéré avec prudence, car dans la mesure ou son produit joue un rôle très important sur le métabolisme aérobie des cellules, il est documenté comme pouvant être tantôt sous forte sélection purifiante (Powell and Somero, 1986) ou subir de fréquents balayages sélectifs chez les organismes présentant une grande taille efficace de population (en débat : Bazin et al., 2006; Berry, 2006). Ces effets de la sélection sont discutés comme pouvant conduire à une diminution de la diversité génétique ou à une accélération du taux d'évolution mitochondrial par rapport à l'évolution des gènes nucléaires et peut conduire à considérer 2 individus comme des espèces différentes alors qu'ils ne le sont pas ou, inversement à regrouper des espèces différentes sous le même nom (Berry 2006).

2.2. Capitellacine

A partir des données de RNAseq réalisé sur des *Capitella spp.* du port de Roscoff et du Laber, et de la séquence de la préprocapitellacine récupérée du génome de *C. teleta* (JGI), l'orthologue de l'alvinellacine a été identifiée sur l'une des différentes espèces de *Capitella* du port de Roscoff (Figure 4). Le peptide antimicrobien (23 acides aminés) est clivé d'un

précurseur protéique plus large (préprocapitellacine) comprenant également un peptide signal, une prorégion anionique ainsi qu'un domaine BRICHOS. Les deux peptides capitellacine et alvinellacine ne montrent aucune homologie de séquence mais ils appartiennent à la même famille : les peptides 'arenicine-like'. Ils se replient tous les 2 en feuillet beta avec 2 ponts disulfures.

Figure 4. Séquence nucléotidique codant la préprocapitellacine et positionnement des amorces présentées dans le Tableau 2 (en noir : séquences 5'F et 5'R et en gris : séquences 3'R et 3'R). En gras se trouve la région codante pour le peptide antimicrobien capitellacine, en bleu : le peptide signal, jaune : BRICHOS, le reste constitue la prorégion divisée en deux parties appelées après proregion 1 et proregion2 (cette dernière était monomorphe au chapitre 2).

A partir de cette séquence, des amorces ont été synthétisées pour amplifier le gène codant pour le précurseur protéique de la capitellacine (préprocapitellacine) en deux morceaux chevauchant : la région 5' (910pb) et la région 3' (930pb) (Tableau 2).

COI	
capitella COI-R	CCACCACCAGTAGGATCAAA
capitella COI F	GTACAGAACTTGCGGTTCCT
Genotypage PAM	
capitella genotypage F	GTCAAGCACGAGAATGGCTC
capitella genotypage F	ATTTCTTTCAGAGCAAAGTGGA
Capitellacine	
Capitellacine_5F	GGACATGGATAGCATGGAGA
Capitellacine_5R	GGAGAAGGGAGGTTGCG
Capitellacine_3F	GGAGCCATTCTCGTGCT
Capitellacine_3R	CGGTGTCGTCTTAAGTGTTCCA

Tableau 2. Amorces utilisées pour l'amplification des 2 gènes et du génotypage des individus au locus capitellacine (i.e. fragment contenant les sites diagnostiques des différentes lignées mitochondriales trouvées.

3. ACQUISITION DES SEQUENCES

3.1. Extraction de l'ADN génomique

L'ADN génomique total a été extrait sur 154 individus provenant des sites du Laber (Roscoff, Manche occidentale) du Port de Roscoff, de Boulogne (Normandie, Manche orientale) et de Dunkerque (Pas de Calais, Mer du nord) grâce au kit NucleoSpin Plasma XS de Macherey-Nagel en utilisant le protocole du fournisseur.

3.2. Amplification par PCR

Pour le gène mitochondrial *Cox-1*, l'ADN mitochondrial a été amplifié sur tous les individus échantillonnés avec l'enzyme Gotaq (Promega) en utilisant des amorces *Capitella* spécifiques (Tableau 2) définies à partir des séquences issues de l'assemblage des données RNAseq sur plusieurs individus échantillonnés sur le site du Laber (Roscoff). Les amplifications ont été réalisées dans un volume final de 25µL avec: 1X de tampon, 2mM MgCL2, 0.05mM dNTP, 0.4µM de chaque primer, 1U de Taq polymerase (Uptitherm, Interchim^{TM}). Le protocole d'amplification est le suivant : 95°C pendant 3 min ; suivi de 40 cycles de 95°C pendant 30 s, 56°C pendant 30 s, 72°C pendant 1 min et une élongation finale de 72°C pendant 10 min. Les amplificats sont visualisés sous lampe UV après électrophorèse sur gel d'agarose à 1% contenant du bromure d'éthidium.

Pour la préprocapitellacine, les régions 5‘ et 3’ du gène ont été amplifiées sur seulement 14 individus du Laber, 4 de Boulogne, 4 de Dunkerque, 11 individus de Port-la-Forêt pour limiter le nombre de clonages individuels. Les amorces utilisées pour amplifier ces deux parties du gène sont présentées dans le Tableau 2. Le gène a été amplifié avec la Taq Uptitherm (Promega) en suivant le protocole précédant et le protocole d’amplification suivant : un cycle de dénaturation à 95°C pendant 5 min suivi de 39 cycles de 95°C pendant 45 s, 55°C pendant 45 s, 72°C pendant 2 min et une phase d’élongation finale de 72°C pendant 10 min.

3.3. Séquençage direct pour les produits d’amplification Cox-1

Les produits PCR ont été séquencés sur un séquenceur automatique à 16 capillaires ABI 3100 (Applied Biosystems) au laboratoire EEP et les séquences obtenues dans les 2 sens ont été assemblées à l’aide du logiciel De Novo Assemble de Geneious (Drummond et al., 2010). Les séquences *Cox-1* consensus sont générées pour chaque individu et alignées entre elles pour l’ensemble des individus échantillonnés entre Concarneau et Dunkerque avec le module PairWise/Multiple alignement du logiciel Geneious.

3.4. Clonage et séquençage des clones pour la capitellacine

Pour la capitellacine, un clonage de produit de PCR a été effectué pour chaque individu avec un effort de recapture de 8 clones par individu et par région du gène (5’ et 3’) en partant du postulat de départ d’un système plus simple que décrit au chapitre 1 c’est-à-dire en considérant un nombre de duplications plus faible que celui de l’alvinellacine. Le clonage a été effectué avec un kit de TA cloning kit (Invitrogen) avec le vecteur pcr2.1 selon le protocole du fournisseur. Les clones ont été criblés avec les amorces M13F/M13reverse et séquencés à l’aide du séquenceur automatique à 16 capillaires ABI 3100. Les séquences F et R de chaque clone ont été assemblées à l’aide du package De Novo Assemble de Geneious. Les séquences consensus obtenues ont été toutes alignées à l’aide du package PairWise/Multiple alignement du même logiciel. Un alignement a été effectué pour chaque individu et les séquences identiques ont été enlevées du jeu de données. Sur la base des séquences recapturées, les recombinants artéfactuels entre deux séquences d’un même individu ont été visuellement enlevés également du jeu de données. Une fois le jeu de données épuré des recombinants artéfactuels, un alignement a été effectué sur l’ensemble des individus des différentes localités et les recombinants précédemment enlevé ont été

comparés entre tous les individus pour chercher d'éventuels recombinants naturels en considérant qu'un recombinant retrouvé chez au moins 2 individus différents n'était pas un artéfact de PCR/clonage.

Une élimination des mutations artéfactuelles a également été effectuée pour chaque individu en supprimant les singletons retrouvés au sein de chaque groupe allélique de séquences. Lorsque deux séquences appartenant à un même groupe d'allèles par une seule et unique mutation au sein d'un même individu, celle-ci a toutes les chances d'être une mutation artéfactuelle. Ceci permet d'obtenir un pourcentage moyen de mutations artéfactuelles par individu et d'appliquer un filtre sur les singletons dans l'alignement final à n individus.

3.5. Génotypage des individus sur le gène capitellacine par séquençage direct

Pour la capitellacine, un génotypage des individus a été effectué par séquençage direct sur un fragment diagnostique du gène ne contenant pas d'indel et sur lequel il est possible de distinguer les différentes lignées alléliques ('clades') observés. Des amorces spécifiques ont donc été produites dans une région conservée de la partie 5' qui possède les séquences diagnostiques choisies (18 sites polymorphes couvrant l'ensemble des divergences inter-clades) sur une longueur totale de 120 pb pour éviter des problèmes de recombinaison. Les séquences forward et reverse de chaque individu sont assemblées à l'aide du logiciel Geneious (module De Novo Assemble) et la présence de double-pics à chaque site polymorphe est évaluée visuellement pour quantifier le nombre de séquences différentes à chaque individu. Les séquences prédites par le génotypage sont ensuite alignées à celles issues des clones pour chaque individu afin d'estimer la diversité allélique individuelle au sein de chaque lignée mitochondriale.

4. Analyse des données

4.1. Phylogéographie du complexe Capitella spp. à l'aide du marqueur Cox-1

Toutes les séquences du gène mitochondrial de la cytochrome c oxydase I actuellement disponibles pour le genre *Capitella* (espèces *teleta* et *capitata*) au niveau mondial ont été récupérées dans les bases de données GenBank et DDBJ (DNA Data Bank of Japan) et un arbre des relations phylogénétiques entre les différents individus/espèces a été obtenu à l'aide du logiciel MEGA 6. Tout d'abord le meilleur modèle de substitution a été trouvé grâce au critère BIC du logiciel jModeltest : il s'agit du HKY+I+G. L'arbre en Maximum Likelihood a ensuite été réalisé sous Mega6.0 avec l'option 'Complete deletion'. Un arbre ML a été également été obtenu avec la même méthode en prenant uniquement en compte les deux premières bases de chaque codon.

La divergence nette entre les différents clades a été estimée en utilisant le modèle de substitutions K2P (Kimura, 1980) calculé sous MEGA v6 avec le package « ComputeNet between group mean distances » qui permet de calculer une estimation du nombre net de substitutions par site par groupe de séquences, ces groupes étant prédéfinis à l'aide de l'arbre obtenu précédemment.

4.2. Réseaux d'haplotypes de C. capitata-like Cox-1 sous PoPart

Dans un deuxième temps, le logiciel PoPart (Leigh and Bryant, 2015) a permis d'analyser et d'illustrer les différentes relations phylogénétiques qui existent entre les haplotypes du gène mitochondrial dans le cadre des populations françaises : Boulogne – Dunkerque – Roscoff (Laber et Port) – Port la Forêt – Concarneau, uniquement pour les espèces qui ne sont pas apparentées à l'espèce *C. teleta*. Pour cela, un réseau en « Neighbor Joining » est construit à partir de l'alignement réalisé sous Geneious. Un deuxième réseau est également réalisé à l'aide du logiciel Network (Bandelt et al., 1999) pour comparer les réseaux obtenus.

4.3. Barcode gap du complexe C. capitata sur les données Cox-1

L'analyse du barcode-gap du complexe *C. capitata* au sein des localités françaises a été effectué à l'aide du logiciel ABGD (Puillandre et al., 2012). Ce logiciel permet de sortir des groupes de séquences qui seraient des espèces/sous-espèces hypothétiques basés sur une augmentation brutale de la divergence entre groupes d'individus, et ce, à partir des

distances génétiques, lorsque la variation intra-spécifique est inférieure à la divergence inter-spécifique.

4.4. Courbes de mésappariement

Une représentation des données de polymorphisme de séquences en courbes de mésappariement (histogrammes de fréquences des différences observées entre paires de séquences) a été produite pour le gène *Cox-1* uniquement pour regarder l'histoire évolutive de chaque lignée mitochondriale de façon globale puis pour chaque lignée et chaque localité dans le cas où une structuration génétique significative puisse être rapportée. Cette analyse permet d'avoir une idée de la dynamique démographique des différentes « espèces » sous l'hypothèse d'une accumulation neutre des mutations dans le polymorphisme (sous l'hypothèse d'une absence de balayage sélectif sur le génome mitochondrial).

De la même façon, pour tester si la distribution des mutations accumulées dans le polymorphisme d'une espèce (courbes de mésappariement) s'écarte de l'attendu neutre dans un modèle de population Wright-Fisher, on utilise à la statistique R^2 de Ramos-Onsins and Rozas, 2002. Ce test compare les différences entre le nombre de mutations de type singletons et la moyenne des différences nucléotidiques qui en fait un puissant test pour détecter la croissance des populations.

4.5. Diversité génétique

Pour les deux gènes, les indices de diversité nucléotidique π (Nei, 1987) et θ_W (Watterson, 1975), le nombre de sites ségrégant (S) et la diversité haplotypique (Hd) ont été estimés à l'aide du programme DNAsp v5 (Librado and Rozas, 2009).

Pour le gène mitochondrial, les descripteurs suivants ont été calculés par lignée/clade et au sein de chaque localité dans chaque lignée. Pour la préprocapitellacine, ceux-ci ont été calculés par groupe de séquences alléliques/clade puis par localité au sein de chaque clade sous l'hypothèse qu'un clade représente une espèce.

L'indice de diversité S représente le nombre de sites variants dans un échantillon représentatif d'allèles trouvés dans la lignée et/ou la population. L'estimateur S (ou Eta) permet de calculer le théta de Watterson (θ_W) en rapportant le nombre de sites variables S

dans l'échantillon à la longueur de l'alignement. Cet estimateur est égal à 4Ne.μ sous l'hypothèse d'une accumulation neutre des mutations dans le polymorphisme. La diversité nucléotidique peut également être estimée à partir du nombre moyen de différences nucléotidiques observées entre paires de séquences pour un site nucléotidique donné. Cet estimateur noté π est lui aussi équivalent à 4Ne.μ sous l'hypothèse neutraliste mais est beaucoup moins sensible à un effet sélectif et/ou démographique passé.

Sur ces jeux de données (*Cox-1* et préprocapitellacine), plusieurs tests (D de Tajima, 1989) et Fs de Fu and Li, 1993 visant à détecter un écart à l'accumulation neutre des mutations dans le polymorphisme des *Capitella* ont été effectués à l'aide du logiciel DNAsp. Tajima (1989) a proposé un premier test de neutralité basé sur la différence des estimateurs θ_W et π sachant que θ_W est plus sensible à l'accumulation de variants rares que l'estimateur π. Sous hypothèse de neutralité, les deux estimateurs mesurent la même chose (θ=4Ne.μ), et la différence pondérée par la variance $D=(\pi-\theta_W)/\sigma 2(\pi-\theta_W)$ de Tajima est égale à 0. Dans le cas d'une population en expansion suivant un goulot d'étranglement et/ou d'une population ayant subi un balayage sélectif au locus considéré (i.e. fixation d'un allèle avantageux), cette valeur est fortement négative (excès de mutations rares). Dans le cas d'une population subissant un goulot d'étranglement suffisamment récent et faible pour avoir permis la subsistance de plusieurs lignées et/ou un effet de sélection balancée à ce locus (maintien de plusieurs lignées alléliques), cette valeur est positive (excès de mutations en fréquence intermédiaire).

Le test de Fu & Li (1993) est analogue au test de Tajima, mais au lieu de regarder si le nombre de différences par paires est compatible avec le nombre de sites polymorphes, il se base sur la relation attendue entre θ_W et le nombre d'allèles de l'échantillon (*k*).

Bien que les calculs diffèrent légèrement, conférant à ces tests des propriétés de robustesse et de puissance légèrement différentes d'un cas à l'autre, ils sont basés sur un principe similaire. Ces tests permettent de détecter deux types d'évènements présentant une signature similaire: la sélection positive ou un événement démographique de réduction de taille de populations.

La diversité nucléotidique π a également été calculée le long du gène (région exonique et intronique) en estimant celle-ci dans une fenêtre mobile (50pb, pas de 10pb) à l'aide du logiciel DNAsp.

4.6. Arbre phylogénétique et réseau d'allèles pour le gène de la préprocapitellacine.

Les relations phylogénétiques entre allèles de la préprocapitellacine ont été retracées à l'aide du logiciel MEGA 6.0 en utilisant la méthode du Maximum likelihood (modèle Jukes-Cantor, 1000 bootstrap en « complete deletion ») et ce, à partir du jeu de données brut (ensemble des séquences obtenues par clonage). Ces arbres ont été obtenus pour les deux régions du gène afin de tester la congruence des résultats. Chaque individu a été assigné à son type mitochondrial à l'aide d'une pastille de couleur.

Un réseau d'allèles a également été effectué avec le logiciel PoPart (http://popart.otago.ac.nz) en utilisant les données issues du clonage individuel avec un algorithme d'association du type « Minimum Spanning » en assignant les individus soit à leur lignée mitochondriale, soit à leur provenance géographique.

4.7. Tableau des génotypes de la préprocapitellacine – Assignation aux différents clades

Un Tableau récapitulatif du nombre d'allèles de la préprocapitellacine par individu est effectué en résumant dans un premier temps le nombre d'allèles retrouvés en région 5' à partir des données de clonage pour chaque individu en s'appuyant sur les données de génotypage. Dans un deuxième temps, la même analyse est effectuée pour la région 3' sur les données de clonage (assignation à un clade par comparaison de la séquence de l'allèle à l'alignement global) sans recherche d'allèles diagnostiques dans le génotypage puisque celui-ci n'a été fait que pour la région 5'. Finalement, la comparaison des génotypes obtenus dans les 2 régions est effectuée pour définir un génotype consensus pour chaque individu en fonction de sa localité et son type mitochondrial.

4.8. Tests visant à détecter l'action de la sélection positive

Pour tester plus spécifiquement l'hypothèse d'un effet sélectif positif et observer sa localisation le long du gène de la préprocapitellacine, nous avons estimé les ratios de la divergence non synonyme sur la divergence synonyme entre les principaux clades de la

capitellacine et les ratios de polymorphisme non-synonyme sur le polymorphisme synonyme au sein de ces mêmes groupes d'allèles, en partant du postulat que ces groupes représentent les lignées ancestrales des différentes espèces avant la remise en contact des populations. Cette analyse permet donc de visualiser l'évolution des pressions de sélection le long du gène au sein et entre les différents groupes (espèces présumées) en utilisant une fenêtre mobile (50 bp, pas de 10) à l'aide du logiciel DNAsp.

Nous avons ensuite testé les différences observées à l'aide d'un test de McDonald-Kreitman en faisant attention à ne prendre que deux allèles par individu (choix des allèles les plus divergents) (McDonald et al., 1991). Ce test est le seul test qui permette de détecter l'action de la sélection positive sur un gène donné en réalisant une comparaison entre les ratios du nombre de mutations synonymes (supposées neutres: d_S) et du nombre de mutations non-synonymes (potentiellement soumises à la sélection: d_N) qui se sont accumulées à la fois dans la divergence entre espèces et dans le polymorphisme des deux espèces (π_N et π_S). Sous l'hypothèse neutre, ces ratios sont identiques. L'égalité des ratios π_N/π_S et d_N/d_S est testée à l'aide d'un Tableau de contingence sur les mutations observées à l'aide d'un test de Fisher. Lorsque le test est significatif, il peut être dû à $\pi_N/\pi_S > d_N/d_S$ et donc à un effet de la sélection sur le polymorphisme (accumulation de mutations délétères dans le polymorphisme d'une des deux espèces) ou indiquer une accumulation de mutations positives dans la divergence des espèces après la spéciation ($d_N/d_S > \pi_N/\pi_S$). Ce test repose donc sur la comparaison intra et inter-spécifique de séquences en comparant deux classes de mutations : les mutations non synonymes et les mutations synonymes. Alors que les mutations non synonymes sont considérées comme soumises à la sélection, les synonymes sont supposées neutres. Si le locus (il s'agit d'un test mono-locus) est soumis à sélection positive alors $\pi_N/\pi_S << d_N/d_S$: les mutations non synonymes se sont accumulées principalement entre espèces. De l'autre côté, si $\pi_N/\pi_S > d_N/d_S$ alors les mutations non synonymes ségrégent dans les populations mais ne se fixent pas. Il peut s'agir alors d'un maintien de lignées anciennes dans le polymorphisme sous sélection balancée si le polymorphisme non-synonyme apparaît en fréquence intermédiaire.

Il reste cependant important de garder à l'esprit que ce test est à prendre avec précaution dans le cas où le jeu de données utilisé contient des duplications et/ou de l'hybridation entre espèces, ce qui peut être documenté lorsque des espèces cryptiques sont étudiées en

parallèle. En effet, la recombinaison pourra propager des mutations non synonymes fixées dans la divergence des clades dans le polymorphisme de ceux-ci. En conséquence, ceci pourrait permettre de réfuter l'hypothèse de sélection positive agissant sur la divergence des clades.

4.9. Structure spatiale des populations de Capitella spp. : Fréquences haplotypiques (Cox-1) et alléliques (préprocapitellacine)

Les fréquences haplotypiques ont été calculées à l'aide du logiciel Fstat (Goudet, 1995) et reportées sur une carte géographique à l'aide de camemberts permettant de visualiser l'information sur l'aire de distribution échantillonnée sous R avec un script « home-made ». Les fréquences alléliques de la préprocapitellacine (calculées à l'aide du logiciel Fstat également) au sein des différentes localités sont également calculées à partir du jeu de données brut issu du clonage individuel encore une fois en ne gardant que les deux allèles les plus divergents pour chaque individu lorsque cela est possible et en regroupant les individus de Boulogne et Dunkerque. Différentes couleurs ont été choisies pour renseigner de la répartition des différents clades. Ce choix fait l'hypothèse que les clades d'allèles trouvés correspondent à une signature d'espèce et non à des duplications du gène qui seraient antérieures aux événements de spéciation.

4.10. Calcul des indices de différenciation génétique

4.10.1 Structure génétique avec le marqueur *Cox-1*

Au sein de chaque lignée mitochondriale, les calculs de la statistique –ici uniquement- Fst (sur données haplotypiques) (Wright, 1951) par paire de localités ont été réalisés à l'aide du logiciel Arlequin (Excoffier et al., 2005). L'écart des valeurs de Fst à zéro a été testé au seuil de 5% en effectuant 1000 permutations des génotypes/haplotypes entre populations pour chaque comparaison avec le même logiciel.

4.10.2. Structure génétique de la capitellacine

Une analyse de la différenciation génétique a été menée entre les populations Boulogne/Roscoff/Port-la-Forêt. A partir des données de séquences, le logiciel Arlequin a été utilisé pour quantifier le degré de différenciation génétique par paires de populations, et ont

été testés contre l'hypothèse de non-différenciation par un procédé de permutations des allèles entre populations (fixé à 10 000) permettant d'obtenir une valeur de significativité pour chaque valeur d'indice de fixation. Ce calcul a été réalisé/basé à la fois sur les fréquences des haplotypes dans les populations (Fst ; Wright 1951) et sur les fréquences des sites variants dans l'alignement (le Phi_{ST} : Excoffier et al., 1992). Ceci a été effectué globalement pour chacun des clades, puis par paires de localités au sein des clades.

4.11. Analyse IMa2

L'utilisation de marqueurs génétiques, que ce soit à partir de séquences de gènes, de données SNP ou encore des microsatellites, et le développement de méthodes analytiques (d'autant plus robustes qu'il y a de locus à analyser) ont permis l'amélioration de l'étude des flux de gènes entre populations après ou pendant un épisode d'isolement (Pinho and Hey, 2010). Ces logiciels utilisent des modèles de migration avec isolement utilisant un certain nombre de paramètres de population (tailles efficaces, taux de migration et temps depuis la séparation des populations), essentiels pour la compréhension des processus de divergence avec flux de gènes. Des programmes basés sur la théorie de la coalescence se sont ainsi développés, et permettent d'estimer ces paramètres démographiques (Nielsen and Wakeley, 2001). Le logiciel IMa2 utilise des données séquences à plusieurs locus pour inférer les paramètres démographiques associés à n populations en phase d'isolement, et notamment les flux migratoires après séparation des populations selon un arbre de référence retraçant l'histoire de la séparation des populations. Dans notre cas, le logiciel IMa2 a été utilisé pour estimer le degré d'introgression des allèles de la préprocapitellacine entre les trois clades mitochondriaux Cc-Manche1 (pop 0), Cc-Atlantique (pop 1) et Cc-Manche2 (pop 2) et leurs populations ancestrales (pop 3 et 4) en utilisant les données des locus COI et capitellacine en estimant conjointement la taille efficace de populations, les taux de migration orientée entre populations et le temps de divergence).

Un premier scénario (Figure 5) repose sur l'hypothèse que chaque clade nucléaire est corrélé au fond génétique des espèces et, que des échanges d'allèles de la capitellacine aient pu avoir lieu entre les génomes de ces lignées mitochondriales.

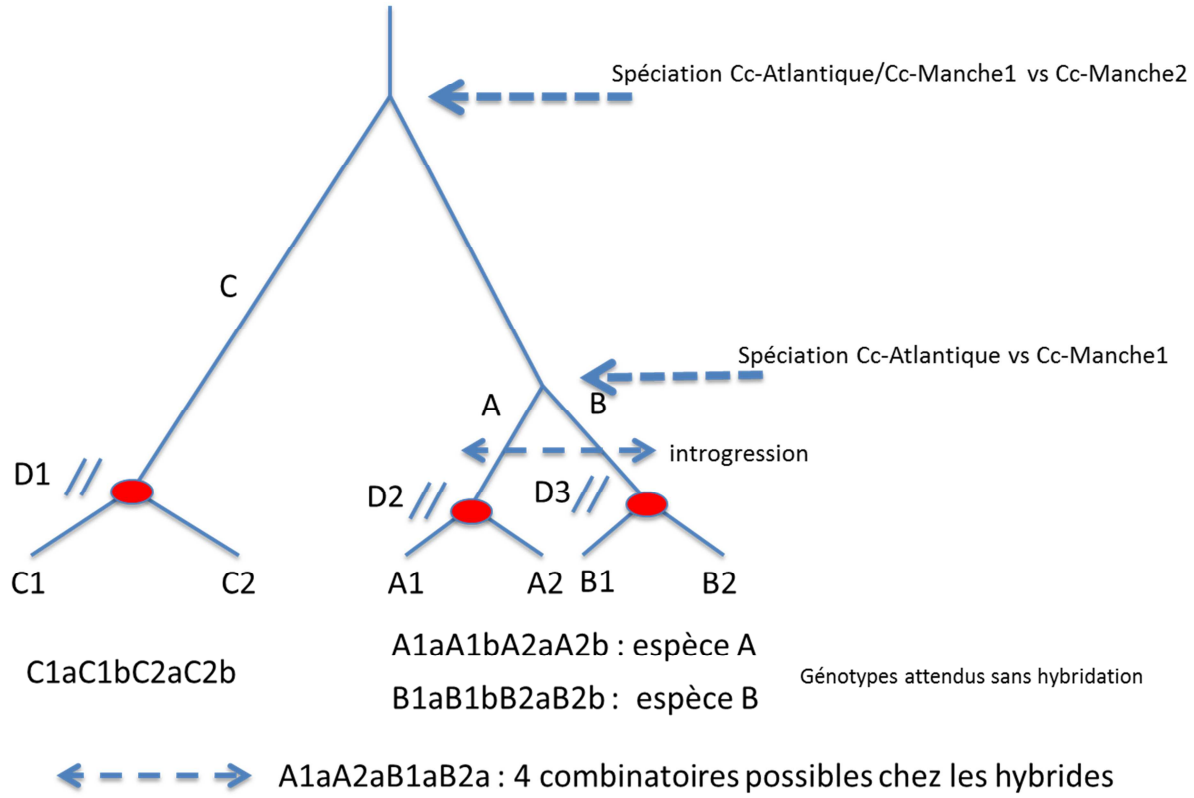


Figure 5. Schéma récapitulatif du scénario évolutif proposé et testé dans lequel les clades nucléaires correspondent au fond génétique des lignées mitochondriales.

Un deuxième scénario (Figure 6) a été testé dans lequel les clades A et B représenteraient un évènement de duplication survenu chez l'espèce ancestrale aux deux lignées mitochondriales actuelles Cc-Manche1 et Cc-Atlantique.

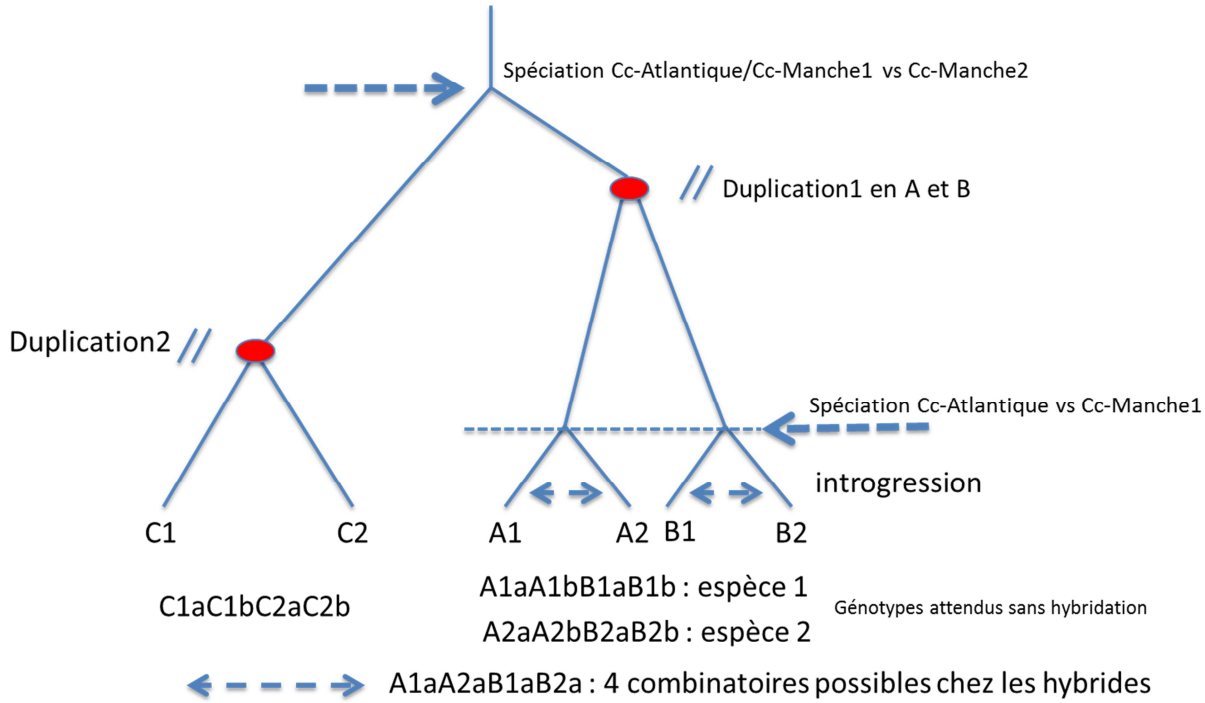


Figure 6. Scénario 2 testé ou A et B sont issus d'un évènement de duplication ancestrale à la spéciation des espèces Cc-Atlantique/ Cc-Manche1 (survenu après la spéciation de l'espèce ancestrale avec l'espèce actuelle Cc-Manche2.

Les alignements de la capitellacine utilisés constituent des blocs sans indel en supprimant les parties introniques trop variables autour des zones contenant ces derniers. Les individus sont ensuite regroupés par lignée mitochondriale en gardant une correspondance parfaite entre les noms des individus sur les 2 alignements *Cox-1* et capitellacine, et en prenant systématiquement les 2 allèles les plus divergents de la capitellacine pour chaque individu génotypé dans le cas du scénario 1. Les alignements ont pour le scénario 2 été regroupé par lignée mitochondriale en attribuant à chaque individu deux allèles pour chaque clade. Dans le cas où l'individu a été déterminé comme étant homozygote au sein d'un clade, l'allèle caractérisé par le clonage individuel a été doublé. Comme le logiciel IMa2 exclue toute recombinaison entre allèles au sein d'un locus (effet de réticulation dans le coalescent), le package IMgc (Woerner et al., 2007) a été utilisé sur les 2 alignements pour ne garder que la portion du gène sans recombinaison entre les différents allèles.

Paramètres estimés

Le paramètre Θ (ici q = 4Ne.μ où Ne représente la taille efficace de la population) a été estimé pour les différents clades mitochondriaux (espèces) et populations des espèces ancestrales. Le paramètre 2N.m est également estimé pour chaque paire de populations (pour lequel N représente la taille de la population qui émet les migrants et m la proportion de la population qui immigre à chaque génération). Finalement, le paramètre t (temps depuis la séparation des populations) est également estimé et peut être défini par l'équation t = T/μ ou T représente le nombre d'années depuis la divergence des clades. Le temps de divergence fait référence au temps où une structure panmictique globale a été remplacée par des populations génétiquement structurées avec peu ou pas de connectivité.

Le programme IMa2 utilise une méthode de Monte Carlo par chaînes de Markov pour déterminer la distribution *a posteriori* des paramètres cités précédemment. La chaîne de Markov a été lancée sur 10 millions d'itérations, et les valeurs des paramètres enregistrées avec un pas d'échantillonnage de 100 itérations (100 000 données) après un période de lancement (burn-in) de 10 000 itérations. A chaque itération, les valeurs des paramètres sont tirées au hasard dans une distribution uniforme (priors) comprise de zéro à 40 pour θ, de zéro à 5 pour m et de zéro à 10 pour t afin de reconstruire les probabilités postérieures des paramètres.

Résultats

1. Analyse phylogéographique du complexe *Capitella spp.* avec *Cox-1* – Assignation des espèces françaises par rapport au complexe mondial d'espèces cryptiques.

L'alignement final du *Cox-1* sur une longueur de 522 nucléotides (174 acides aminés) a été effectué sur 95 individus issus du complexe d'espèces *Capitella spp* au niveau mondial sur la base de 10 localités provenant d'Europe, Japon et Amérique du sud. Sur les 523 sites, 189 (36%) sont variables avec un nombre total de 283 mutations réparties sur 48 haplotypes. L'arbre ML réalisé à partir de l'alignement complet permet de discriminer huit clades principaux (Figure 7). Les mêmes clades peuvent être mis en évidence en utilisant l'arbre réalisé à partir des deux premiers nucléotides de chaque codon (données non montrées). Ces clades sont soutenus par des valeurs de bootstrap élevées (comprises entre 90-100%). Trois groupes géographiques principaux sont mis en évidence : le premier regroupe les clades 1 à 3 qui sont les espèces retrouvées sur le littoral français ; le second regroupe les clades 4, 5 et 6 composés des individus de l'espèce *C. teleta-like* français, américain et japonais, exception faite du clade 4 pour lequel certains individus ont été identifiés comme appartenant à l'espèce *C. capitata* par (Carr et al., 2011). Le dernier groupe comprend des individus apparentés à l'espèce *C. capitata* provenant de Méditerranée, du Canada et de l'Inde. Cependant ces regroupements ne sont pas supportés statistiquement avec des nœuds internes présentant des valeurs de bootstrap comprises entre 25 et 48%, exception faite du nœud (88%) regroupant les clades français Cc-Atlantique, Cc-Manche1 et Cc-Manche2 et, laissant supposer que celles-ci ont effectivement une histoire commune.

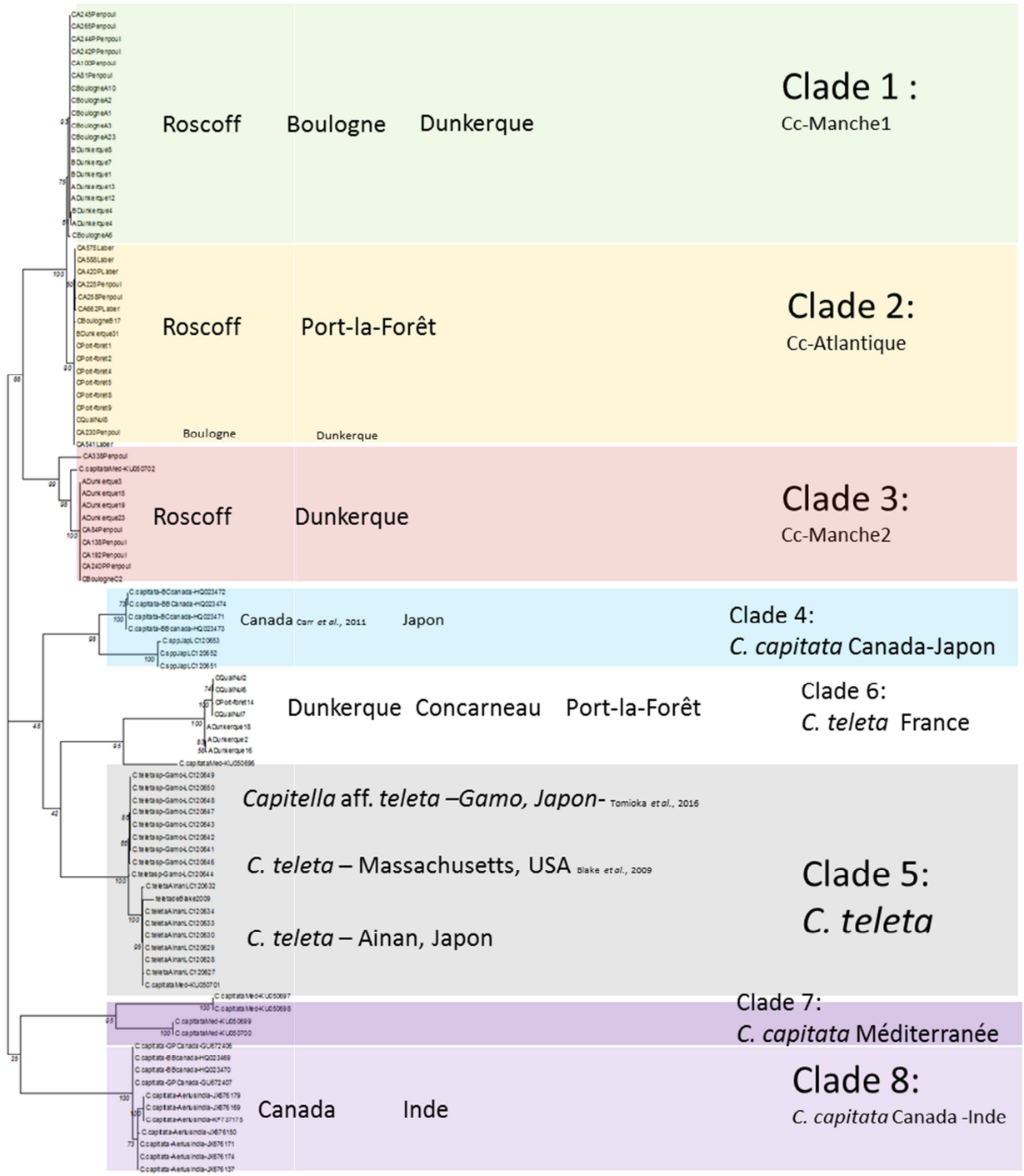


Figure 7. Arbre ML des individus de *Capitella spp.* réalisé avec le modèle de substitution HKY G+I: (meilleur modèle retenu par BIC avec jModelTest) à partir du gène *Cox-1* avec les séquences de notre jeu de données contenant des espèces de *Capitella spp.* de Manche orientale (Dunkerque-Boulogne), Manche occidentale (Roscoff) et Atlantique (Port-la-Forêt–Concarneau) et celles des espèces apparentées aux espèces *C. capitata* et *C. teleta* provenant de plusieurs régions du monde (Japon, Canada, Inde, Etats Unis, Méditerranée).

Dans l'espèce « *capitata-like »*, 3 lignées divergentes ont pu être échantillonnées en Manche et en Atlantique. Ces lignées sont très différentes des lignées échantillonnées en Inde, Canada et Méditerranée. Quatre clades se répartissent dans l'espèce *teleta :* deux lignées divergentes au Japon, une aux Etats Unis, et la dernière en Europe *(*la lignée Française regroupant des individus de Concarneau, Dunkerque et de la Méditerranée). Enfin, une lignée intermédiaire référencée comme *capitata* mais plutôt affiliée génétiquement à *teleta* est présente au Japon et au Canada uniquement.

Au sein de la Manche et de l'Atlantique Nord, la relation clade-géographie est évidente : le clade « Cc-Atlantique » possède des individus apparentés à l'espèce *C. capitata-like* à Port-la-Forêt majoritairement avec quelques individus en Manche et le clade « Cc-Manche1 » contient des individus de Dunkerque-Boulogne à Roscoff (Manche occidentale). Le clade nommé « Cc-Manche2 » n'est retrouvé qu'à Dunkerque et Roscoff et semble absent des sites Boulogne et Concarneau même si l'effort d'échantillonnage reste insuffisant pour conclure quant à leur absence réelle. La lignée Méditerranée référencée comme appartenant à l'espèce *C. capitata* est bien distincte des lignées Manche-Atlantique.

Au niveau mondial, l'association clade-géographie est moins évidente : des individus proches phylogénétiquement sont retrouvés à la fois au Canada mais aussi au Japon ou en Inde (Clade 4 et Clade 8) bien que des divergences soient observées entre individus selon les pays échantillonnés au sein de ces clades. L'analyse montre que les 2 espèces morphologiques *teleta* et *capitella* présentent toutes les 2 une répartition cosmopolite juxtaposée mais avec des espèces géographiques cryptiques.

Pour l'espèce *C. teleta*, une quatrième lignée proche mais suffisamment divergente de l'espèce *C. teleta* américaine et japonaise est présente à Dunkerque et Concarneau et dans les données RNAseq d'individus prélevés à Roscoff.

2. Divergence nette inter clades

Les divergences nettes entre les différents clades sont présentées dans le Tableau 3. Elles montrent que les lignées géographiques Cc-Manche1–Cc-Atlantique sont faiblement divergentes (2%) comparées à celles trouvées entre les autres lignées mondiales. La lignée Cc-Manche2 présente quant-à-elle une divergence beaucoup plus élevée (14%) avec les 2 lignées précédentes avec lesquelles elle partage la même zone géographique. Cette divergence est cependant nettement inférieure à celles trouvées entre les lignées des autres zones géographiques (plus de 21%). La moyenne des divergences intra *C. capitata* (populations échantillonnées au Canada, Inde, Méditerranée et Japon) est de 33% et la divergence moyenne entre les populations de *C. capitata* au niveau mondial et les lignées Cc-Manche1-Cc-Atlantique est de 26%. Les divergences moyennées entre les clades *teleta* et *capitata* sont du même ordre de grandeur (24% : *C. teleta* vs Lignées *capitata* FR et 36% : *C. teleta* FR et *C. capitata*) qu'entre celles estimées au sein du complexe *capitata* suggérant une longue histoire évolutive de spéciation et que ces espèces ne constituent pas des espèces cosmopolites *per se*.

	Cc-Manche	Cc-Atlantique	Cc-Manche2	*C. capitata* Canada/Inde	*C. capitata* Méditerranée	*C. capitata* Canada/Japon	*C. teleta* FR	*C. teleta*
Cc-Manche	0,001							
Cc-Atlantique	0,021	0,002						
Cc-Manche2	0,142	0,141	0,000					
C. capitata Canada/Inde	0,257	0,262	0,227	0,011				
C. capitata Méditerranée	0,286	0,291	0,311	0,349	0,124			
C. capitata Canada/Japon	0,245	0,238	0,229	0,299	0,338	0,067		
C. teleta FR	0,308	0,326	0,259	0,403	0,395	0,292	0,011	
*C. teleta**	0,238	0,236	0,259	0,288	0,337	0,225	**0,265**	0,018

Tableau 3. Divergences nettes entre les différents groupes de séquences (les clades/lignées) pour le gène mitochondrial *Cox-1* calculé à l'aide du logiciel MEGA6.

3. Description de la diversité génétique mitochondriale dans les populations Françaises

3.1. Fréquences haplotypiques du Cox-1 dans les populations françaises

Un total de 23 haplotypes a été identifié au sein des 145 individus assignés à *C. aff. capitata* dans les 5 populations françaises (Boulogne, Dunkerque, Roscoff, Port-la-Forêt, Concarneau). Les fréquences de ces haplotypes sont présentées sous la forme de camemberts aux 5 localités (Figure 8C). Le réseau réalisé sous PoPart (Figure 8A) montre que trois lignées mitochondriales co-existent (Clade 1, Clade 2, Clade 3) dans le jeu de données avec des distributions géographiques plus ou moins disjointes. D'après le réseau, le clade 1 (Cc-Manche1) occupe une position centrale et est le plus diversifié, ce qui laisse penser que le clade 2 (Cc-Atlantique) pourrait être récemment dérivé lors des derniers épisodes glaciaires.

Le clade 1 (vert) forme une structure en étoile caractéristique composé d'un haplotype majoritaire (Hap6) autour duquel gravitent 9 singletons et quelques haplotypes plus divergents (Hap3, 4, 9, 15) (Figure 8A). Cet haplotype majoritaire se retrouve en fréquence élevée (73%, 63% et 54%) dans toutes les populations de Manche (Dunkerque, Boulogne et Roscoff respectivement). Deux haplotypes secondaires (les haplotypes 1 et 7) se retrouvent partagés également entre Dunkerque et Boulogne à des fréquences de 10% et 25% respectivement.

Au sein du clade 2 (jaune), un haplotype principal peut être observé (Hap16 : 93%) dont dérivent plusieurs haplotypes. Les haplotypes 17 (blanc), 18 (rosé) et 21 (blanc) dérivent de cet haplotype principal d'une seule et unique mutation. L'haplotype 18 n'est retrouvé qu'à Roscoff et possède quant-à-lui plusieurs haplotypes dérivés qui sont eux-aussi retrouvés uniquement au sein de la population de Roscoff. Le clade 3 (rouge) possède un seul et unique haplotype qui se retrouve en fréquence intermédiaire au sein des populations Dunkerque-Boulogne-Roscoff (13, 25, 10%). Aucun haplotype n'est partagé entre toutes les localités (l'haplotype majoritaire de la lignée Atlantique ne se retrouvant pas à Boulogne). Le long de la Manche, seuls les 2 haplotypes majoritaires de la lignée Manche1 (vert) et de l'autre lignée Manche2 (rouge) sont présents dans les trois populations. Les populations de Roscoff et d'Atlantique ne partage qu'un seul haplotype avec un fort différentiel de fréquences.

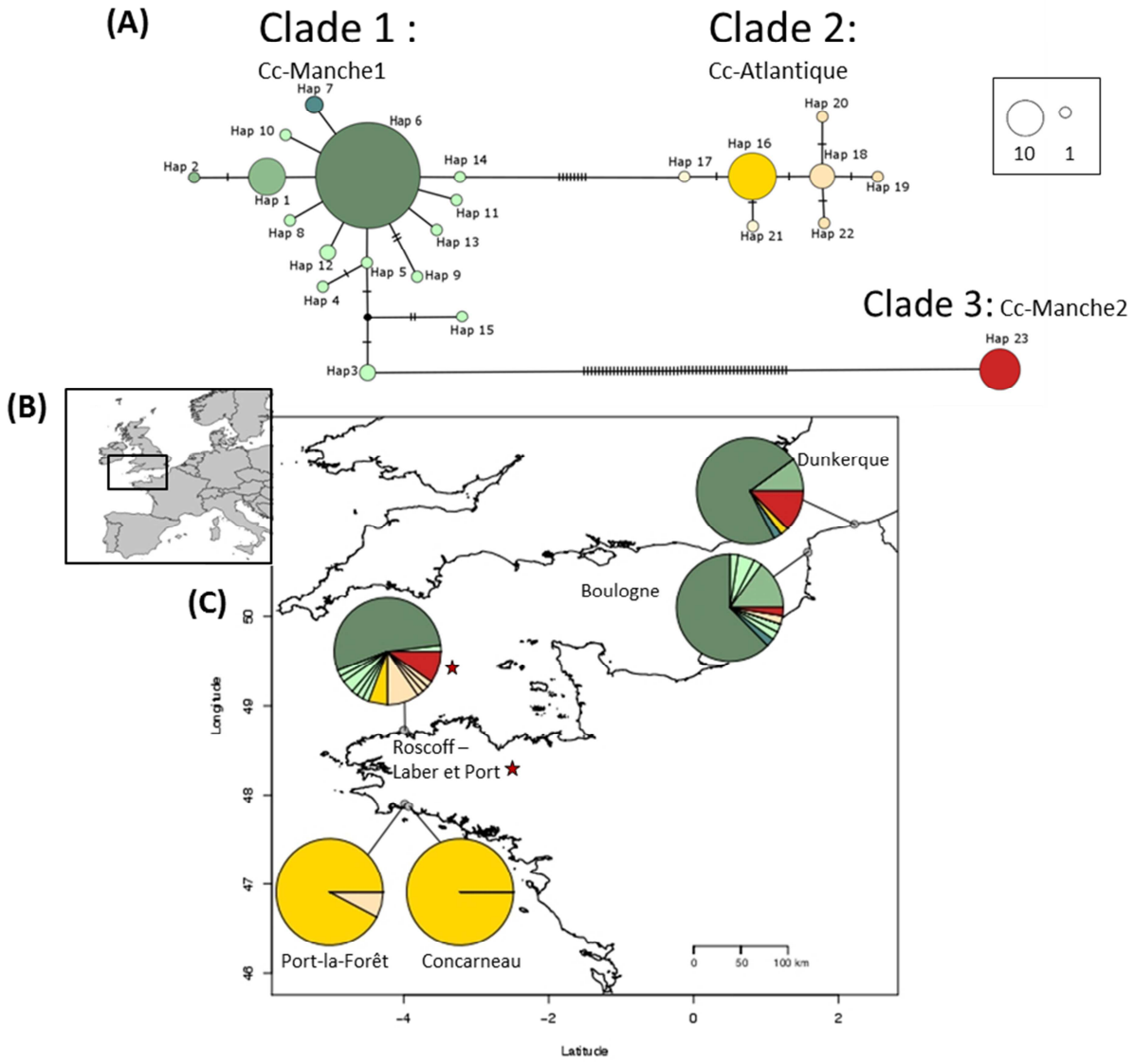


Figure 8. Réseau et distribution des haplotypes mitochondriaux de *Capitella spp.* le long de la Manche (Dunkerque-Boulogne-Roscoff) et de l'Atlantique nord (Concarneau – Port-la-Forêt). **A)** Réseau d'haplotypes en Neighbor joining. La taille du cercle est proportionnelle au nombre d'haplotypes. Les liens et les barres noires représentent le nombre de mutations entre deux haplotypes. Les cercles noirs représentent un haplotype manquant. Les phylogroupes inferrés sont matérialisés sous forme de couleur (jaune/rouge/vert) et les couleurs pales représentent les haplotypes dérivés des trois principaux clades **B)** Zone d'étude (entourée en noir) en relation avec l'Europe de l'ouest **C)** Camemberts des fréquences haplotypiques par localité. La présence de l'étoile rouge indique que ces haplotypes se retrouvent seulement au sein du Port de Roscoff et aucun dans le laber.

Ces résultats suggèrent une bonne connectivité des populations de la Manche orientale et occidentale et l'existence d'une barrière qui isole les populations Atlantique de celles de la Manche. Un seul camembert a été représenté à Roscoff (Laber et Port) puisque les fréquences des haplotypes sont les mêmes au sein des deux localités pour les lignées Cc-Manche1 et Cc-Atlantique. Pour l'espèce Cc-Manche2 cependant, les individus n'ont été retrouvés qu'au sein du port de Roscoff (étoile rouge) et suggère que cette espèce pourrait être plus spécifiquement inféodés aux sédiments anoxiques des ports. Des individus parasités sont retrouvés dans tous les clades (25% dans Cc-Atlantique, 43% dans Cc-Manche1 et 20% dans Cc-Manche2).

Pour l'espèce assignée à *C. teleta*, plus rarement échantillonnée dans nos prélèvements, elle n'est retrouvée qu'au sein des localités situées aux extrémités de notre continuum, à savoir Port-la-Forêt et Dunkerque avec une fréquence peu élevée (inférieure ou égale à 10% de la population) mais représente l'espèce majoritaire à Quai Nul (Concarneau) ou quelques individus ont pu être échantillonnés ensemble dans un habitat sédimentaire plus grossier.

3.2. Barcode-Gap sur le complexe C. capitata-like

Les distances intra-clade et inter-clades sont présentées dans la Figure 9. Les distances génétiques calculées intra-clades (0 à 0,001) sont nettement inférieures aux distances génétiques inter-clades (0,021 à 0,143) et conforment aux données de polymorphisme intra-espèce. Pour rappel, c'est la comparaison inter-clade Cc-Atlantique/Cc-Manche1 qui montre la plus petite divergence inter-clade alors que les comparaisons avec Cc-Manche2 sont les plus élevées malgré l'absence de divergence intra-clade. L'identification de « gaps » dans les distances nucléotidiques au sein du jeu de données contenant toutes les espèces des côtes françaises permet de mettre en évidence que les séquences peuvent être regroupées en 3 unités taxonomiques bien séparées : un groupe contenant les comparaisons de séquences des clades Cc-Manche1 et Cc-Atlantique (47 et 53% d'individus respectivement), un autre contenant toutes les comparaisons avec le clade Cc-Manche2 et le dernier contenant les comparaisons de séquences avec les individus du clade *C. teleta FR*. La présence du premier gap se fait après 0,02 de divergence et sépare le clade Cc-Manche2 des deux phylogroupes regroupés Cc-Manche1 et Cc Atlantique pour lesquelles la divergence inter-clade n'est que légèrement supérieure à la divergence intra-clade et ne permet pas de les assigner à deux

espèces différentes. Le deuxième gap se fait après 0,13 et permet de discriminer l'espèce *C teleta* FR.

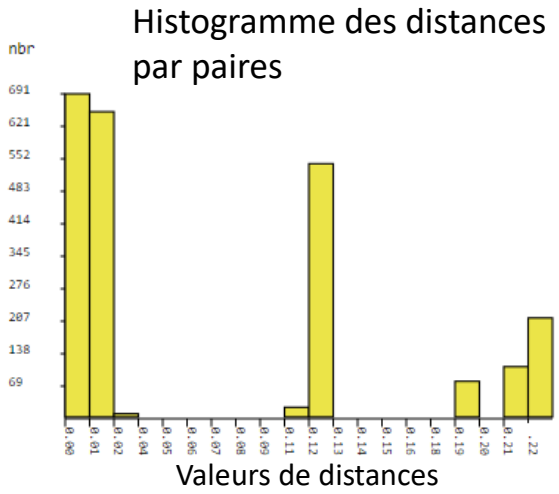


Figure 9. Histogramme des distances K2P calculées à partir du logiciel ABGD sur les espèces de localités françaises.

3.3. Diversité génétique dans le complexe C. capitata français (Cox-1)

La diversité nucléotidique du gène mitochondrial pour les 'espèces' françaises varie beaucoup entre les 3 clades mitochondriaux examinés avec des valeurs de π comprise entre 0 et 0,0016 et des diversités haplotypiques variant de 0 à 0,6 (Tableau 4). Il est important de noter que, même si l'effectif échantillonné est restreint, il n'existe aucune variation génétique dans le clade Cc-Manche2, laissant supposer l'existence d'un goulot d'étranglement et/ou un balayage sélectif mitochondrial.

Clade	N	S	Hd	π	θ_w	Fu et Li D	Fu et Li F	Tajima D
Cc-Manche1	108	17	0,417 ± 0,059	0,00123 ± 0,00024	0,00620 ± 0,00207	-5,13**	-4,89**	-2,44**
Cc-Atlantique	26	6	0,6 ± 0,0098	0,00158 ± 0,00034	0,00301 ± 0,0015	-2,24*	-2,53*	-1,21(NS)
Cc-Manche2	11	0	N.A.	N.A.	N.A.	N.A.	N.A.	N.A

Tableau 4. Descripteurs de la diversité génétique pour les différents clades et valeurs des tests statistiques visant à détecter un écart à la neutralité pour le gène mitochondrial *Cox-1* dans les 3 clades (espèces). N : nombre de séquences, S : nombre de sites variant, ***p<0.01 **p<0.02 * p<0.05, NS: non significatif, N.A. non applicable.

Pour l'espèce Cc-Manche1, la diversité génétique π est plus faible à Dunkerque (π=0,005 Hd=0,26) qu'à Boulogne (π= 0,0018/Hd=0,56) ou à Roscoff- Laber et Port : π= 0,0013/Hd=0,41). Pour la lignée Cc-Atlantique, la diversité génétique est beaucoup plus élevée à Roscoff (π=0,0018, Hd=0,76 équivalente entre Laber et Port) qu'à Port-la-Forêt (π=0,0003, Hd=0,15). Les diversités nucléotidiques des autres espèces recensées comme appartenant à *C. capitata* sont, par comparaison et logiquement (intégration de la différenciation spatiale), presque 10 fois plus élevées avec des valeurs comprises entre 0,07 à 0,11 mais englobent des individus de régions nettement plus éloignées (échelle spatiale d'échantillonnage plus grande).

Les valeurs des tests de Tajima et de Fu & Li sont toutes négatives pour les populations françaises et présentent un écart significatif à l'attendu neutre pour la lignée Cc-Manche1 qui est beaucoup moins marqué pour la lignée Cc-Atlantique. Sous DNAsp v5, une simulation de coalescents neutres à partir du jeu de données pour chaque type mitochondrial a été réalisée permettant de mettre en évidence que la lignée Cc-Manche1 s'écarte significativement d'un coalescent neutre, ce qui n'est pas le cas de la lignée Atlantique chez qui les tests ne sont pas significatifs.

Les divergences synonyme et non-synonyme entre lignées mitochondriales ont également été calculées (Ka et Ks par paire de lignées : Tableau 5). Seules quelques mutations synonymes sont fixées entre la lignée Manche et la lignée Atlantique (8 substitutions, Ks=0,08). Pour la lignée Cc-Manche2 également, seules des mutations synonymes sont fixées (50 et 56 entre CC-Manche1 et CC-Atlantique et Cc-Manche1 et Cc-Manche2 respectivement (Ks=0,63). Les comparaisons entre l'espèce française *C. teleta* et les autres lignées (Cc-Manche1; Cc-Atlantique et Cc-Manche2) montrent des valeurs de Ka logiquement plus élevées (0,02 en moyenne) pour des valeurs de Ks très élevées (1,3 minimum), laissant supposer que le synonyme est saturé et donc que cette divergence est le fruit d'une histoire évolutive très ancienne sur un gène très contraint par la sélection purifiante (résultats non montrés).

	Cc-Manche		Cc-Atlantique		Cc-manche2		*C. teleta* FR
Cc-Manche	*ka*	*Ks*					
	Ka/Ks						
Cc-Atlantique	*0,000*	*0,080*					
	0,000						
Cc-manche2	*0,000*	*0,624*	*0,000*	*0,630*			
	0,000		**0,000**				
C. teleta FR	*0,025*	*1,822*	*0,024*	*2,290*	*0,027*	*1,270*	
	0,013		**0,011**		**0,021**		

Tableau 5. Divergences non synonyme (Ka) et synonyme (Ks) et leur rapport Ka/Ks entre les différentes lignées françaises pour le gène mitochondrial COI calculé avec le logiciel DNAsp v5.

3.4. Différenciation génétique au sein des différentes lignées de C. capitata-like

Un test de différenciation génétique entre les différentes populations géographiques de chaque clade/lignée a été réalisé à l'aide du logiciel Arlequin en calculant un F_{ST} par paire de populations sur les données de fréquences haplotypiques (Tableau 6). Ce test permet de mettre en évidence qu'il n'existe pas de structure géographique au sein de la lignée Manche Cc-Manche1 (F_{ST} global 0,028 ; p-value=0,06) avec des valeurs de F_{ST} par paire de populations non- significatives voire marginalement significative dans le cas de Boulogne-Roscoff.

	Dunkerque	Boulogne	Roscoff
Dunkerque			
Boulogne	0,011 (0,21)		
Roscoff	0,030 (0,11)	0.036 (0,051)	

Tableau 6. Calcul de l'indice de fixation F_{ST} par paire de populations. P-values fournies entre parenthèses

Au sein de la lignée Atlantique, la seule valeur de F_{ST} calculable est celle entre Port-la-Forêt et Roscoff. Celle-ci correspond à un test exact de différenciation significativement différent de zéro (0,41 p-value <1%) soulignant l'isolement géographique de la population de Roscoff vis-à-vis des populations de Bretagne Sud (1 seul site polymorphe avec une mutation fixée entre les deux populations).

3.5. Courbes de mésappariement et histoire démographique des lignées mitochondriales de Capitella

Les courbes de mésappariement ont été réalisées au sein de chaque lignée mitochondriale et sont représentées sur la Figure 10. Pour la lignée Cc-Manche1, la distribution du nombre de mutations entre paires de séquences s'ajuste à la fois à la courbe attendue sous le modèle d'une population démographiquement stable mais aussi à celle attendue pour une population en expansion (R2=0,037, p=0,057). Il semblerait cependant pour cette lignée que ces courbes présentent un excès de mutations rares suggérant une évolution qui pourrait s'écarter du coalescent neutre mais avec un Tau égal à zéro.

De la même façon, la courbe de distribution des mutations observées par paire de séquences s'ajuste sur la distribution théorique d'une population en croissance/déclin pour la lignée Cc-Atlantique avec une plus forte fréquence des mutations rares (R2=0,074, p=0,06) et fournit un coefficient Tau d'accroissement de la population en unité mutationnel de 0,80. L'excès de mutations rares par rapport à l'attendu d'une population démographiquement stable est cependant biaisé par la structure génétique sous-jacente liée à la différenciation Roscoff/Port-la-Forêt (chaque localité possède en effet un fort pourcentage d'allèles privés).

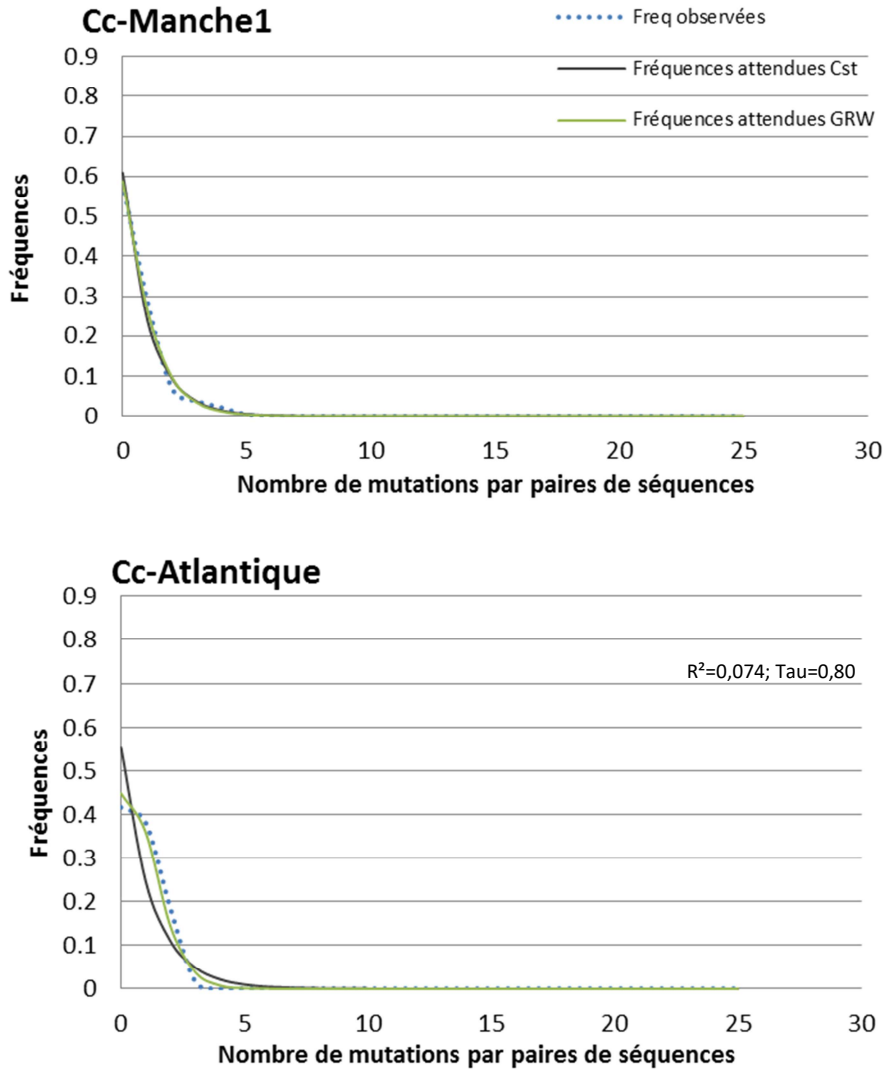
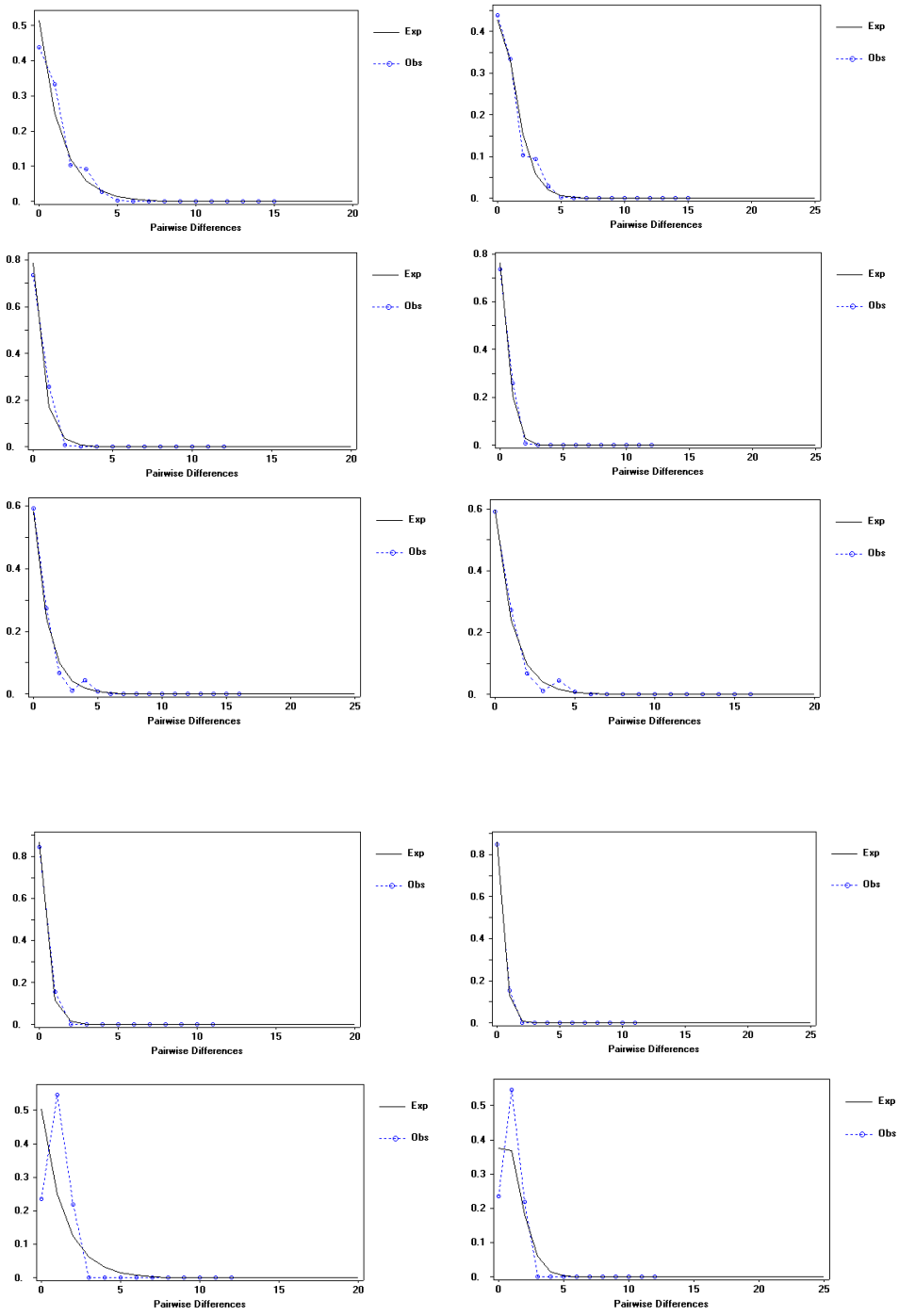


Figure10. Distributions en fréquence du nombre de différences nucléotidiques par paire de séquences par type mitochondrial (courbe de mésappariement) et ajustement des données observées aux modèles théoriques attendus de population démographiquement stable (cst) et de population en expansion/déclin (grw pour growth/decline)

La même analyse a ensuite été effectuée par localité dans l'éventualité d'une différenciation inter-localités (Figure 11). Cette analyse réalisée sous DNAsp v5 a permis de mettre en évidence que les courbes s'écartent d'un coalescent neutre (SCN non significatif) à Dunkerque/Cc-Manche1. Pour cette localité, les courbes s'ajustent au modèle attendu sous une population en déclin/croissance (non démographiquement stable). L'ensemble des données supporte donc plus l'hypothèse d'une population en expansion au niveau de la lignée Cc-Manche1 mais ne permet pas de conclure quant à la lignée Atlantique.



BOULOGNE Cc-Manche1
Constant -> Growth
(R^2=0,054; Tau=0,43
SCN: p-value<0,05*)

DUNKERQUE Cc-Manche1
Constant -> Growth
(R^2=0,01; Tau=0,27
SCN: p-value>0,05)

ROSCOFF Cc-Manche1
Constant -> Growth
(R^2=0,065; Tau=0
SCN: p-value<0,05*)

PORT-LA-FORET Cc-Atlantique
Constant -> Growth
(R^2=0,27; Tau=0,15
SCN: p-value>0,05)

ROSCOFF Cc-Atlantique
Cst -> Growth
(R^2=0,12; Tau=0,98
SCN: p-value<0,01**)

Figure 11. Distributions en fréquence du nombre de différences nucléotidiques par paire de séquences par type mitochondrial au sein des différentes localités (courbe de mésappariement) et ajustement des données observées aux modèles théoriques attendus de population démographiquement stable (cst) et de population en expansion/déclin (grw pour growth/decline).

4. Diversité génétique du précurseur protéique de la Capitellacine

4.1. Description de la structure du gène

Le gène de la préprocapitellacine se compose de 6 exons (672 bp) et 5 introns (715 bp) pour une longueur totale de 1387 bp. L'étude de la diversité moléculaire de ce gène a été réalisée en séparant le gène en deux régions à peu près égales (avec une zone chevauchante commune). La région 5' du gène est composée de 962 nucléotides dont 413 bp dans les régions exoniques et la présence d'un microsatellite dans l'intron 3 en fin de région 5'. La zone chevauchante avec la région 3' quant-à-elle est constituée de près de 450 nucléotides qui englobent des zones diagnostiques (intronique et exonique) des différents groupes de séquences avec leurs propres signatures mutationnelles. La région 3' est composée de 929 nucléotides dont plus de 450 sont extérieurs à la région chevauchante. La région du gène codant pour le PAM est constituée de 69 nucléotides soit 23 acides aminés codés en région C-terminale. La région BRICHOS quant à elle contient 267 nucléotides soit 89 acides aminés codant pour une protéine chaperonne de petite taille (plus petite que celle de la préproalvinellacine).

4.2. Analyse des séquences et filtrage des séquences chimériques artéfactuelles

Pour les 33 individus étudiés, les recaptures de séquence issues du clonage individuel ont été alignées (8 séquences par individu soit 260 séquences). Les séquences rigoureusement identique au sein d'un individu mais capturées plusieurs fois ont permis de valider de nombreux allèles et ont ensuite été enlevées du jeu de données dans le but d'obtenir un nombre minimum de séquences/allèles par individu pour chaque partie du gène. Le nombre total d'allèles recapturés est *in fine* de 76 sur les 33 individus après élimination des mutations artéfactuelles (i.e. singletons propre à un individu). Ce filtrage (séquences identiques/singletons) a permis de retirer un quart des séquences du jeu de données. Les recombinants artéfactuels (entre allèles au sein d'individu) ont également été enlevés du jeu de données quand ils ne présentaient pas de point de recombinaison entre les allèles de 2 individus différents déjà recapturés dans le jeu de données. Puisque chaque individu ne montrait globalement pas plus de 2 ou 3 séquences et que le clonage a été fait individu par individu, cette étape a été réalisée visuellement puisque le jeu de données individuel était moins complexe que celui décrit dans le cas de l'alvinellacine. Au final, deux allèles recombinants naturels ont été retrouvés entre deux clades chez des individus différents au

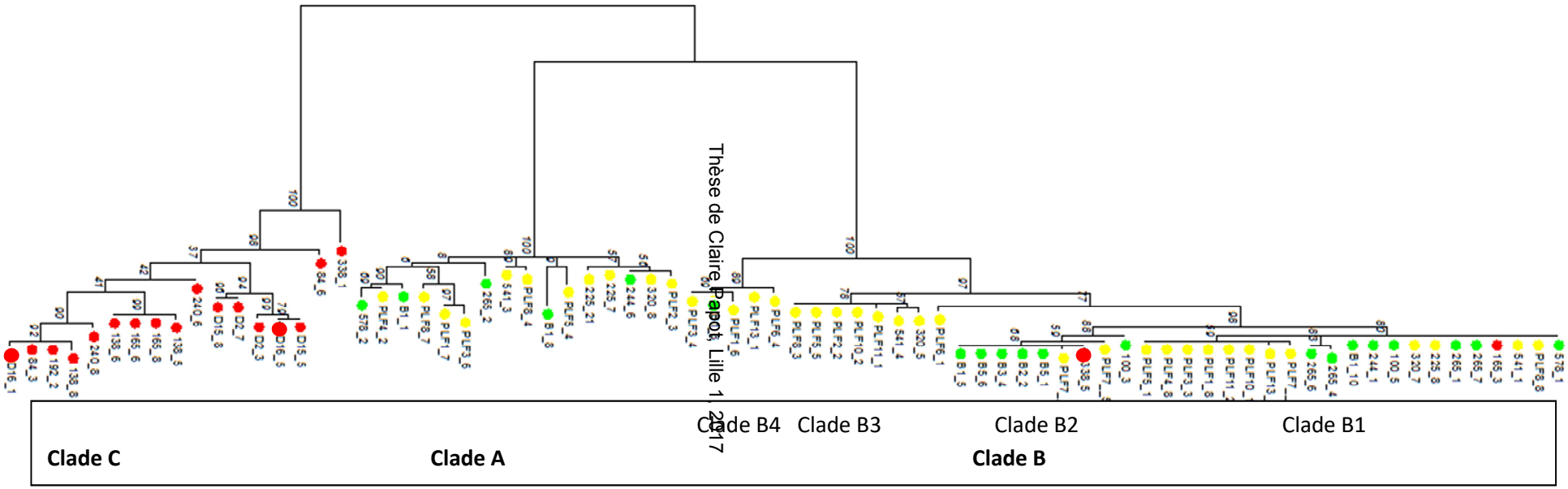
niveau de la région 3', avec leurs propres mutations (en 138.1 et 541.3) comme cela a pu être décrit chez *Alvinella pompejana* précédemment.

Tous les individus ont ensuite été alignés pour chaque région du gène et un deuxième filtrage a également été effectué pour enlever du jeu de données global (tous les individus) un nombre de singletons équivalent au pourcentage moyen de mutations artéfactuelles de type singleton calculé au sein des individus (de l'ordre de 1%). Ce jeu de données épuré (réalisé pour les régions 5' et 3') a ensuite été utilisé pour construire un arbre des allèles de la préprocapitellacine sur l'ensemble des individus échantillonnés.

4.3. Diversification du gène de la capitellacine

L'arbre des allèles du gène codant pour la partie 5' de la préprocapitellacine (régions introniques incluses) est présenté dans la Figure 12. Il a été obtenu à partir de l'alignement des données « filtrées » issues du clonage et montre trois clades principaux, appelés A, B et C rappelant les trois clades précédemment trouvés au niveau du gène mitochondrial. Le Tableau 7 récapitule les divergences nettes entre les différents groupes de séquences au sein des régions 3' et 5'. Le clade C apparaît comme étant le clade le plus divergent pour les deux régions du gène et pour les deux comparaisons avec les autres clades A et B (18 et 20% de divergence en région 5'). Les clades A et B présentent quant-à-eux 16% de divergence nette dans la région 5' et 9,6% dans la région 3'.

Dans la région exonique (entre parenthèses), les divergences nettes sont plus faibles et atteignent 8,7 et 10% pour les comparaisons avec le clade C et 8,2% entre les clades A et B en région 5'. En région 3', les niveaux de divergence chutent également à 12,8/12,9% pour les comparaisons avec le clade C et à 7% entre les clades A et B.



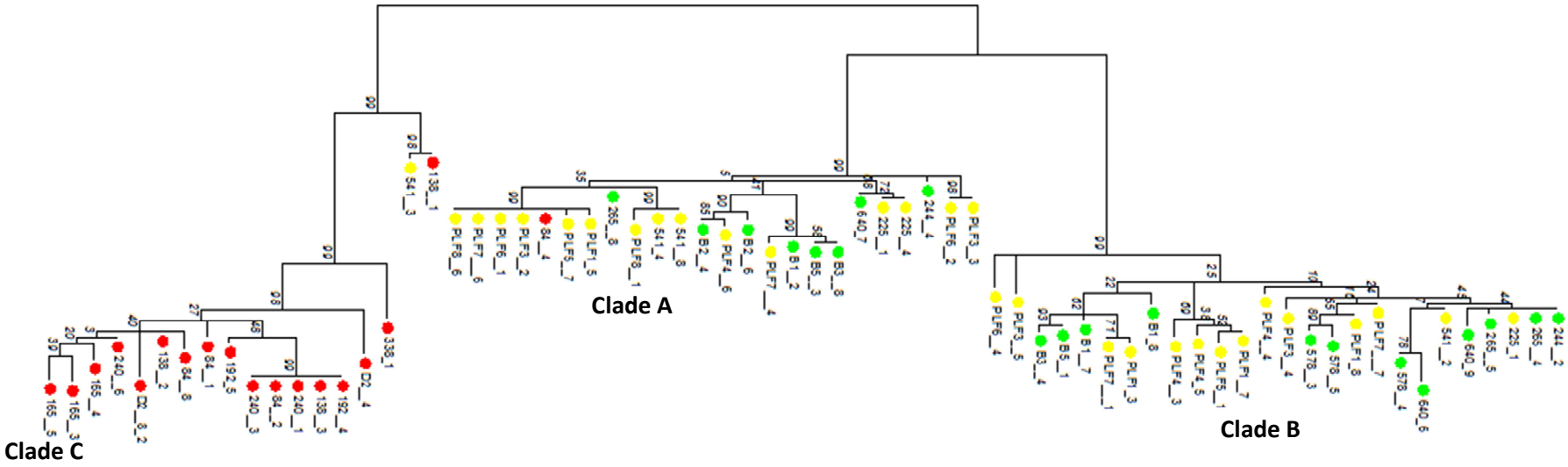


Figure 12. Arbres ML des relations phylogénétiques entre les différents allèles de la préprocapitellacine pour tous les individus de Boulogne (B), Dunkerque (D), Port-la-Forêt (PLF) et Roscoff (chiffres) pour les deux régions 5' (haut) et région 3' (bas). Le type mitochondrial a été assigné sous forme de pastille de couleur: Rouge: Cc-Manche2; Vert: Cc-Manche1 et Jaune: Cc-Atlantique

Région 5'

	Clade A	**Clade B**	**Clade C**
Clade A			
Clade B	0,16 (0,82)		
Clade C	0,18 (0,87)	0,20 (0,10)	

Région 3'

	Clade A	**Clade B**	**Clade C**
Clade A			
Clade B	0,096 (0,07)		
Clade C	0,13 (0,13)	0,15 (0,13)	

Tableau 7. Divergences nettes entre les différents groupes de séquences (clades/lignées A,B,C) calculées à l'aide du logiciel MEGA6 pour la région 5' et la région 3' (entre parenthèses = calcul de la divergence dans la région exonique).

4.4. Nombre d'allèles par individu et hypothèse de duplications du gène

En partant du principe que chaque clade correspond au fond génétique de l'espèce, l'arbre ML des allèles recapturés de la capitellacine montre, qu'outre des phénomènes d'introgression d'allèles de la capitellacine entre espèces (définies selon le type mitochondrial), il existe un nombre d'allèles par individu généralement supérieur à 2, ce qui laisse supposer la présence de duplications au sein du gène de la capitellacine, voire d'un polymorphisme du nombre de copies du gène selon les individus génotypés). Le Tableau 8 permet de récapituler le nombre d'allèles de la préprocapitellacine obtenus après un génotypage en séquençage direct de chaque individu. En assignant chaque allèle trouvé aux différents clades décrits dans la Figure 12, on peut évaluer le nombre d'allèles par individu et le statut d'hybride potentiel entre les 3 clades, toujours sous l'hypothèse que ces clades correspondent bien au « background » génétique des espèces avant hybridation à partir des informations du génotypage en région 5' et du clonage en région 3'. Ce Tableau résulte donc du croisement des données entre les 2 régions et un génotype final multi-allélique est ensuite déduit. Le nombre d'allèles total par individu varie de 2 à 5 allèles avec de nombreux individus présentant un nombre impair d'allèles laissant supposer à l'existence d'un polymorphisme du nombre de copies entre les individus faisant suite à des duplications récentes après spéciation.

Nom individu	Localité	Type mito	Capitellacine : region 5'	Na	Capitellacine : Région 3'	Na	**Recapitulatif**			Génotype final:
B1	Boulogne	M1	A1A2B1B2	4	A1 B1B2	3	**2**	**2**	**0**	A1A2B1B2
B2	Boulogne	M1	B1B2	2	A1A2B1	3	**2**	**2**	**0**	A1A2B1B2
B3	Boulogne	M1	B2**B4**	2	A1B1B2	3	**1**	**2**	**0**	A1B2B4
B5	Boulogne	M1	A1B1	2	B1	1	**1**	**1**	**0**	A1B1
244	Roscoff	M1	A1B1	2	A1B1B2	3	**1**	**2**	**0**	A1B1B2
265	Roscoff	M1	A1B1	2	A1B1	2	**1**	**1**	**0**	A1B1
578	Roscoff	M1	A1B1	2	A1B1B2	3	**1**	**2**	**0**	A1B1B2
100	Roscoff	M1	A1B1B2	3		3	**1**	**2**	**0**	A1B1B2
640	Roscoff	M1	A1B1B4	3	A1B1B2	3	**1**	**2**	**0**	A1B1B2B4
PLF1	Port la foret	Atl	A1B1B4	3	A1B1B4	3	**1**	**2**	**0**	A1B1B4
PLF3	Port la foret	Atl	A1A2 B1B4	4	A1A2B1B4	4	**2**	**2**	**0**	A1A2B1B4
PLF4	Port la foret	Atl	A1B1B1'B2	4	A1A2B1B2B3	5	**2**	**3**	**0**	A1A2B1B1'B2
PLF5	Port la foret	Atl	A1A2B1B1'B4	5	A1A2B1B2B3	5	**2**	**3**	**0**	A1A2B1B1'B4
PLF6	Port la foret	Atl	A1B3B4	3	A1A2B1	3	**2**	**2**	**0**	A1A2B3B4
PLF7	Port la foret	Atl	A1B1B1'B2	4	A1A2B1B2	4	**2**	**3**	**0**	A1A2B1B1'B2
PLF8	Port la foret	Atl	A1A2B1B3	4	A1A2B1	3	**2**	**2**	**0**	A1A2B1B3
320	Roscoff	Atl	A1B1B3	3		3	**1**	**2**	**0**	A1B1B3
541	Roscoff	Atl	A1B1B3	3	A1A2B1B2	4	**2**	**2**	**0**	A1A2B1B3
225	Roscoff	Atl	A1B1B4	3	A1B1	2	**1**	**2**	**0**	A1B1B4
PLF10	Port la foret	Atl	A1B1B3	3		/	**1**	**2**	**0**	A1B1B3
PLF11	Port la foret	Atl	A1B1B4	3		/	**1**	**2**	**0**	A1B1B4
PLF2	Port la foret	Atl	A1A2B1B3	3		/	**2**	**2**	**0**	A1A2B1B3
PLF13	Port la foret	Atl	A1B1B4	3		/	**1**	**2**	**0**	A1B1B4
240	Roscoff	M2	C1C2	2	C1C2	2	**0**	**0**	**2**	C1C2
338	Roscoff	M2	B2C1	2	C1	1	**0**	**1**	**1**	**B2**C1
84	Roscoff	M2	C1 C3	2	A1**C1C2C3C4**	5	**1**	**0**	**4**	**A1C1C2C3C4**
138	Roscoff	M2	C1C2C2'	3	A1C1C2C3	4	**1**	**0**	**3**	**A1**C1C2C3
192	Roscoff	M2	C1C2	2	C1C2	2	**0**	**0**	**2**	C1C2
165	Roscoff	M2	B1 C1	2	B1C1C2	3	**0**	**1**	**2**	B1C1C2
D2	Dunkerque	M2	C1C2	2	A1C1C2	3	**1**	**0**	**2**	**A1**C1C2
D15	Dunkerque	M2	C1C2	2	C1C2C3	3	**0**	**0**	**3**	C1C2C3
D20	Dunkerque	M2		/	C1C2	2	**0**	**0**	**2**	C1C2

Tableau 8. Nombre d'allèles par individu (Na) et génotypes synthétiques pour la préprocapitellacine chez les individus de Boulogne (B), de Dunkerque (D), de Port-la-Forêt (PLF) et de Roscoff (chiffres). Les types mitochondriaux sont : M1 : Cc-Manche1 ; Atl : Cc-Atlantique ; M2 : Cc-Manche2. Au sein des colonnes « récapitulatif » sont présentés le nombre d'allèles trouvés dans chacun des 3 clades trouvés pour la préprocapitellacine (A/B/C).

Dans le clade A, au maximum 2 copies du gène sont retrouvées au sein d'un individu, 3 copies pour le clade B et jusqu'à 4 copies pour le clade C avec dans tous les cas un polymorphisme dans le nombre de copies. Sur la base de ce génotypage synthétique des clades de la capitellacine, on observe que 100% des individus des lignées mitochondriales Cc-Manche1 et Cc-Atlantique sont hétérozygotes aux clades A et B.

Pour le clade mitochondrial Cc-Manche2, environ 50% des individus apparaissent homozygotes pour le clade C de la capitellacine. Ces individus sont retrouvés soit à Roscoff soit à Dunkerque. Cependant, il apparaît également que les individus de Roscoff de ce type mitochondrial puissent présenter des allèles du clade A ou du clade B en situation d'hétérozygotie avec le clade C. A l'inverse, aucun allèle appartenant au clade C n'est observé dans les lignées mitochondriales Cc-Manche1 et Cc-Atlantique. Le nombre de copies de chaque clade de la capitellacine varie entre individus même s'il apparaît que les individus de Port-la-Forêt (PLF7-PLF3-PLF4) aient un nombre de copies d'allèles pour le clade B plus élevé que dans les autres localités. Les copies surnuméraires B3 et B4 ont été en effet retrouvés uniquement au sein des individus de type mitochondrial Cc-Atlantique.

4.5. Réseau d'allèles de la préprocapitellacine

Le réseau d'allèles effectué dans la Figure 13 permet de montrer en plus de la présence des trois clades précédemment décrits comment les allèles se distribuent au sein de chaque clade et au sein des localités échantillonnées. Il semblerait que pour le clade C les allèles forment deux groupes entre ceux échantillonnés à Dunkerque et ceux retrouvés à Roscoff. Cette différence de fréquences au sein des localités est moins évidente pour les deux autres clades bien qu'au sein du clade B certains allèles (entourés en noir) soient plus spécifiques de Port-la-Forêt. De plus, les quelques allèles échantillonnés au sein de l'espèce *C. teleta* et ajoutés au jeu de données sont pour certains strictement identiques à ceux décrits pour l'espèce Cc-Manche2 (en violet et jaune) suggérant la encore des échanges inter-génomiques entre ces 2 espèces.

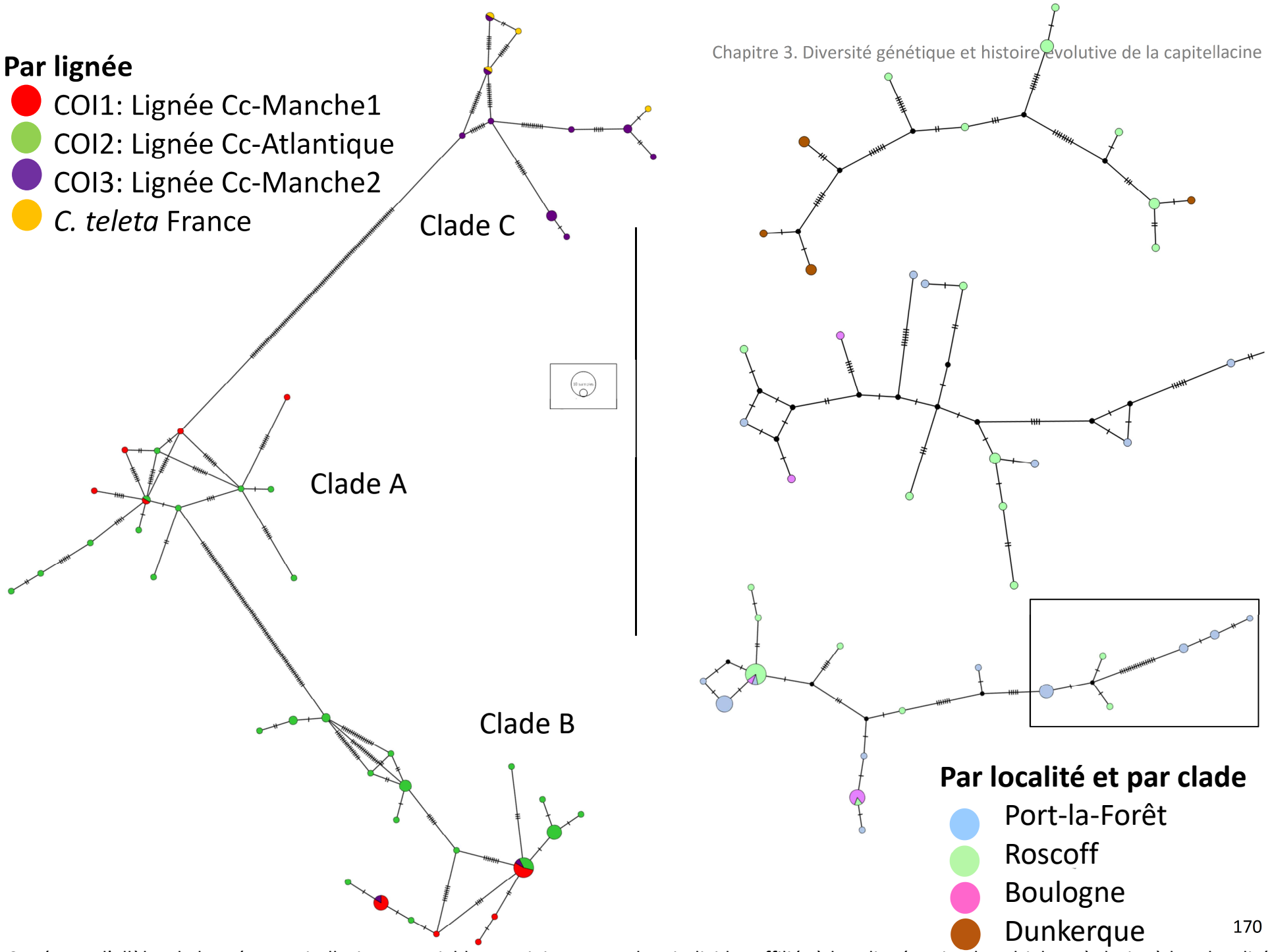


Figure 13. Réseau d'allèles de la préprocapitellacine en Neighbour Joining. A gauche : individus affiliés à leur lignée mitochondriale et à droite à leur localité pour chaque clade.

4.6. Diversité génétique de la préprocapitellacine

4.6.1. le long du gène

La Figure 14 montre que la diversité nucléotidique, en plus d'être plus forte dans les régions intronique**,** oscille autour de 0.08 (moyenne de 0.088). Dans les régions exonique, cette diversité nucléotidique diminue et oscille autour de 0.06. La région codant le PAM en lui-même montre, cette fois-ci, une diversité nucléotidique non nulle, dans un contexte d'espèces cryptique en mélange.

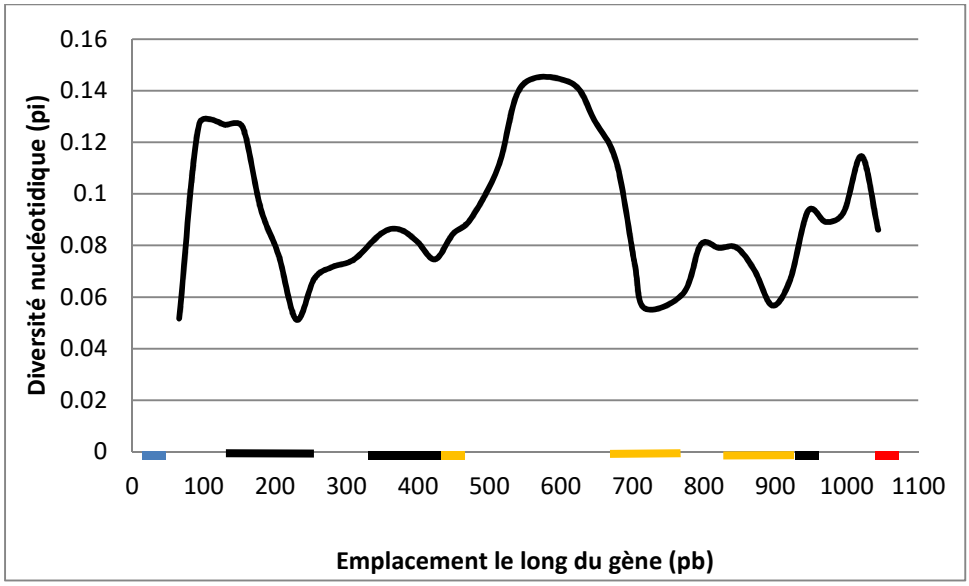


Figure 14 : Distribution de la diversité nucléotidique le long du gène (en pb) chez toutes les espèces cryptiques de *Capitella spp.* avec l'association des différents domaines associés (régions exoniques). En bleu : peptide signal, en noir: prorégion, en jaune : BRICHOS et en rouge : le PAM mature.

4.6.2. Globale par clade et localité

Pour ne pas biaiser les résultats des calculs de diversité et des tests de neutralité, deux allèles par individu ont été conservés en prenant les allèles les plus divergents pour chaque individu. Les résultats de diversités génétiques des clades A, B et C pour les régions 5' et 3' sont présentés dans le Tableau 9. La diversité génétique retrouvée en région 3' est du même ordre pour les trois clades. La diversité nucléotidique est comprise entre 0,013 (clade B) et 0,029 (clade C) en région 5' contre 0,019 et 0,023 en région 3'. Les diversités haplotypiques sont supérieures à 0,87 (clade B en région 5'). Les tests de Fu et Li et le D de Tajima montrent de valeurs oscillant autour de zéro avec des p-values non significatives.

Les diversités haplotypiques par localité sont plus faibles pour le clade B (0,49 à Roscoff et 0,53 à Boulogne) en région 5', ce qui n'est pas retrouvé dans la région 3' avec des diversités toutes supérieures à 0,99. Les diversités nucléotidiques au sein des deux régions sont les plus faibles à Roscoff et à Boulogne pour les deux clades B et C, ce qui pourrait s'expliquer si ces localités représentent une limite d'aire géographique pour des espèces qui seraient associées à ces clades. A Port-la-Forêt, les diversités (nucléotidique et haplotypique) sont nettement plus élevées pour les clades A et B au niveau des deux régions 5' et 3'.

REGION 5	Localités	N	S	Hd	Pi	θ_w	N	S	Hd	Pi	θ_w	Fu et Li D	Fu et Li F	Tajima D
Clade A	Port-la-Forêt	6	23	1,0 ± 0,096	0,015 ± 0,0021	0,014 ± 0,0072	13	34	1 ± 0,031	0,014 ± 0,0015	0,015 ± 0,0061	-0,02	-0,18	-0,53
	Roscoff	7	18	0,95 ± 0,096	**0,0087 ± 0,0017**	0,010 ± 0,0051								
Clade B	Port-la-Forêt	16	40	0,87 ± 0,079	0,016 ± 0,0038	0,016 ± 0,0064	34	40	0,87 ± 0,039	0,013 ± 0,0025	0,0135 ± 0,0045	1,1	0,85	-0,04
	Roscoff	11	18	*0,49 ± 0,175*	**0,0058 ± 0,0026**	0,0085 ± 0,0037								
	Boulogne	7	31	*0,53 ± 0,21*	0,013 ± 0,0072	0,017 ± 0,0083								
Clade C	Roscoff	10	54	0,956 ± 0,059	0,0271 ± 0,004	0,027 ± 0,012	16	63	0,97 ± 0,031	0,029 ± 0,0025	0,027 ± 0,010	-0,31	-0,19	0,2
	Dunkerque	6	30	0,87 ± 0,017	0,017 ± 0,0058	0,019 ± 0,0093								

REGION 3	Localités	N	S	Hd	Pi	θ_w	N	S	Hd	Pi	θ_w	Fu et Li D	Fu et Li F	Tajima D
Clade A	Boulogne	4	23	1 ± 0,177	0,019 ± 0,0057	0,020 ± 0,011	17	48	0,98 ± 0,031	0,019 ± 0,0021	0,023 ± 0,0085	-1,05	-1,13	-0,8
	Port-la-Forêt	7	30	0,952 ± 0,096	0,017 ± 0,0041	0,019 ± 0,0093								
	Roscoff	6	17	1 ± 0,096	**0,011 ± 0,0017**	0,012 ± 0,0061								
Clade B	Port-la-Forêt	6	38	1 ± 0,096	0,023 ±0,0046	0,024 ± 0,012	15	61	1 ± 0,024	0,023 ± 0,0025	0,027 ± 0,010	-0,57	-0,71	-0,72
	Roscoff	6	23	1 ± 0,092	**0,014 ± 0,0034**	0,014 ± 0,0072								
	Boulogne	3	6	1 ± 0,27	**0,0058 ± 0,0020**	0,0058 ± 0,0039								
Clade C	Roscoff	12	31	0,91 ± 0,079	0,014 ± 0,0019	0,016 ± 0,0069	-	-	-	-	-	-0,83	-0,83	-0.44

Tableau 9. Descripteurs de la diversité pour les différents clades de la préprocapitellacine (pour chaque population et chaque clade toutes localités confondues) avec les tests statistiques visant à détecter un écart à la neutralité, pour les régions 5' et 3' du gène. N : nombre de séquences analysées, S : nombre de sites variants, Hd : diversité haplotypique, Pi : diversité nucléotidique, qW : théta de Watterson. Les p-values des trois tests de détection d'écart à la neutralité sont toutes non significatives.

4.7. Fréquences des différents clades de la capitellacine le long de la côte.

La Figure 15 récapitule les fréquences des différents clades de la capitellacine au sein des différentes localités échantillonnées : Port-la-Forêt, Roscoff et Boulogne/Dunkerque. Les différentes couleurs représentent l'appartenance des allèles aux différents clades : violet pour le clade C (le plus divergent), bleu pour le clade A et grisé à noir pour le clade B (gris clair=B1, gris=B2, gris foncé=B3, noir=B4). Le nombre d'individus étudié pour chaque lignée mitochondrial est également présenté pour chaque localité. Un biais d'échantillonnage existe cependant puisque le gène codant pour la préprocapitellacine a été amplifié seulement pour des individus du type mitochondrial Cc-Manche2 à Dunkerque. En conséquence, seuls des allèles appartenant au clade C sont retrouvés pour cette localité. A Boulogne, seuls des individus appartenant à l'espèce Cc-Manche1 (espèce majoritaire) ont été amplifiés au gène de la préprocapitellacine. Ces deux localités ont donc été regroupées comme une seule et unique localité Nord-de-la-France. En effet, les données mitochondriales montrent que les clades Cc-Manche1 et Cc-Manche2 sont tous les 2 bien représentés à Dunkerque. À Roscoff, des individus de tous les types mitochondriaux ont été trouvés. A Port-la-Forêt, seuls des individus du type mitochondrial Cc-Atlantique sont présents (le type Cc-Manche1 étant extrêmement rare et le type Cc-Manche2 absent à cette localité).

Il apparaît que les allèles du clade C ne sont présents (Roscoff et Dunkerque), que dans le cas ou des individus du type mitochondrial Cc-Manche2 ont été observés. A Boulogne et Dunkerque, la fréquence des allèles est de presque 20% pour le clade A et de 35% pour le clade B (B1 étant minoritaire, et B2 majoritaire), les allèles du clade C sont majoritaires à 45%. La localité de Roscoff montre le même patron de fréquences avec 25% d'allèles A et 35% d'allèles des clades B et C (B1 étant majoritaire). A Port-la-Forêt, on retrouve une prédominance du clade B (>75%) avec la présence quasi « privée » de B4 (allèle spécifique à l'espèce Cc-Atlantique). Ces données montrent qu'il existe une structuration spatiale des populations de *Capitella capitata* à ce locus qui permet de réfuter l'hypothèse que les clades A, B et C de la capitellacine puissent être dus à des duplications antérieures aux événements de spéciation ayant conduit aux 3 lignées mitochondriales.

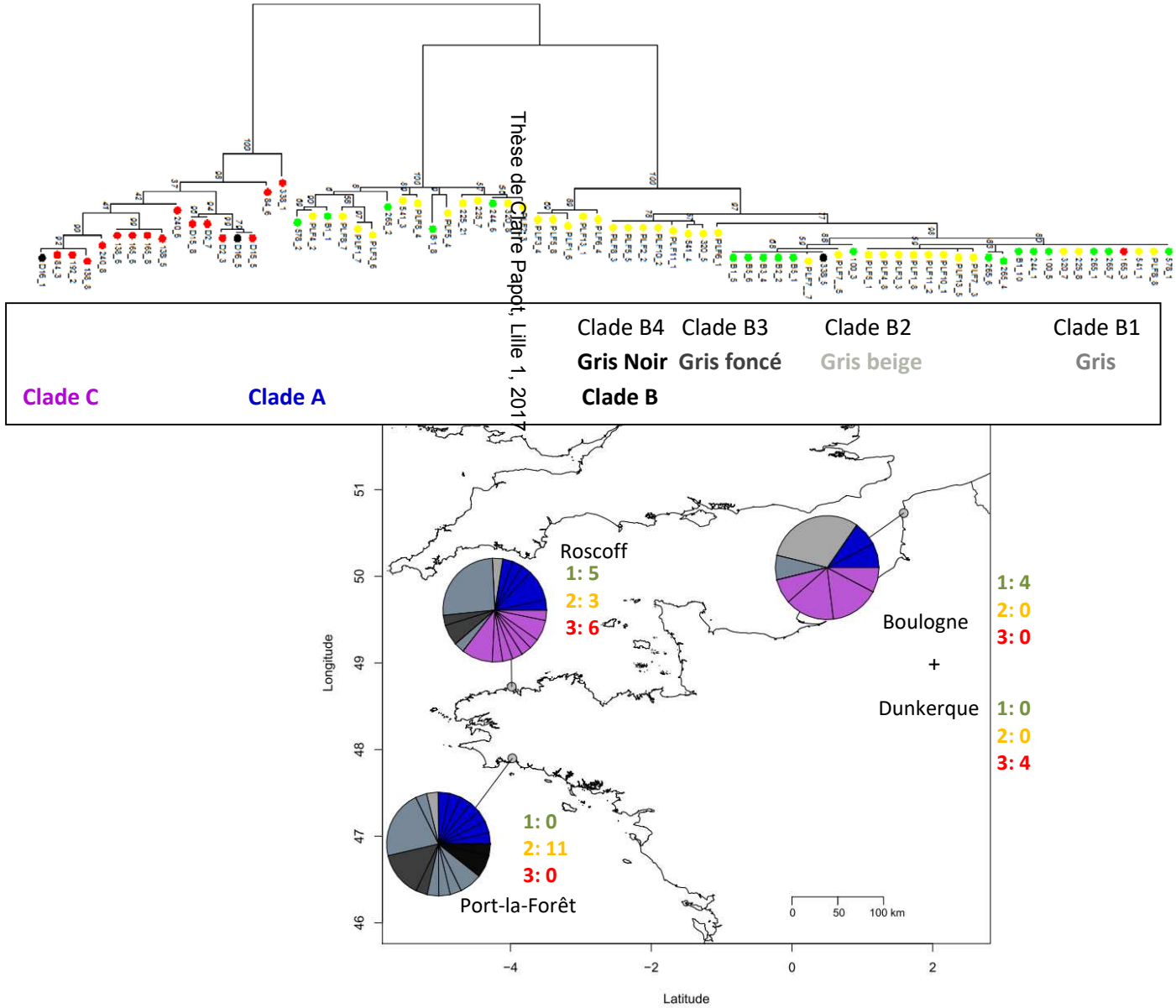


Figure 15. Carte des fréquences alléliques de la capitellacine au sein des différentes localités échantillonnées (Boulogne/Dunkerque, Roscoff, Port-la-Forêt). Les chiffres correspondent au nombre d'individus pour lesquels la capitellacine a été échantillonnée par type mitochondrial (1: Cc-Manche1; 2: Cc-Atlantique; 3: Cc-Manche2).

4.8. Analyse de la différenciation génétique (Phi_{ST}).

Les analyses de différenciation génétique ont été faites au sein de chaque clade car ces clades représentent soit des espèces, soit des duplicats de gènes. Ces résultats sont présentés dans le Tableau 10. Pour le clade B tout d'abord, le nombre de recaptures de ce clade est plus élevé en région 5' qu'en région 3', qui est en plus la région la plus longue et donc la plus résolutive pour quantifier la différenciation en terme de Phi_{ST} notamment en contenant plus de polymorphisme dans les zones introniques (les valeurs de Fst sont en Annexe 3). Le test de différenciation global est significativement différent de zéro sur les deux régions indiquant la présence d'une structuration génétique entre les populations échantillonnées. En région 5', les comparaisons entre Roscoff/Port-la-Forêt (PLF) et la population de Boulogne sont significatives au seuil de 5%. La comparaison entre PLF et Roscoff est marginalement significative à 5% (0,055) en région 5' et significative en région 3'. Ainsi ce clade B semble être différencié génétiquement de manière significative entre l'Atlantique, la Manche Orientale et la Manche Occidentale.

Pour le clade A, la tendance est plus diffuse. En effet, ce clade a été moins recapturé au sein des deux régions du gène. Même si les deux tests de différenciations génétique sont significatifs pour les deux régions 5' et 3', (0,24 p-value 0,014* et 0,34 p-value 0,006***), il semble qu'en région 5' seule la comparaison PLF – Roscoff est significative (bien que la comparaison Roscoff – Boulogne montre une p-value de 0.077) alors qu'en région 3' il s'agit des comparaisons Roscoff/PLF contre Boulogne. Pour ce clade, le nombre d'allèles recapturés est faible pouvant expliquer le manque de robustesse des résultats mais les deux régions du gène s'accordent sur le faite qu'une structuration génétique puisse être significativement retrouvée entre certaines populations de l'échantillonnage.

Le clade C quant à lui montre un test de différenciation significatif entre les deux populations Roscoff et Dunkerque.

REGION 5'

Clade B Phi_{ST} global = 0,24 pvalue=0,001***

	Boulogne	Roscoff	PLF	Boulogne	Roscoff	PLF
Boulogne (5)						
Roscoff (12)	**0,46**			**0,01**		
PLF (20)	**0,37**	0,09		**0,004**	0,055	

Clade A Phi_{ST} global = 0,24 pvalue=0,014*

	Boulogne	Roscoff	PLF	Boulogne	Roscoff	PLF
Boulogne (2)						
Roscoff (7)	0,26			0,077		
PLF (6)	0,14	**0,25**		0,18	**0,0096**	

Clade C Roscoff vs Dunkerque
Phi_{ST} global = 0,30 pvalue=0,007

REGION 3'

Clade B Phi_{ST} global= 0,36 p value 0,003**

	Boulogne	Roscoff	PLF	Boulogne	Roscoff	PLF
Boulogne (5)						
Roscoff (12)	**0,52**			**0,013**		
PLF (20)	0,23	**0,31**		0,11	**0,0013**	

Clade A Phi_{ST} global= 0,34 p value 0,006**

	Boulogne	Roscoff	PLF	Boulogne	Roscoff	PLF
Boulogne (2)						
Roscoff (7)	**0,59**			**0,0048**		
PLF (6)	**0,36**	0,045		**0,041**	0,21	

Tableau 10. Différenciation génétique des différentes populations (PLF pour Port-La-Fôret) géographiques en terme de Phi_{ST} (et p-value) calculées globalement (toutes localités confondues) et par paire de localités pour chaque région du gène (5') et (3'). Les valeurs entre parenthèses correspondent au nombre de séquence utilisées pour chaque localité.

5. Analyse IMa2- Test de l'hypothèse d'isolement des populations avec migration

Le modèle IMa2 a été utilisé pour quantifier les flux de gènes entre les différentes espèces de *Capitella spp.* étudiées sous l'hypothèse d'une introgression des allèles de la capitellacine à la suite de contacts secondaires entre les 3 lignées mitochondriales en prenant comme postulat que toutes les duplications en tandem du gène de la capitellacine sont postérieures aux événements de spéciation ayant conduit aux 3 lignées Cc-Atlantique, Cc-Manche1 et Cc-Manche2. Comme l'étude ne porte que sur 2 gènes, l'analyse fait une hypothèse forte que les clades A, B et C de la capitellacine sont effectivement associées aux 3 lignées mitochondriales utilisées pour leur identification et, donc que ce n'est pas le génome mitochondrial qui introgresse entre les différentes espèces. Ce qui compte particulièrement dans cette analyse sera surtout de discuter des flux de gènes entre les génomes des différentes espèces de *Capitella*, en sachant que ces flux sont spécifiques à la capitellacine puisque le COI permet uniquement de marquer l'espèce.

5.1. 1er scénario évolutif:

Le premier scénario évolutif testé correspond à l'hypothèse initiale que les trois clades nucléaire de la capitellacine correspondraient aux trois clades obtenus au niveau du gène mitochondrial *Cox-1*. Ici, les clades nucléaire sont considérés comment correspondant au fond génétique des lignées mitochondriales.

5.1.1. Temps de divergence et taille de population

Les paramètres démographiques estimés à l'aide du logiciel IMa2 sont présentés Tableau 11. Le temps de divergence (t0) le plus récent entre les groupes Cc-Manche1 et Cc-Atlantique est estimé à près de 4500 ans (HiPt : 4563 ans, entre 507 et 33785 ans). Le deuxième temps de divergence (t1) entre le groupe 3 (population ancestrale aux deux lignées mitochondriales Cc-Manche1 et Cc-Atlantique) et le groupe Cc-Manche2 est estimé à plus de 8000 ans (HiPt : 8143, entre 2612 et 30052) mais cette date de divergence est sans doute largement sous-estimée par la méthode de calcul. Dans les deux cas donc, l'étalement des distributions des temps de divergence est tel qu'il est probable qu'ils ne soient pas significativement différents et surement concomitants au dernier maximum glaciaire : le LGM sous l'hypothèse que l'espèce ne présente qu'une génération par an (Figure 16).

Value	t0	t1	q0	q1	q2	q3	q4
HiPt	4563	8143	4321	4705	4283	299437	307119
HPD95Lo	507	2612?	902,6	1171	672,2	59765	180830
HPD95Hi	33785	300052?	20568	19032	27251	307119	307119

Tableau 11. Valeurs modales des probabilités postérieures (HiPt) pour les paramètres temps de divergence (t) et taille de population (q) ainsi que les intervalles contenant 95% des probabilités postérieures (HPD95lo (low) et HPD95Hi (high)). Le symbole ? indique que l'intervalle de confiance est mal estimé.

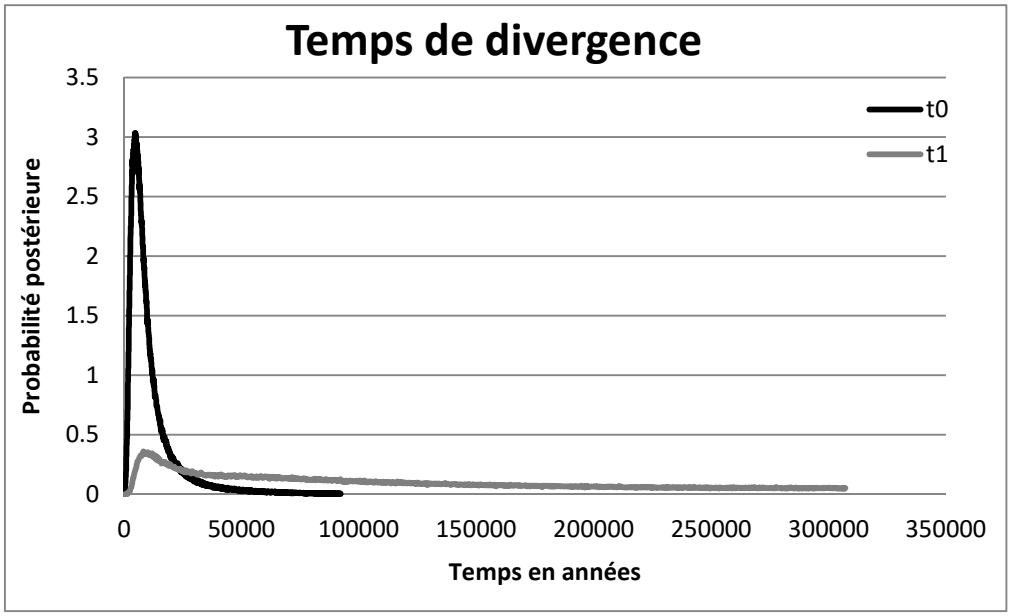


Figure 16. Distribution des probabilités postérieures du paramètre temps de divergence (t) chez *Capitella spp*.

Les tailles modales des lignées Cc-Manche1 ; Cc-Manche2 et Cc-Atlantique sont résumées dans le Tableau 11. Les trois apparaissent comme étant du même ordre de grandeur puisque les tailles de populations estimées oscillent entre 4283 et 4705 avec cependant de forts écarts concernant les HPD95 (low and High) et, sont très inférieures aux tailles des populations ancestrales q3 et q4 (HiPt = 299437 et 307119). Ces données pourraient donc suggérer que les populations actuelles aient subi un goulot d'étranglement assez prononcé. Néanmoins, le fort étalement des probabilités postérieures indiquent que les tailles des populations ancestrales ne sont pas correctement estimées (Figure 17).

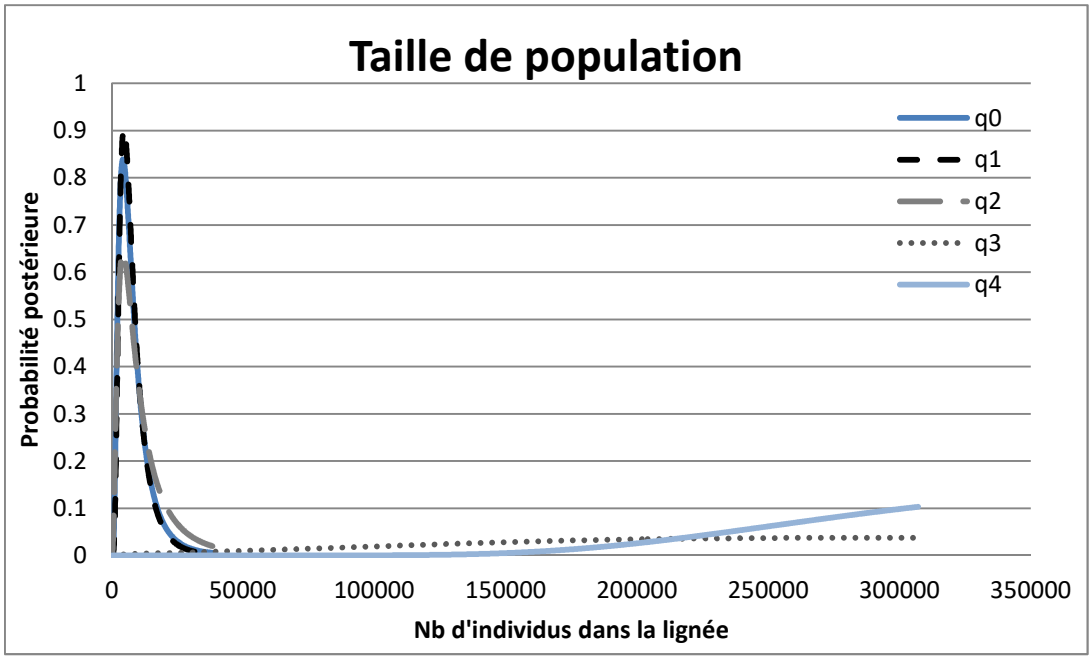


Figure 17. Distribution des probabilités postérieures marginales du paramètre taille de populations (q) chez *Capitella spp.*

5.1.2. Estimation des flux géniques (2N.m et m)

Les estimations d'échanges de migrants orientées sont réalisées 'backward in time', c'est-à-dire que le terme $N_i m_{i>j}$ est un taux d'immigration et désigne un nombre de migrants de la population j vers la population i. Ces résultats sont présentés dans le Tableau 12 et la Figure 18. Seules les Cc-Manche1 et Cc-Atlantique montrent des échanges de migrants ($2N_0m_{0>1}$ et $2N_1m_{1>0}$) significativement différents de zéro (autour de 0,6 migrants par génération). La distribution des probabilités postérieures du nombre de migrants échangés de Cc-Manche1 vers Cc-Manche2 ($2N_2m_{2>0}$) est elle aussi non nulle (autour de 0,4 migrants par génération) bien que non significative car les valeurs de HDP95 englobent zéro.

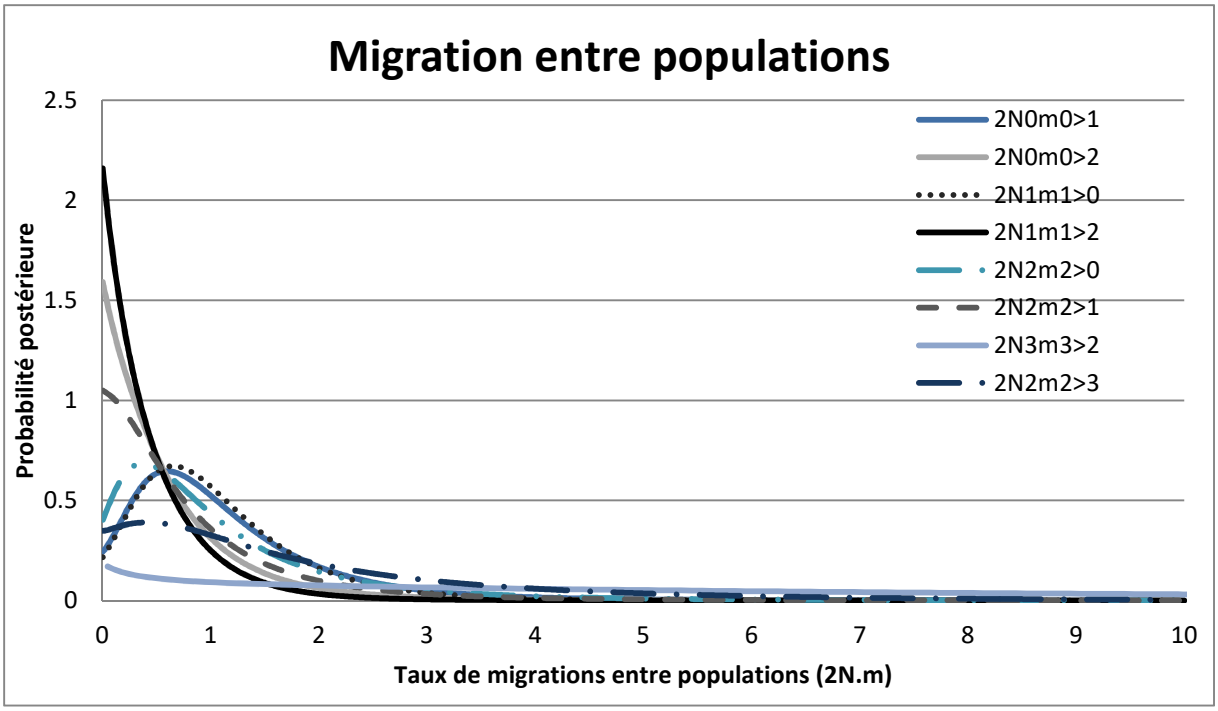


Figure 18. Distribution des probabilités postérieures marginales du paramètre taux de migration entre les différentes lignées mitochondriales de *Capitella spp.* (2N.m)

Dans le Tableau 12, seuls les taux de migrations $m_{0>1}$ et $m_{1>0}$ sont significativement différents de zéro, il est donc possible de conclure quant à un flux de gène non nul significatif et symétrique entre Cc-Manche1 et Cc-Atlantique et faire l'hypothèse que l'espèce Cc-Manche2 ait pu recevoir des allèles de l'espèce Cc-Manche1 uniquement.

Taux de Migration	$m_{0>1}$	$m_{1>0}$	$m_{0>2}$	$m_{2>0}$	$m_{1>2}$	$m_{2>1}$	$m_{2>3}$	$m_{3>2}$
HiPt	2,483	2,429	1,433	2,018	1,131	1,546	2,901	1,086
HPD95Lo	0,3025#	0,2075#	0,0#	0,0#	0,0#	0.0#	0.0#?	0.0#
HPD95Hi	4,978#	4,867#	4,138#	4,527#	3,612#	4,197#	4,997#?	3.803 # ?

Tableau 12. Taux de migration entre populations ($m_{i>j}$) avec leurs intervalles de confiance à 95% (HPD95). La présence d'un # indique que la valeur modale du taux de migration trouvée n'est pas significativement différente de zéro ou de la borne supérieure.

La Figure 19 permet de récapituler les différents résultats en termes de taille de population, divergence et migration. En effet, la largeur des différentes cases est proportionnelle à la taille de population (theta), la hauteur au temps de divergence (t) et les flèches aux événements de migration entre populations lorsqu'ils sont significativement différents de zéro.

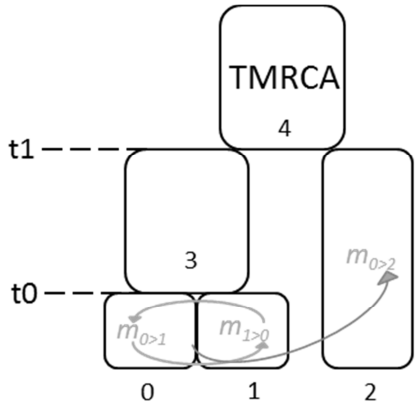


Figure 19. Arbre récapitulatif du scénario évolutif d'échanges de migrants entre les différentes lignées mitochondriales à l'aide du logiciel IMa2. 0: Cc_manche1; 1: Cc_Atlantique; 2: Cc_manche2; 3: population ancestrale à Cc_manche1 et Cc_atlantique et 4: population ancestrale à toutes les lignées mitochondriales contemporaines et passées.

5.2. 2ème scénario évolutif.

Le deuxième scénario évolutif fait l'hypothèse de l'existence d'une duplication qui serait survenue avant la spéciation des deux phylogroupes Cc-Manche1 et Cc-Atlantique mais après la spéciation de l'espèce Cc-Manche.

5.2.1. Temps de divergence et taille de population

Le temps de divergence (t0) le plus récent entre les groupes Cc-Manche1 et Cc-Atlantique est estimé à près de 300 000ans (Moyenne du Tableau 13 et Figure 20). Le deuxième temps de divergence (t1) entre le groupe 3 (population ancestrale aux deux lignées mitochondriales Cc-Manche1 et Cc-Atlantique) et le groupe Cc-Manche2 est estimé quant à lui à près de 1.2millions d'années (500000 à partir de la valeur d'HiPt) bien que dans les deux cas les estimations semblent mal estimées. Ces valeurs sont largement supérieures aux valeurs estimées précédemment

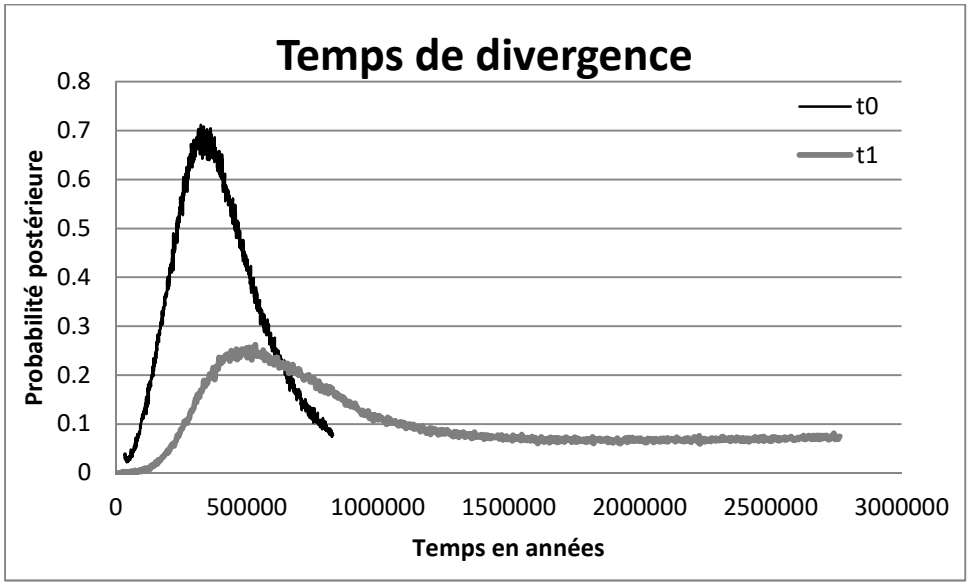


Figure 20. Distribution des probabilités postérieures du paramètre temps de divergence (t) chez *Capitella* spp (Scénario 2).

Les tailles modales des lignées Cc-Manche1 ; Cc-Manche2 et Cc-Atlantique sont résumées dans le Tableau 13 et Figure 21. Elles apparaissent comme étant oscillantes entre 46000 (Cc-Manche1) et 116280 (Cc-Manche2) avec cependant de forts écarts (Tableau 13) et sont quoiqu'il en soit également très inférieures aux tailles des populations ancestrales q3 et q4.

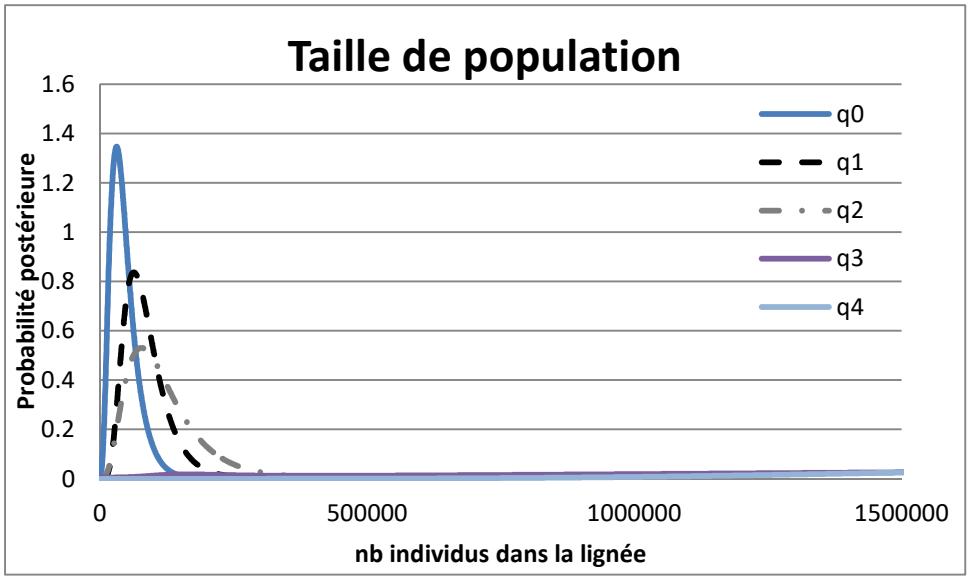


Figure 21. Distribution des probabilités postérieures marginales du paramètre taille de populations (q) chez *Capitella spp.* pour le scénario 2.

Value	t0	t1	q0	q1	q2	q3	q4
HiPt	324755	532271	31625	63769	77940	2763663	2763663
Mean	***388193***	***1176510***	***45955***	***86066***	***116270***	1700410	2076560
95%Lo	27789	264062	11579	29551	30588	161755	1018919
95%Hi	757761	2669651	111120	189233	274258	2724952	2736012
HPD95Lo	1244 #?	230881 ?	6394	21256	19183	103689 #?	1187587
HPD95Hi	743659 #?	2763663 ?	97641	168149	246607	2763663 #?	2763663

Tableau 13. Valeurs modales des probabilités postérieures (HiPt) et moyennes pour les paramètres temps de divergence (t) et taille de population (q) ainsi que les intervalles contenant 95% des probabilités postérieures (HPD95lo (low) et HPD95Hi (high)). Le symbole ? indique que l'intervalle de confiance est mal estimé. Le symbole « # » indique que les valeurs sont non significativement différente de zéro.

5.2.2. Estimation des flux géniques (2N.m et m)

Seules Cc-Manche1 montre des échanges de migrants ($2N_1m_{1>0}$ et $2N_2m_{2>0}$) significativement différents de zéro (compris entre 1.2 et 0.07) avec Cc-Atlantique et Cc-Manche2 respectivement. Les résultats sont présentés dans le Tableau 14 et la Figure 22.

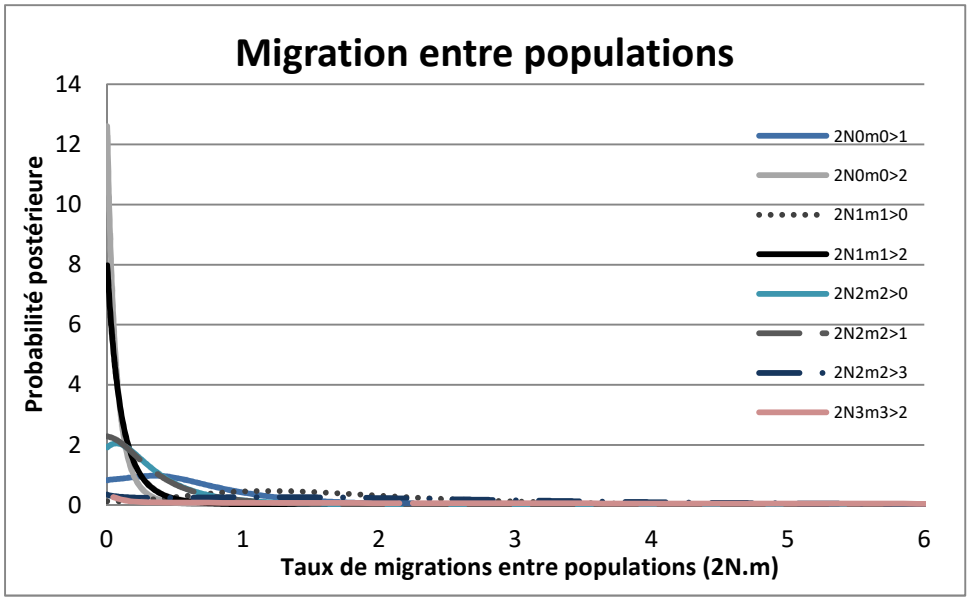


Figure 22. Distribution des probabilités postérieures marginales du paramètre taux de migration entre les différentes lignées mitochondriales de *Capitella spp*. (2N.m) pour le scénario 2.

Value	$m_{0>1}$	$m_{1>0}$	$m_{0>2}$	$m_{2>0}$	$m_{1>2}$	$m_{2>1}$	$m_{2>3}$	$m_{3>2}$
HiPt	0.0025	**3.562**	0.0025	0.1025	0.0025	0.0025	4.997	0.0025
Mean	2.343	**2.902**	0.3587	0.5481	0.2864	0.5469	2.817	1.293
95%Lo	0.0775	**0.2825**	0.0075	0.0175	0.0075	0.0125	0.1325	0.0175
95%Hi	4.867	**4.893**	1.877	2.303	1.377	2.388	4.912	4.562
HPD95Lo	0.0#?	**0.5475 #**	0.0#	0.0#	0.0#	0.0#	0.0#?	0.0#
HPD95Hi	4.997 #?	**4.997 #**	1.288 #	1.667 #	0.9725 #	1.722 #	4.997 #?	4.147 #

Tableau 14. Taux de migration entre populations ($m_{i>j}$) (moyenne et valeur modale) avec leurs intervalles de confiance à 95%. La présence d'un # indique que la valeur modale du taux de migration trouvée n'est pas significativement différente de zéro ou de la borne supérieure.

Seul le taux de migrations $m_{1>0}$ (en gras) montre une valeur différente de zéro bien que la présence d'un # indique l'absence de significativité. Avec les données précédentes, il est cependant possible de conclure quant à l'existence de façon significative d'un flux de gène entre Cc-Manche1 et Cc-Atlantique. Ainsi, cette analyse indique l'existence de flux inter-génomiques qui est cette fois-ci orienté de Manche1 vers Atlantique (non symétrique donc)

de façon significative mais aussi de Cc-Manche1 vers Cc-Manche2 bien que les valeurs estimés soient beaucoup plus faible.

6. Evolution du ratio d_N/d_S au sein et entre les clades de la capitellacine.

6.1. Comparaisons inter-clade (toutes régions confondues)

L'analyse des ratios de la divergence non synonyme sur la divergence synonyme permet de renseigner sur l'action de la sélection positive qui peut opérer le long du gène dans une lignée ou un duplicat particulier(e). Pour chaque comparaison, les valeurs de divergence d_N/d_S ont été calculées par paire de séquences entre 2 clades correspondants et moyennés sur l'ensemble des paires de séquences. Les résultats sont présentés dans le Tableau 15. Ceci permet de mettre en évidence que la plupart des mutations qui sont fixées dans la divergence entre clades sont principalement des mutations synonymes donnant des valeurs de d_N/d_S de l'ordre de 0,23 et 0,25 pour les comparaisons Clades A et B avec le clade C et une valeur 3 fois plus élevée (d_N/d_S: 0,636) pour la comparaison entre les Clades A et B. Cette valeur semblerait traduire d'un relâchement des pressions de sélection sur l'évolution des clades A et B par rapport au clade C, qui pourrait s'expliquer soit par le fait que les lignées Cc-Manche1 et Cc-Atlantique soient associées aux mêmes contraintes environnementales, soit que les clades A et B constituent en fait des paralogues issus d'une duplication antérieure à la spéciation des lignées Cc-Manche1 et Cc-Atlantique.

	Clade A		Clade B		Clade C
Clade A	d_N	d_S			
	Ratio				
Clade B	*0,082*	*0,12*			
	0,636				
Clade C	0,06	0,25	0,083	0,28	
	0,232		**0,249**		

Tableau 15. Taux de mutations non synonymes par sites NS (d_N) et de mutations synonymes par sites S (d_S) en utilisant un modèle K2P) et ratio de ces taux entre les différents clades.

6.2. Par région

Par équivalence avec l'étude réalisée au chapitre précédent, ce même ratio moyen a été calculé par région dans le but d'étudier l'action potentielle de la sélection sur ces différentes régions du préprocapitellacine (Tableau 16). Alors que le peptide signal est rigoureusement monomorphe sur l'ensemble des individus séquencés, un seul d_N/d_S est supérieur à 1 au sein de la pro-region2 (partie liant le BRICHOS au PAM) au niveau de la comparaison des clades A et B (1,71) (6 sites impactés sur une longueur de 13 acides aminés). A l'exception du peptide signal, les régions présentent toutes des ratios inférieurs à 1 mais relativement élevés pour une protéine, suggérant un relâchement des pressions de sélection entre clades mais les valeurs observées entre les clades A et B sont systématiquement plus élevées que celles associant le clade C. La région du PAM mature est très polymorphe et présentent des valeurs de d_N/d_S élevées entre le clade A et les autres clades (0,54 et 0,46 respectivement contre 0,182). Toutes les comparaisons ont été réalisées en l'absence des recombinants inter-clades 'naturels' précédemment mis en évidence.

Peptide signal		Clade A	Clade B	Clade C
	Clade A			
	Clade B	M		
	Clade C	M	M	

Proregion-1		Clade A	Clade B	Clade C
	Clade A			
	Clade B	0,672		
	Clade C	0,274	0,291	

BRICHOS		Clade A	Clade B	Clade C
	Clade A			
	Clade B	0,615		
	Clade C	0,163	0,223	

Proregion-2		Clade A	Clade B	Clade C
	Clade A			
	Clade B	1,710		
	Clade C	0,288	0,390	

PAM		Clade A	Clade B	Clade C
	Clade A			
	Clade B	0,540		
	Clade C	0,460	0,182	

Tableau 16. d_N/d_S moyen par région pour chaque paire de clades (Peptide signal, Prorégion, BRICHOS et PAM). M : séquences monomorphes.

6.3. Le long du gène

Ces variations du d_N/d_S se traduisent le long du gène par deux pics de divergence non-synonyme élevée (>3,5) dans la première partie de la prorégion pour les comparaisons des clades A et B avec le clade C (Figure 23).

La comparaison des clades A et B révèle l'existence d'un pic de d_N/d_S au niveau du domaine BRICHOS (jusqu'à 9,3) et de la prorégion2 mais également sur le PAM lui-même (pic à 2,2). A l'inverse, les comparaisons des clades A et B avec le clade C montre un pic de d_N/d_S (>3) plutôt situé au début de la prorégion1, indiquant que les pressions de sélection ne sont pas les mêmes selon les clades comparés.

En termes d'évolution du π_N/π_S, le clade A (dans le graphique en titre souligné) ne montre pas de variation remarquable de ce ratio, compris entre 0 et 1, avec une absence de variation non-synonyme de la prorégion1 au BRICHOS et sur le PAM lui-même. Le clade B montre plusieurs pics excédant 1 dans la région BRICHOS. Le clade C montre quant à lui le même patron que le clade A avec cependant un pic très élevé (4,4) dans la prorégion presque synchrone avec la divergence non-synonyme décrite précédemment entre ce clade et les 2 autres clades. La région du PAM pour ce clade apparait comme étant monomorphe (ratio de π_N/π_S nulle).

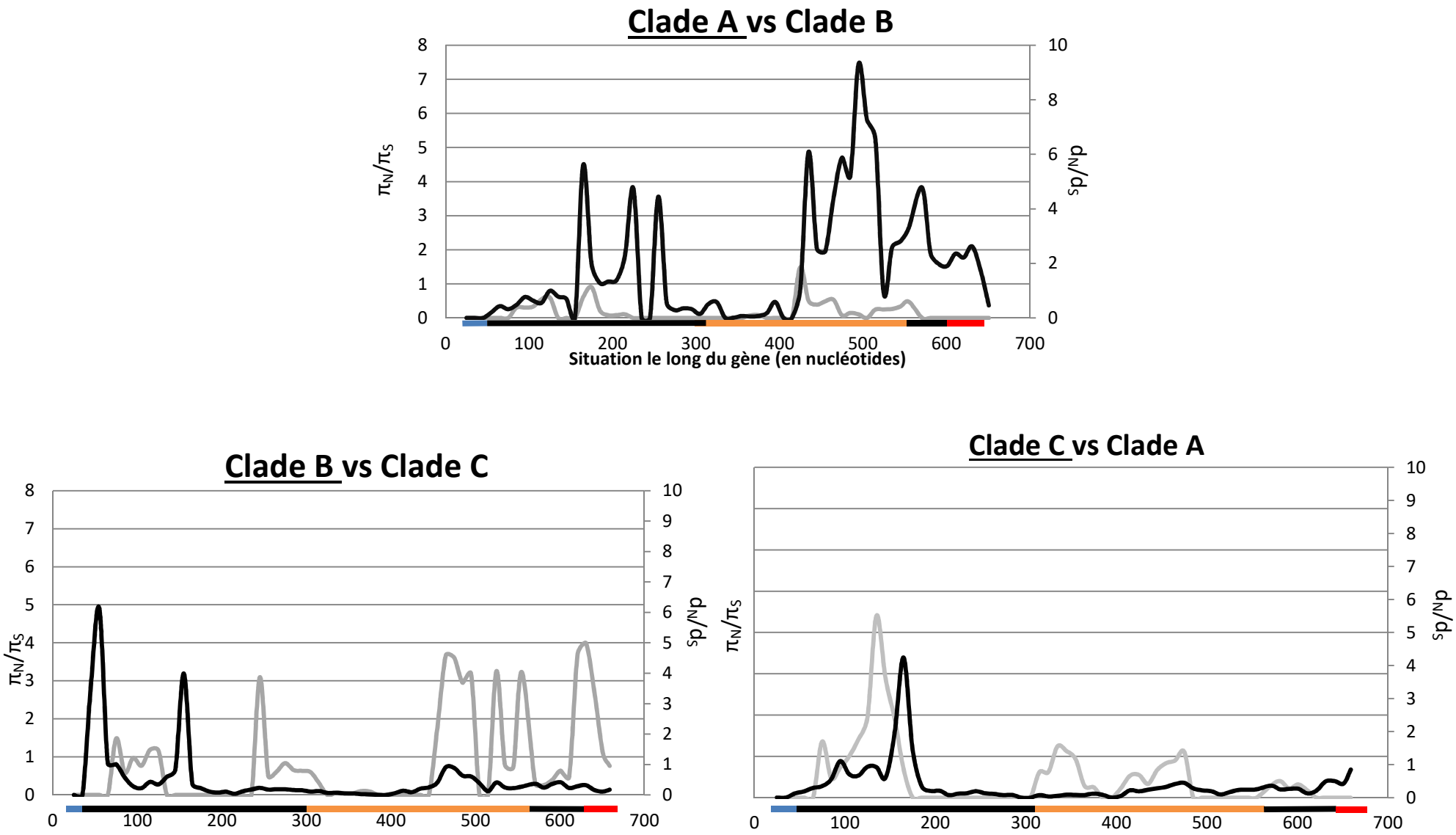


Figure 23. Evolution du d_N/d_S (échelle de droite, courbe noire) et du π_N/π_S (échelle de gauche, courbe grisée) le long du gène pour les comparaisons inter-clades de la préprocapitellacine (3 graphes). Les barres de couleurs correspondent aux différents domaines de la preprocapitellacine: en bleu, le peptide signal, en noir la prorégion, en jaune le BRICHOS et en rouge le PAM. Chaque courbe π_N/π_S est montrée pour chaque clade et correspond au clade qui est surligné dans le titre du graphique.

La Figure 24 permet d'étudier plus en détails l'évolution de ces 2 ratios au sein du clade B sachant que ce clade est beaucoup plus diversifié que les 2 autres, avec une forte composante géographique (récapitulé en Figure 12 : clades B1 à B4). Pour la suite, ont été ensuite appelés clade BA le regroupement des clade B1 et B2 et le clade BB le regroupement des clade B3 et B4. Les calculs de d_N/d_S entre clades montrent de façon quasi-systématique, mais avec des intensités différentes, 3 zones de la préprocapitellacine pour lesquelles on observe une augmentation nette de ce ratio. Ces zones correspondent au début et à la fin de la prorégion1 et à la fin du BRICHOS, avec une augmentation très forte de la divergence non-synonyme dans cette dernière zone lorsque les clades BA et BB sont comparés. Pour rappel les clades B3 et B4 (BB) sont retrouvés majoritairement au sein d'individus Cc-atlantique (Port-la-Forêt/Roscoff) et un individu de Boulogne (Cc-Manche1). Les autres clades B1 et B2 (BA) sont retrouvés chez les individus de toutes les localités mais plus spécifiquement dans la lignée Cc-Manche1 et un individu de Cc-Manche2.

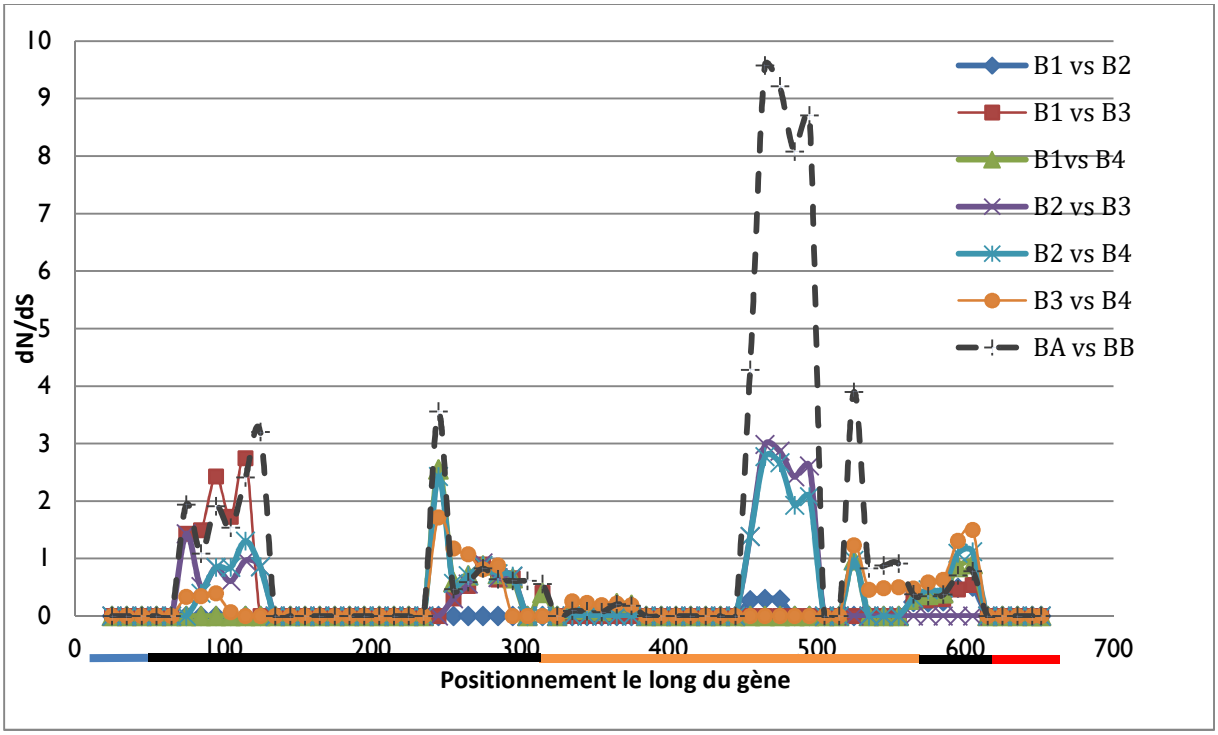


Figure 24. Evolution du d_N/d_S le long du gène codant le précurseur protéique (en bleu: peptide signal; en noir: prorégion; jaune: BRICHOS; rouge: PAM) entre les variants alléliques des sous-clades du clade B. (BA : B1 et B2 ; BB : B3 et B4)

6.4. Tests de MacDonald et Kreitman

Des tests MK ont été réalisés sous l'hypothèse que les clades A,B et C représentent bien les lignées Cc-Atlantique, Cc-Manche1 et Cc-Manche2, ces résultats sont donc conditionnés par cette hypothèse et pourrait être biaisée par la présence de paralogues. Les résultats de ce test sont présentés dans le Tableau 17 et permettent de montrer des p-value significatives pour la comparaison des clades B et C pour les deux régions et une p-value marginalement significative pour la comparaison des Clades A et C pour la région 3'. Ainsi dans les comparaisons avec le clade C, la valeur de d_N/d_S est inférieure à la valeur de π_N/π_S ce qui pourrait corroborer l'hypothèse d'une action de sélection balancée sur les polymorphismes des différents clades ou d'un relâchement des pressions de sélection sur le gène entraînant un excès de mutations faiblement délétères dans le polymorphisme. Pour la comparaison des clades A et B, la valeur de d_N/d_S est supérieure à la valeur de π_N/π_S et laisse supposer qu'une grande partie de la variation non-synonyme trouvée est impliquée dans la divergence de ces 2 clades, et ce malgré la forte variabilité géographique non-synonyme détectée au clade B. Un test a été effectué en utilisant seulement les clades B1 et B2 versus le clade A et la p-value est significative (p-value=0,044) entre ces deux clades avec toujours un $d_N/d_S>\pi_N/\pi_S$.

	Région 5'	Non syn	Syn	p-value (Fisher's exact test)
Clade A vs B	Polymorphisme au sein du clade A	6	6	
	Polymorphisme au sein du clade B	12	5	
	Divergence entre A et B	**13**	**11**	**0.56**
Clade B vs C	Polymorphisme au sein du clade B	12	5	
	Polymorphisme au sein du clade C	13	12	
	Divergence entre B et C	**8**	**19**	**0.026***
Clade A vs C	Polymorphisme au sein du clade A	6	6	
	Polymorphisme au sein du clade C	13	12	
	Divergence entre A et C	**9**	**16**	**0.23**

	Region 3'	Non syn	Syn	p-value (Fisher's exact test)
Clade A vs B	Polymorphisme au sein du clade A	14	12	
	Polymorphisme au sein du clade B	23	12	
	Divergence entre A et B	**12**	**3**	**0.20**
Clade B vs C	Polymorphisme au sein du clade B	23	11	
	Polymorphisme au sein du clade C	8	7	
	Divergence entre B et C	**12**	**21**	**0.022***
Clade A vs C	Polymorphisme au sein du clade A	14	12	
	Polymorphisme au sein du clade C	8	7	
	Divergence entre A et C	**12**	**23**	**0.055**

Tableau 17. Résultats des tests de Mac Donald et Kreitman par paire de clades et valeur du test de Fisher associée à chaque comparaison.

7. Polymorphisme et divergence au niveau de la région du PAM.

La région qui code le PAM (23 codons en région N terminale) est particulièrement polymorphe avec onze variants séquencés (haplotypes). Chaque variant de la Figure 25 a été retrouvé chez au moins deux individus.

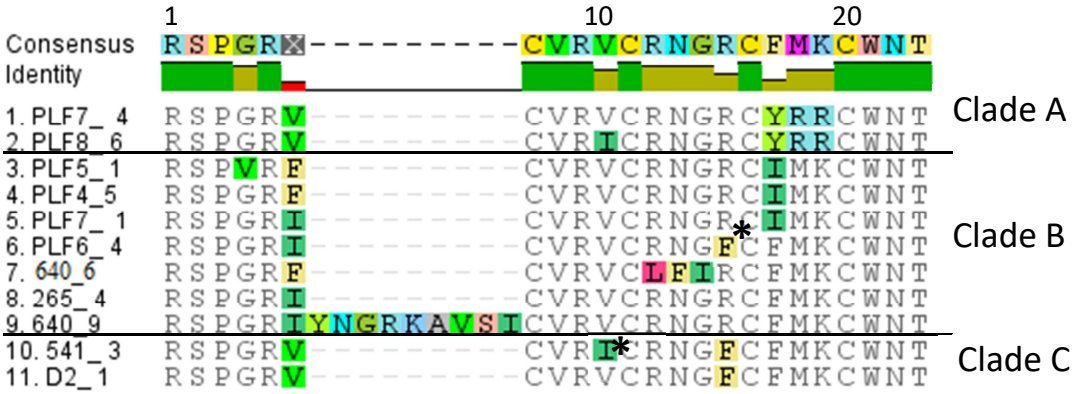


Figure 25. Séquences des PAM appartenant aux clades A, B et C. Les étoiles indiquent des points de convergence entre clades pouvant être du à de la conversion génique ou de la recombinaison inter-clades.

7.1. Description et fréquences des PAMs dans les clades et types mitochondriaux

Le clade A de la capitellacine possède deux variants YRR (mutations fixées) avec un seul site non synonyme polymorphe (I10V). Ce site I10V est retrouvé à 25% à Boulogne (1/4) le reste des indivus possédant l'autre variant de ce clade. Ce variant est présent à 86% au sein de Port-La-Forêt (6/7). A Roscoff, ce variant est présent à 66% (4/6). Au sein des deux lignées mitochondriales : le variant est présent à 42% chez Cc-Manche1 et 78% chez l'espèce Cc-Atlantique.

Le clade B présente trois mutations synonymes fixées dans la divergence des clades et une mutation non synonyme polymorphe (F/I) propre au clade en position 6, les autres mutations polymorphes sont toutes non-synonymes pour coder 6 peptides antimicrobiens différents (tous recapturés chez plus d'un individu) dont un peptide possède une insertion de 9 acides aminés (appelé « X+9 »). Ce dernier peptide n'est cependant retrouvé que chez un individu (individu 640 « parasité » par *Thiomargarita*). Un autre variant diverge dans sa structure primaire de 3 acides aminés contigües (LFI en position 12-13-14). Les autres variants de PAM pour le clade B sont décrits dans le Tableau 18. La diversité est telle que

tous ses PAMs semblent équitablement répartis entre espèces et localités bien que le variant LFI 12-13-14 est présent uniquement de Roscoff (lignée mitochondriale Cc-Manche1).

Le PAM du clade C ne diffère des PAMs du clade B que par une seule modification en acide aminé fixée en position 15 (F15R) plus deux mutations synonymes (T51C et T57C). Les PAMs des clades B et C (2 mutations NS fixées) sont donc plus proches entre eux qu'ils ne le sont des PAMs du clade A (4 mutations NS fixées). Les mutations qui ségrégent dans le polymorphisme du clade C sont des mutations synonymes (A9G et T87C).

Finalement, l'existence de 2 mutations NS polymorphes inter-clades (étoiles dans la Figure 25 ; retrouvés chez deux individus à chaque fois) pourrait être la résurgence d'un polymorphisme ancestral et/ou de la conversion génique entre allèles.

Une de ces mutations (I10V) est responsable de l'existence d'un PAM recombiné entre clade C et Clade A qui se retrouve chez des individus de l'espèce Cc-Manche 2 et Cc-Atlantique (138 et 541). Un autre polymorphisme inter-clade peut également être souligné puisqu'un peptide possède une mutation non synonyme (F15R) diagnostique du clade C alors que le reste du peptide est diagnostique du clade B (V6I notamment). Ce patron ayant été retrouvé avec les mêmes points de recombinaison chez deux individus de Port-La-Forêt appartenant au même clade Cc-Atlantique cette fois-ci. De plus, un autre PAM (non présenté dans la Figure 25 mais répertorié dans le Tableau 18) issu d'un évènement de recombinaison peut être mis en évidence chez les individus de Dunkerque uniquement puisqu'il apparait dans le Tableau 18 qu'un PAM caractérisé chez 3 individus (D2, D15 et D16 de la lignée mitochondriale Cc-Manche2 donc) possède la mutation I6V retrouvée au sein du clade B ainsi qu'une mutation cette fois qui lui est propre (N22K).

Les variants considérés comme recombinants sont retrouvés chez les individus de la lignée Cc-Atlantique mais aussi Cc-Manche2 de Dunkerque à Port-La-Forêt.

ind	Localité	Clade A (YRR)		Clade B (F/I 6)					RECOMB		Clade C (F15)	
				F6			I6		I6	V6	V6	I6
Mutations (type et position)				LFI	I17	I17	I17	F17	F15	F15		K22
		I10	V10			V4				I10		
B1	B		X				X					
B2	B	X			X							
B3	B		X		X							
B5	B		X		X							
244	R-P P*	X						X				
265	R-P	X						X				
578	R-L			X		X						
640	R-L P*		X	X				(+6)				
PLF1	PLF	X				X	X					
PLF3	PLF	X					X		X			
PLF4	PLF	X			X							
PLF5	PLF	X				X						
PLF6	PLF	X							X			
PLF7	PLF	X	X				X					
PLF8	PLF	X				X						
541	R-L	X						X		X		
225	R-P		X					X				
240	R-P P*										X (2)	
84	R-P	X									X (2)	
138	R-P									X	X (2)	
192	R-P										X(1)	
165	R-P										X (2)	
D2	D										X (2)	X
D15	D											X
D16	D											X

		G9		A9
Nom ind	Localités	T87	C87	
240	R-P P*	X		X
84	R-P	X		X
138	R-P	X		X
192	R-P			X
165	R-P	X	X	
D2	D	X	X	
D15	D		X	
D16	D			X

Tableau 18. Présence/Absence des variants de capitellacine au sein de chaque individu

7.2. Nombre de variants du PAM par individus

Ainsi, le nombre de variants par individu est dans beaucoup de cas supérieur à 2 dans la lignée mitochondriale Cc-Atlantique mais reste globalement égal à 2 dans la lignée Cc-Manche1 sauf pour l'individu 640 décrit en association avec la bactérie sulfo-oxydante. Tous les individus possèdent un PAM des clades A et B dans les lignées Cc-Atlantique et Cc-Manche1. A Roscoff, certains individus possèdent trois variants du PAM bien que cela ne soit pas la majorité alors que les individus de Boulogne/Dunkerque ne présentent systématiquement que 2 variants. Finalement à Port-la-Forêt, trois individus sur 7 (PLF1 – PLF3 – PLF7) possèdent trois variants du PAM (dont toujours un variant du clade A et un du clade B) (Tableau 18). Les individus du clade C possède tous un PAM diagnostique purement du clade C et pour certains un PAM issus d'évènements de recombinaison avec les autres clades A et B.

DISCUSSION

L'utilisation de plusieurs marqueurs génétiques pour inférer l'histoire évolutive d'un organisme est de plus en plus courante puisque les inférences réalisées sur un seul et unique gène permettent de retracer l'histoire évolutive du gène mais pas forcément celle de l'organisme. En effet, l'histoire des gènes et l'histoire des organismes ne sont pas toujours parallèles (Funk and Omland, 2003). Au niveau mitochondrial par exemple l'existence de numts (nuclear mitochondrial pseudogenes : Song et al. 2008), ou encore d'introgression mitochondriale comme chez le gastéropode *Littorina obtusata* (Kemppainen et al., 2009) représentent autant de mécanismes évolutifs qui peuvent biaiser ces inférences. C'est par exemple le cas d'*Hediste* et de *Littorina* ou un seul génome mitochondrial est présent au sein d'un complexe d'espèces s'hybridant (Audzijonyte et al., 2008) – on parle alors d'un balayage sélectif interspécifique. De plus, des effets de sélection dirigée sont souvent détectés sur le génome mitochondrial des espèces (Bazin et al., 2006; Meiklejohn et al., 2007) et l'utilisation du gène mitochondrial *Cox-1* comme marqueur d'identification des espèces peut alors être remis en question puisque l'assignation d'individus à une espèce ou à une autre peut en conséquence être biaisée. Par exemple chez trois espèces du genre *Nasonia*, la comparaison des ratios de substitutions non synonymes/synonymes entre espèces (Ka/Ks) par rapport aux ratios des substitutions polymorphes (π_N/π_s) a permis de révéler que, alors que les gènes mitochondriaux atp6 et atp8 montrent l'action d'une

sélection diversifiante, le gène *Cox-1* montre quant-à-lui une absence de polymorphisme en acides aminés suggérant l'action de la sélection purifiante (Oliveira et al., 2008). Ceci illustre la possibilité d'évènements de sélection –non homogène- sur l'évolution du génome mitochondrial déjà montré chez la Drosophile (Ballard and Kreitman, 1994) et l'intérêt de confronter des données de plusieurs locus dans le but d'inférer l'histoire de l'espèce de façon robuste. Dans le cas de cette étude, les faibles valeurs de d_N/d_S calculés entre les différentes lignées mitochondriales de *Capitella*, même au niveau mondial, suggère que le gène *Cox-1* serait dans ce cas sous forte sélection purifiante. Les divergences observées entre nos espèces sont uniquement basées sur des mutations synonymes et sont donc la marque d'une accumulation de mutations après des événements d'isolements géographiques laissant supposer à l'existence de 3 espèces cryptiques (cf. barcode gap) dans le complexe d'espèces françaises affilié à *Capitella capitata*. En effet, les clades français sont plus proches phylogénétiquement des clades mondiaux identifiés comme appartenant à l'espèce *C. capitata*, bien que présentant une divergence moyenne de 23% par rapport aux espèces *C. capitata* retrouvées au Canada, Inde et Méditerranée. De plus, une espèce retrouvée à Concarneau est affiliée à *C. teleta* puisque plus proche phylogénétiquement de ce complexe au niveau mondial (26% de divergence avec teleta contre 31% avec *C. capitata*).

Ces résultats montrent que la diversification du genre *Capitella* est très ancienne malgré l'absence de critères morphologiques diagnostiques et, suggèrent que ces espèces ont pu avoir des histoires parallèles au niveau mondial. Les annélides du genre *Capitella* ont en effet été longtemps considérées comme des espèces cosmopolites présentes du littoral aux plaines abyssales (Silva et al., 2016). Nos résultats s'accordent donc avec les travaux de Grassle and Grassle, 1976; Blake et al., 2009; Silva et al., 2016; Tomioka et al., 2016 ayant décrit plus d'une vingtaine d'espèce du genre *Capitella* sur des critères génétiques

De nombreuses espèces cryptiques associées à l'histoire climatique de l'Atlantique nord-est

La caractérisation de la diversité génétique des différentes 'espèces' de *Capitella spp.* à l'aide du *Cox-1* a permis de montrer qu'en Atlantique nord est, plusieurs clades peuvent être mis en évidence dans l'espèce dénommée *C. capitata.* Les 3 lignées mitochondriales (Cc-Manche1, Cc-Atlantique et Cc- Manche2) présentent entre elles une divergence allant de 2% (Cc-Manche1/Cc-Atlantique) à 14% (avec Cc-Manche2). De telles divergences sont

habituellement retrouvées entre espèces cryptiques chez les polychètes (Jolly et al., 2006; Nygren, 2014; Tomioka et al., 2016). Dans la plupart des cas, les descriptions d'espèces sœurs de *Capitella spp*. ont été réalisées à partir d'échantillons issus de régions géographiques distinctes. Seuls Grassle & Grassle (1976) présentent des données génétiques démontrant la présence d'espèces sympatriques chez *C. teleta*. La présence d'espèces cryptiques est cependant souvent rapportée dans le cadre de structure géographique complexe ou les espèces se superposent dans des zones de contact secondaire (Pérez-Portela et al., 2013) voire sous l'action d'une dispersion orientée liée à l'activité anthropique (pêche, aquaculture, transport…) pouvant rendre l'histoire plus complexe. En Atlantique Nord-Est, une explication à la présence d'espèces cryptiques en sympatrie montrant des niveaux de divergence de plus de 15% au niveau du mitochondrial met l'accent sur la recolonisation par des espèces du Pacifique il y a quelques 3.5 Ma lors de l'ouverture du détroit de Bering (Vermeij, 1991). Jolly et al., 2006 ont retracé la phylogéographie de deux espèces de polychètes de sables envasés le long des côtes européennes d'Atlantique et nos données sont concordantes avec leurs résultats. Les patrons phylogéographiques décrits et la présence de deux clades divergents de 16% entre la Manche et l'Atlantique pour des espèces proches du genre *Pectinaria* et celles d'*Owenia fusiformis*, leur ont permis de conclure que ce niveau de divergence correspondrait à une séparation il y a quelques 3.5 Ma (période Mio-Pliocène). Ce temps de divergence correspondrait à une période pouvant coïncider avec l'ouverture du détroit de Béring et à l'introduction de migrants depuis le Pacifique nord ou encore à une période plus lointaine (5 Ma) représentant également une période de forts changements climatiques associée à des mouvements tectoniques (Brault et al., 2004). La divergence de 14% des lignées Cc-Manche1/Cc-Atlantique avec la lignée Cc-Manche2 pourrait ainsi être expliquée par cette histoire d'ouverture entre Pacifique et Atlantique, période pendant laquelle aucun corridor marin n'assurait la communication entre Atlantique et mer du Nord. L'étude réalisée par (Muths et al., 2006) sur l'ophiure *Acronida brachiata* à développement direct montre également deux écotypes divergents (intertidal versus subtidal) de 20%. Le fait que ces études montrent le même type de patrons de divergence chez les espèces littorales a permis à ces auteurs de conclure quant à des évènements de vicariance à la transition Mio-Pliocène plutôt qu'à de la spéciation écologique (Muths et al., 2010). Chez *Capitella spp*., on observe bien une divergence de 14% entre la lignée Atlantique et l'une des 2 lignées Manche (lignée

Manche2) mais la divergence entre les lignées Cc-Atlantique et Cc-Manche1 n'est que de 2%, laissant penser qu'un deuxième événement climatique plus récent, comme le dernier maximum glaciaire (10 000 ans) ai pu laisser également une empreinte à la différenciation génétique de ce complexe lors de la recolonisation des côtes européennes à partir des zones refuges glaciaires (Maggs et al., 2008). Il est en effet possible que la recolonisation des côtes suite au LGM ait été beaucoup plus rapides dans le cas des polychètes *Owenia* et *Pectinaria* à larves pélagiques que dans le cas de *Capitella spp.* ou le développement semble direct. Il faut néanmoins nuancer ce propos car le développement des *Capitella* d'Europe peut être direct, incubant ou pélagique (Méndez et al., 2000). Néanmoins, Muths et al. (2006) en plus de montrer l'existence de deux écotypes divergents de 20% mettent également en évidence un isolement géographique Manche/Atlantique avec des lignées divergentes de 1% (et des valeurs de D de Tajima significativement négatives ainsi que des réseaux en forme d'étoiles). Ces auteurs en concluent que cela pourrait provenir, sous hypothèse de neutralité, à la colonisation de nouveaux territoires après le dernier maximum glaciaire et /ou suite à la formation de la Manche (10-8Ka).

Expansion récente des *Capitella spp.* ?

Au sein de toutes les populations, la diversité génétique globale (π et Hd) est faible jusqu'à être nulle pour l'espèce Cc-Manche2 (un haplotype majoritaire retrouvé de Dunkerque à Roscoff) avec des valeurs des tests de Tajima et Fu & Li significativement négatifs qui pourrait indiquer soit un début d'expansion après un goulot d'étranglement récent des populations de l'espèce, soit que le génome mitochondrial a balayé suite à un épisode de sélection directionelle. Chez les deux phylogroupes Cc-Manche1 et Cc-Atlantique qui ont une histoire séparée récente, la forme des réseaux pour chaque lignée est aussi en étoile avec au moins un haplotype majoritaire et un fort pourcentage de singletons. Ceci suggère également une expansion récente des populations (sous l'hypothèse de neutralité). Les courbes de mésappariements s'accordent avec un coalescent neutre pour la lignée Cc-Manche1 sauf au sein de la population Dunkerque qui montre un excès de mutations rares (bottleneck secondaire récent ?). Pour la lignée Atlantique, on observe le même effet, probablement atténué par la structure géographique sous-jacente, la population de Roscoff étant affectée mais pas celle de Port-la-Forêt qui montre un même patron de population stable tel que celui décrit pour la lignée Cc-Manche1 à Dunkerque.

De plus, il a été montré une faible divergence entre les deux phylogroupes Cc-Manche1 et Cc-Atlantique (2% environ) suggérant une divergence récente entre ces lignées mitochondriales. Ceci pourrait correspondre à une histoire glaciaire récente : celle du Last Glacial maximum (LGM, 10 000-21 000 ans) ou l'épisode glaciaire précédent (il y a 500 000 ans). Ces données s'accordent avec les études phylogéographiques réalisés sur les taxons marins le long des côtes de l'Atlantique nord-est qui sont actuellement en expansion depuis plusieurs refuges glaciaires du dernier maximum glaciaire. Parmi ceux-ci, la péninsule ibérique, les azores au Sud et la mer d'Irlande et d'autres mers plus boréales (Islande, Nord de la Norvège) sont rapportés (Jolly et al., 2006; Maggs et al., 2008; Provan and Bennett, 2008). Les variations climatiques complexes qui ont pu avoir lieu du Pléistocène, entre autres, les oscillations du niveau de la mer et la diminution de la température de surface l'été de 5-6°C, sont documentées comme ayant pu avoir un impact majeur sur la subdivision/différenciation et la variation en taille des populations chez beaucoup d'espèces marines côtières (Dynesius and Jansson, 2000; Hewitt, 2000).

Du point de vue de la structure génétique de chaque lignée mitochondriale, il n'y a pas de différenciation géographique chez l'espèce Cc-Manche1 bien que la comparaison des populations Boulogne avec celles de Roscoff/Port-la-Forêt montre des différences marginalement significatives. Pour l'espèce Cc-Atlantique, une différenciation génétique des populations est beaucoup plus marquée entre la Manche et l'Atlantique avec un haplotype quasiment fixée à Port la Forêt et un haplotype divergent d'une mutation retrouvée uniquement à Roscoff. Ceci semblerait montrer que, après une histoire commune liée à un évènement démographique récent, les lignées ont désormais leur histoire propre de recolonisation de part et d'autre de la pointe de Bretagne.

Pour reprendre l'exemple de (Muths et al., 2006) dans un contexte de lignées hautement divergentes au niveau du mitochondrial, une phase larvaire courte (non planctonique et inférieure à 4 jours) restreint suffisamment les flux géniques entre populations « voisines » (d'une baie à l'autre entre Océan Atlantique, Manche et Mer d'Irlande) pour ne pas générer de barrière spécifique au niveau des principales barrières agissant sur les espèces hautement dispersives telles que le front d'Ouessant. Pour des espèces dispersant de proche en proche, comme suspecté pour *Capitella spp.*, une possibilité est que les espèces puissent avoir divergé suffisamment vite pour avoir accumulé

suffisamment d'incompatibilités génétiques et créer des barrières génétiques dans les zones de contacts secondaires. L'une des principales interrogations est ensuite de déterminer comment et à quelle fréquence des lignées évolutives distinctes sont encore capables de s'hybrider lors des remises en contact lorsque les divergences mitochondriales sont élevées. Lorsque des espèces proches sont capables de s'hybrider localement, la colonisation de nouveaux habitats peut conduire à une introgression adaptative d'allèles venant de l'un ou l'autre des 2 génomes en présence selon que ces allèles sont favorisés dans l'habitat colonisé.

Trois lignées mitochondriales – 3 clades nucléaires de la préprocapitellacine : une évolution parallèle ou des évolutions de gène différentes ?

Les données de diversité génétique sur la préprocapitellacine montrent que trois clades peuvent être mis en évidence au sein de *Capitella spp.* Ces clades pourraient être homologues des trois clades mitochondriaux décrits précédemment pour ce complexe d'espèces. Alors que le clade C est présent presque uniquement chez la lignée mitochondriale Cc-Manche2 la plus divergente, les deux autres clades (A et B) sont partagés entre les phylogroupes Cc-Manche1 et Cc-Atlantique. Le fait que ces 3 clades nucléaires présentent entre eux une forte divergence comprise de 7-10% selon les régions du gène considérées entre A et B, et entre 10 à 13% entre le clade C et ces 2 derniers clades permet de suggérer que ces trois clades pourraient correspondre au fond génétique des trois lignées mitochondriales décrites précédemment et non à une duplication ancestrale antérieure aux événements de spéciation. Néanmoins, alors que le clade C correspond bien au phylogroupe Cc-Manche2, les deux phylogroupes Cc-Manche1 et Cc-Atlantique se partagent un pool d'allèles en commun dans les clades A et B. De plus, chaque individu présente généralement plus de 2 allèles pour le locus de la capitellacine, laissant supposer à l'existence de duplications en tandem, comme dans le cas de l'alvinellacine (cf. chapitre 2). Le nombre de copies du gène varie en effet de 2 à 8 par individu. Cette distribution des allèles au sein des individus et les divergences observées entre clades nous permettent d'émettre 2 scénarios évolutifs dans la diversification de ce gène chez *Capitella spp*.

Dans le premier scénario évolutif étudié, les clades A,B et C correspondent bien à des événements de spéciation ayant conduit aux lignées Cc-Manche1, Cc-Atlantique et Cc-

Manche2. On observe ensuite une diversification récente du gène au sein de chaque espèce par des duplications en tandem qui permettent à la fois d'expliquer les faibles divergences observées entre les copies de chaque clade et le polymorphisme du nombre de copies entre individus. Ce scénario permet aussi d'expliquer pourquoi la structure géographique de la différenciation des populations diffère entre le clade A et le clade B de la préprocapitellacine. Néanmoins, en partant du postulat de base que les clades correspondent donc bien au fond génétique des lignées mitochondriales, le fait que 100% des individus des lignées Cc-Manche1 et Cc-Atlantique possèdent un allèle du clade A et 1 allèle du clade B en statut hétérozygote laisse supposer une introgression massive des allèles de ces 2 clades sur l'ensemble de l'aire géographique entre lignées mitochondriales avec un avantage aux « hétérozygotes » (clade A et clade B) après remise en contact des deux phylogroupes, laissant supposer à une introgression adaptative capable de traverser rapidement la barrière génétique entre les espèces (Barton, 1979; Barton and Hewitt, 1989). Un tel scénario semble difficilement conciliable avec le mode de développement du ver et la colonisation de proche en proche de l'environnement intertidal côtier. En effet, les modèles théoriques postulent que la stabilité et l'étroitesse d'une zone de tension dépend d'un état d'équilibre entre la sélection contre les hybrides et les capacités dispersives des types parentaux (Barton and Hewitt, 1985). L'attendu d'une introgression d'allèles après un événement d'hybridation suite à une zone de contact secondaire est donc d'observer une zone de tension étroite entre les génotypes parentaux relativement localisée dans laquelle les hybrides sont trouvés (Barton and Hewitt, 1985, 1989) et une queue d'introgression plus ou moins directionnelle selon l'orientation des croisements dans la zone de contact (Harrison and Larson, 2014; Parsons et al., 1993). Il est donc assez difficile (voire impossible) de trouver un allèle du clade A et un allèle du clade B dans tous les individus. De plus, les individus peuvent également montrer un statut homozygote/hétérozygote pour chaque clade, laissant supposer que les clades A et B sont bien des locus distincts. Le scénario d'une duplication ancestrale à toutes les lignées mitochondriales décrites ne tient pas non plus puisque la plupart des individus Cc-Manche2 présentent uniquement des allèles du clade C, même si la encore un polymorphisme du nombre de copies peut être mis en évidence ainsi qu'un niveau d'introgression non nul bien que faible avec les autres lignées mitochondriales.

Un deuxième scénario semble plus pertinent pour expliquer les données observées. Dans ce scénario 2, une duplication du gène postérieure à l'événement de spéciation donnant la lignée Cc-Manche2 mais antérieure à la séparation des lignées Cc-Manche1 et Cc-Atlantique permet à la fois d'expliquer les disparités de divergences trouvées entre les clades et les lignées mitochondriales et le nombre élevé de copies du gène trouvé dans les individus des lignées Cc-Manche1 et Cc-Manche2. En effet, la divergence trouvée entre les clades A et B est largement supérieure (7-8%) à la divergence observée entre les lignées Cc-Manche1 et Cc-Atlantique (2%) alors même que le génome mitochondrial évolue généralement plus vite que le génome nucléaire (Brown et al., 1979; Birky et al., 1983). Ce scénario permet aussi d'expliquer pourquoi l'on observe des clades géographiques B1/B2 versus B3/B4 au sein du clade B avec une divergence approchant 1% qui pourrait effectivement correspondre à la séparation initiale des lignées Cc-Manche1 et Cc-Atlantique. En effet, l'étude de la structuration géographique des allèles pour le clade B observée à partir des réseaux montre une prédominance des allèles B3 et B4 (ce dernier étant exclusif de Port-la-Forêt) en Bretagne sud et une prédominance des allèles B2 à Boulogne/Dunkerque (bien que certains allèles soient également retrouvés à Roscoff). Par contre, ce type de structure attendue également dans le clade A (i.e. évolution parallèle des duplicats) n'est pas observé pour celui-ci et suggère que des effets sélectifs (i.e. balayage) aient pu également jouer dans l'évolution des structures génétiques des différentes espèces.

Contrairement aux barrières physiques aux flux de gènes qui sont totalement hermétiques pour l'ensemble des gènes, les barrières génétiques sont souvent « semi-perméables » et vont permettre des échanges génétiques sous conditions dans certaines zones du génome (Harrison, 1990, 1993). Dans notre cas, il semble d'une barrière génétique semi-perméable entre les lignées Cc-Manche1 et Cc-Atlantique mais également, dans une moindre mesure, avec Cc-Manche2 puisse exister, celle-ci étant préférentiellement centrée au niveau de Roscoff, là où tous les types d'allèles sont rencontrés. Barton and Hewitt, 1989 définissent les zones d'hybridation comme des zones dans lesquelles des individus hybrides sont produits, issus de la reproduction de deux entités génétiquement différenciées. Le plus souvent ces zones sont issues d'un contact secondaire apparus suite à l'isolement de populations par une barrière physique (par exemple des populations confinées dans des zones de refuge pendant les glaciations (Hewitt, 2004; Maggs et al., 2008)). Une zone

d'hybridation se caractérise par la création d'un gradient de fréquences alléliques (clines) qui résulte de l'équilibre entre deux forces : la migration et contre sélection des hybrides pour expliquer la persistance de cette zone hybride dans le temps et l'espace. En effet, alors que la migration aura tendance à homogénéiser les fréquences alléliques entre populations, la sélection contre les hybrides va quant à elle privilégier les génotypes parentaux. Ces clines selon la théorie de Barton and Hewitt (1989), sont des sigmoïdes mais des clines moins marquées pourraient être retrouvées notamment en milieu marin avec des espèces ayant un fort potentiel de dispersion qui aurait tendance à aplatir la sigmoïde ou à créer des zones en mosaïque comme cela est documentées chez *Mytilus galloprovinciallis* et *M. edulis* (Bierne et al., 2003). Pour des espèces à faible pouvoir dispersif comme dans notre cas (mais voir Méndez et al. 2000 pour des stratégies de dispersion dépendantes du milieu chez *Capitella spp.*), un des attendus est alors d'observer une zone étroite de tension (front d'hybridation) entre espèces séparées en allopatrie entre régions géographiques. Il est également possible de retrouver (lorsque à la distribution des populations est parcellaire comme expliqué précédemment) des zones de transition tout autant distribuées en 'patch' comme cela a pu par exemple être rapporté en Manche chez les isopodes du genre *Jaera* qui ne possèdent pas que phase larvaire plantonique et donc une dispersion limitée (Ribardière et al., 2017).

Chez les *Capitella spp.*, il semblerait qu'une zone d'hybridation entre les deux espèces Nord Atlantique et Manche (avec création de zone de tension) ne puisse pas être décrite *sensu stricto* avec les seules données de la capitellacine en l'absence d'une définition claire des allèles parentaux. Dans le cas du clade C diagnostique de la lignée mitochondriale Cc-Manche2, cette description est plus facile puisque seuls quelques individus semblent être introgressés par des allèles des clades A et B. Néanmoins, ces individus introgressés sont présents aussi bien à Roscoff qu'à Dunkerque.

Structure génétique des populations

Bien que la distribution des différents haplotypes mitochondriaux montraient une nette dichotomie entre les 2 lignées Manche et la lignée Atlantique, une structuration génétique fine n'a pu être mise en évidence au sein de chaque lignée pour le locus mitochondrial sauf pour la lignée Atlantique et marginalement pour la lignée Manche1 entre Roscoff et Boulogne. Les calculs de différenciation génétique montrent des différences

significatives entre populations pour le gène de la capitellacine. Les indices de fixation Phi_{ST} et Fst ont été calculés sur les 2 régions du gène pour les clades A, B et C. Ceux-ci montrent des différences dans le degré de significativité des tests selon les deux régions du gène et l'indice testé. Ici, les résultats indiquent une différenciation génétique significative des populations échantillonnées pour chaque clade pris indépendamment. Pour les clades A et B, les patrons de différenciation peuvent rappeler les patrons de différenciation décrits au niveau du gène mitochondrial entre les deux phylogroupes Cc-Manche1 et Cc-Atlantique. Il apparaît donc que la structure géographique des clades A et B intègre la différenciation génétique associée à la répartition géographique des lignées Manche1 et Atlantique. Cette différenciation géographique donne du poids à l'hypothèse que les clades A et B pourraient ne pas correspondrent aux deux phylogroupes Cc-Manche1 et Cc-Atlantique mais plutôt qu'ils contiennent l'information de cette divergence au moins à travers la ségrégation des allèles B1/B2 et B3/B4.

Cette étude révèle également l'existence de recombinants dans la région 3' chez deux individus à Roscoff avec les mêmes points de recombinaison et qui ont pu acquérir leurs propres mutations et qui ont été caractérisés chez des individus de la lignée Cc-Atlantique et Cc-Manche2. Ces recombinants sont issus d'une recombinaison entre le clade C et le clade A. L'existence de recombinants inter-clades est une indication qu'un temps suffisamment long depuis la remise en contact des espèces existe pour que de tels évènements aient pu acquérir leurs propres mutations et soient retrouvés au sein de la population à une fréquence suffisante pour être échantillonnés. Puisque cet évènement de recombinaison a eu lieu entre le clade provenant de l'espèce la plus divergente, ceci indiquerait un avantage évolutif à la diversification par recombinaison puisque cet allèle n'a été ni contre-selectionné ni éliminé par dérive. Le fait que cet évènement soit documenté au sein de la région 3' corrobore des résultats trouvés par (Boon et al., 2009) sur la défensine MGD2 chez les moules du genre *Mytilus* pour laquelle un taux de recombinaison élevé est trouvé en région 3' (en amont de la région codant le peptide antimicrobien) qui, de surcroit, masque le signal de différenciation génétique retrouvé sur la région 5' du gène. De plus, il apparaît que les deux clades d'allèles B3 et B4 ont des positions intermédiaires dans l'arbre phylogénétique des allèles indiquant qu'il pourrait s'agir ici de recombinants entre les clades A et B1/B2. Ainsi, ce phénomène décrit presque exclusivement chez la lignée Cc-Atlantique et retrouvé

en majorité au sein de Roscoff (B3) et Port-la-Forêt (B3 et B4), serait apparu tôt dans l'histoire des espèces puisqu'ils ont eu suffisamment de temps pour acquérir leurs propres mutations.

Cependant, la lignée Cc-Atlantique montre une différenciation génétique entre Port-la-Forêt et Roscoff sur le gène mitochondrial laissant supposer à l'existence d'une barrière au flux de gènes à l'entrée de la Manche (front d'Ouessant). Ainsi, l'hypothèse de différenciation géographique avec des pools d'allèles distincts qui ont ensuite recombinés dans la zone de contact suite à de l'hybridation entre les deux lignées mitochondriales distinctes semblerait ici plausible.

Quel que ce soit le scénario évolutif (1 ou 2), l'analyse IMa2 révèle l'existence d'un flux de gènes non nul entre les lignées mitochondriales Cc-Manche1 et Cc-Atlantique depuis leur séparation pour le gène de la préprocapitellacine. Selon le scénario choisi, ce flux est symétrique entre les 2 lignées ou complètement asymétrique de la lignée Cc-Manche1 vers la lignée Cc-Atlantique. Le processus de spéciation entre Cc-Manche1 et Cc-Atlantique semble être purement allopatrique puisque les clades sont bien identifiés géographiquement, avec des flux très réduits entre localités, surtout de part et d'autre de la pointe de Bretagne. D'autres flux asymétriques et faibles peuvent être rapportés de Cc-Manche1 vers Cc-Manche2 et puisque les individus introgressés se retrouvent à Roscoff, cette zone géographique pourrait donc être une zone de contact secondaire majeure entre ces deux espèces. En termes de temps de divergence, malgré de forts écarts types, le scénario 1 semblerait indiquer que la divergence entre les deux populations Cc-Manche1 et Cc-Atlantique pourrait se situer autour du LGM alors que la séparation de ces lignées avec la lignée Cc-Manche2 pourrait être beaucoup plus ancienne (par exemple : glaciations du Pleistocene entre 100 et 500 kya: Ehlers and Gibbard, 2007). Le scénario 2 quant à lui indique des temps de divergence plus lointain avec une divergence entre Cc-Manche1 et Cc-Atlantique autour de 300000ans et 1.2Ma pour la divergence de leur espèce ancestrale avec celle de Cc-Manche2.

Un isolement écologique pourrait expliquer la séparation les deux phylogroupes avec l'espèce Cc-Manche2 puisque ces lignées sont sympatriques mais ne présentent que de rares cas d'introgression. Un mécanisme d'isolement pourrait donc exister ou serait en passe

d'exister entre ces lignées conduisant à un isolement quasi complet de ces lignées. Des études rapportent en effet qu'il existe une corrélation linéaire négative entre la distance génétique et les croisements entre lignées (Mallet, 2005, 2007) et (Muths et al., 2010) rapportent que très peu d'hybrides sont retrouvées entre des lignées divergentes de 20% au niveau du mitochondrial. Lorsque des espèces cryptiques se retrouvent en contact après isolement, il n'est en effet pas rare qu'un isolement écologique puisse être décrit avec des espèces présentant des préférences pour des microhabitats différents (Knowlton, 1993). De nombreux mécanismes pré-zygotiques peuvent alors expliquer cet isolement reproducteur (séparation des habitats, non synchronisme de la reproduction…).

Du point de vue populationnel, l'importance du mode de reproduction et plus particulièrement de la durée de la phase larvaire planctonique est souvent avancée comme mécanisme principal d'isolement pré-zygotique et de nombreuses études ont étudiés le lien entre le mode de reproduction, la dispersion effective et la structure génétique de populations (Goldson et al., 2001). Les espèces du genre *Capitella* sont gonochoriques et les embryons et étapes précoces larvaires se développent au sein d'un tube construit par les femelles autour de leur corps (Méndez et al., 2000). Les larves, qui ne peuvent pas se nourrir, émergent des tubes approximativement au bout de neuf jours et celles-ci se font leurs propres tunnels rapidement après leur métamorphose en juvéniles fouisseurs (Seaver, 2016). Le même type de reproduction est documenté chez *Hediste diversicolor* : les larves sont incubées par la femelle après fécondation et s'enfouissent directement après leur émergence du tube maternel (Bartels-Hardege and Zeeck, 1990; Einfeldt et al., 2014) et explique parfaitement les patrons de différenciation génétique observés à petites échelles spatiales le long de la côte Nord Adriatique (Virgilio and Abbiati, 2004; Virgilio et al., 2006). Ces études montrent également une forte hétérogénéité génétique des individus au sein même des estuaires suggérant une importante structure à micro-échelle (appelée chaotic genetic patchiness ; voir aussi Eldon et al., 2016 pour revue). A macro-échelle, un isolement de type « isolement par la distance » (attendu pour des espèces à faible capacité de dispersion) ne peut alors pas être mis en évidence puisque des processus agissant à échelle locale masquent des patrons de différenciation à large échelle (Hellberg, 2009; Muths et al., 2010). Dans le cas des *Capitella spp.* sur les côtes Françaises de la Manche et Nord Atlantique, il semblerait donc que la courte phase larvaire non sessile se traduise par un

isolement génétique des populations le long des côtes françaises mais que cela ne limite pas les phénomènes d'introgression alléliques pour le gène de la préprocapitellacine apparu tôt dans l'histoire évolutive des deux phylogroupes Cc-Manche1 et Cc-Atlantique et/ou favorisé par de la sélection positive.

Dynamique évolutive du précurseur protéique : comparaison inter-espèces.

Les comparaisons avec le Clade C, montrent en région 5' les plus fortes divergences avec 18% et 20% (ou 11 et 12% dans les régions exoniques) contre 16% (8.2% en régions exoniques) pour la comparaison Clade A-Clade B. Cette divergence est proche de celle trouvée entre les deux clades les plus divergents au niveau mitochondrial (18%). Ceci se traduit, le long du gène par des d_N/d_S relativement faibles (en moyenne 0,09) et par la présence de deux pics de divergence non synonyme au sein de la prorégion. Le clade C semble donc évoluer plutôt à travers un relâchement de la pression de selection le long du gène excepté au début de la prorégion. Le test de MacDonald et Kreitman corrobore ces observations en montrant des valeurs de π_N/π_S globalement supérieures aux valeurs de d_N/d_S (presque significatives en région 3' pour la comparaison Clade A/Clade C) pouvant suggérer l'action d'une sélection balancée sur le polymorphisme ou à une accumulation des mutations légèrement délétères dans le polymorphisme qui ne contribuent pas à la divergence des clades étudiés.

La comparaison des Clade A et Clade B quant à elle montre de nombreux pics de divergence non synonyme localisés dans la prorégion, le domaine BRICHOS et, dans une moindre mesure, la région du PAM. Pour ces clades, l'action d'une sélection diversifiante au sein de ces trois régions peut donc être suggérée et semble assez proche de celle trouvée entre les gènes paralogues de l'alvinellacine. Ces résultats pourraient donc accréditer l'idée que les clades A et B sont bien des locus paralogues et non des clades ancestraux d'allèles associés aux espèces, même si une sélection diversifiante par l'habitat puisse avoir eu lieu lors de la période d'isolement desdites espèces comme cela a été suggéré pour le clade C. Cette interprétation est confirmé par le test de MacDonald and Kreitman, 1991 qui montre des valeurs de d_N/d_S supérieures aux π_N/π_S (la valeur du test devient significative en n'utilisant que les clades B1/B2 versus clade A).

La distribution des mutations non-synonymes a été étudiée plus en détail pour le clade B puisque celui-ci montre des patrons de différenciation génétique entre les différentes localités géographiques au moins au sein d'un phylogroupe (Clade B3 et B4 au sein des individus de Port-la-Forêt du phylogroupe Cc-Atlantique). Les résultats indiquent que la divergence entre les deux principaux groupes d'allèles (B1/B2 et B3/B4 : BA vs BB) du clade B se ferait également à travers l'action d'une sélection diversifiante au sein de la prorégion et du BRICHOS. Ici, l'introgression et la recombinaison entre les lignées Cc-Manche1 et Cc-Atlantique peut également jouer un rôle pour permettre à certaines mutations avantageuses de se propager dans le polymorphisme des 2 lignées sous l'hypothèse que les sous-clades BA et BB représentent bien les 2 lignées géographiques. En effet, la valeur nulle de d_N/d_S entre les clades B pour la région du PAM peut s'expliquer par le fait que les différents variants du peptide antimicrobien *sensu stricto* sont échangés entre les différents sous-clades sans aucune barrière. Cette large distribution géographique des variants pourrait être favorisée par de l'introgression adaptative dans le cas où les sous-clades BA et BB représentent bien les lignées Cc-Manche1 et Cc-Atlantique (scénario 2) mais ne peuvent s'expliquer que par de la recombinaison/conversion inter-génique dans le cas ou les clades B1, B2, B3 et B4 représenteraient des copies du gène obtenus par duplication en tandem. Ceci est d'autant plus plausible que, en plus de l'existence d'un PAM recombiné entre clades, c'est en région 3' que deux séquences recombinantes ont été mises en évidence indiquant que cette région montrerait effectivement un taux de recombinaison plus élevé qu'en région 5'. Ceci pourrait également expliquer la plus faible divergence des clades dans la région 3' qui échangent/ont échangés de l'information génétique par recombinaison. L'existence d'un taux très élevé de recombinaison au sein d'un gène codant pour un peptide antimicrobien a été montrée dans le cas de la défensine MGD2 de Mytilidae notamment en région 3'. Ce gène montrerait également l'action d'une sélection positive (directe ou indirecte) agissant sur la région 3' proche de la région codant pour le PAM *sensu stricto* avec au moins un site qui pourrait être la cible de cette sélection (Boon et al., 2009).

Des recombinants ont été également décrits entre le clade A et le clade C et bien que n'étant pas retrouvés en fréquence élevée (mais notre effort d'échantillonnage est faible) dans la population, ces recombinants sont suffisamment anciens pour avoir accumulé leurs propres

mutations. Ceci constitue un argument supplémentaire pour proposer un avantage évolutif à la diversification du gène par recombinaison et/ou introgression entre lignées.

Au niveau du peptide signal, l'absence de variation génétique pourrait provenir d'un intérêt à ne pas modifier le mode d'adressage vers un compartiment conservé (réticulum endoplasmique : Zhang and Gallo, 2016) ainsi que le site de clivage de cette région (Martoglio and Dobberstein, 1998).

Dans la région du PAM, les valeurs de d_N/d_S sont supérieures à 1 seulement dans le cadre des comparaisons entre les clades A et B, A l'inverse, les comparaisons avec le clade C montrent des valeurs comprises entre 0 et 1 plutôt diagnostique d'une évolution neutre du PAM avec trois mutations synonymes fixées. Il apparaît néanmoins que, si l'on élimine les recombinants inter-clades, le PAM n'est vraiment polymorphe que pour le clade B (le A également mais dans une moindre mesure). Ainsi, l'évolution globalement neutre/légèrement conservée documentée pour le clade C (dupliqué selon les données de génotypage avec certains individus possédant allèles mais non polymorphe) diagnostique de la lignée mitochondriale la plus divergente (Cc-Manche2) s'oppose donc à la forte diversification du peptide antimicrobien (par différents mécanismes) entre et au sein des clades A et B. Du point de vue quantitatif, ceci se traduit effectivement par des individus possédant au moins deux et au maximum 3 peptides antimicrobiens pour les deux lignées Cc-Manche1 et Cc-Atlantique (1 PAM diagnostique de chaque clade). Les individus appartenant à l'espèce Cc-Manche2 quant à eux disposent tous d'un seul peptide antimicrobien diagnostique du clade C et avec 3 individus porteur d'un PAM diagnostique du clade A ou du clade B laissant penser à des événements épisodiques d'introgression d'allèles avec cette lignée divergente. Ainsi, un avantage sélectif à la conservation de la séquence codant le peptide antimicrobien pourrait caractériser la lignée Cc-Manche2 par comparaison avec la diversité allélique observée aux clades A et B, et pour lesquels un avantage à l'état hétérozygote (individus sont tous porteurs d'un allèle A et d'un allèle B) pourrait exister sous l'hypothèse que les clades A et B représentent bien les états ancestraux des lignées Cc-Manche1 et Cc-Atlantique : scénario 1).

Un des résultats de cette étude (bien que manquant de robustesse) concerne une éventuelle absence de relation entre l'arsenal des capitellacines des individus et le statut

« parasité » ou non des individus. De plus, des travaux ont également permis de mettre en évidence que les animaux associés placés en environnement décontaminé, sans sulfure et aseptisé, mourraient en moyenne plus rapidement que ceux non associé (Aurélie Tasiemski, *comm pers*). Finalement, il a été montré que la capitellacine s'accumule dans le mucus produit par les cellules tégumentaires indiquant une sécrétion du PAM dans le milieu extracellulaire (milieu environnant du ver dans le cas des cellules tégumentaires), où il pourrait donc exercer son activité antimicrobienne et ainsi jouer un rôle dans l'immunité externe de l'hôte. Ainsi, puisqu'il n'est pas à exclure que toute la diversité génétique n'a pas été recapturée pour le clade C de la capitellacine (quasi monomorphe au niveau du PAM), le niveau de polymorphisme documenté dans la région du PAM chez les espèces *Capitella spp*. pourrait provenir d'une fonction qui ne serait pas, comme pour *Alvinella*, pour le maintien d'un consortium bactérien (qui a un rôle crucial pour l'organisme) stable dans le temps, mais plutôt traduirait de l'avantage à diversifier l'arsenal immunitaire pour la défense de l'organisme (fonction de défense donc). Ceci est d'autant plus probable qu'il a été montré que ces organismes sont opportunistes et les populations se déplacent dans le sédiment faisant que ces organismes sont forcément en contact avec une large gamme d'environnement microbien fluctuants dans le temps et l'espace. Effectivement, au sein des phylogroupes Cc-Manche1 et Cc-Atlantique, les mécanismes de création de la diversité du peptide antimicrobien se traduisent *in fine* par une augmentation du nombre de copies du gène dans le génome des individus. L'avantage à augmenter son arsenal immunitaire pourrait alors permettre de diversifier la réponse antimicrobienne générale (au niveau des cibles : Yang et al., 2006), voire générer des effets synergiques. Une évolution par duplications peut alors être extrêmement bénéfique dans un environnement dans lequel les pathogènes évoluent rapidement en permettant à la population d'être plus flexible à des changements spatio-temporels des communautés microbiennes (Du Pasquier, 2006). De plus, il a pu être montré chez l'espèce Cc-Manche2 que les évènements d'introgression n'étaient pas contre sélectionnés puisque retrouvés en fréquence non négligeable dans cette étude indiquant bien cette avantage à la diversification de l'effecteur immunitaire.

Chez *Capitella spp.*, il a quand été montré une plus forte survie et un taux de croissance plus élevé dans des environnements riches en sulfures en se nourrissant de bactéries chimiosynthétiques (Tsutsumi et al., 2001). Ces auteurs ont également mis en

évidence que la disponibilité en matières organiques produites à partir de la chimiosynthèse des bactéries du milieu peut également constituer un facteur important dans le contrôle de la distribution des espèces de *Capitella spp*. et de la dynamique de leurs populations. Dans le cas des *Capitella spp.* de Manche (pour toutes les lignées mitochondriales), une proportion (bien que faible et fluctuante en fonction des saisons) de la population est retrouvée en association facultative avec des bactéries sulfo-oxydantes suggérant que cette association pourrait être coûteuse pour l'hôte ce qui a été montré puisque les individus en association meurent plus rapidement que des individus non parasité. Ainsi à partir de ces observations, la présence des bactéries dans l'environnement immédiat des vers pourrait fournir des avantages aux *Capitella spp.* qu'ils soient nutritif ou de protection contre des pathogènes de l'environnement. Cependant, il est possible que selon les conditions environnementales, cette association devienne délétère pour l'hôte qui se retrouve alors parasité par ces bactéries sulfo-oxydante au moins transitoirement réduisant sa fitness (par un cout du à la réponse immunologique par exemple). L'étude effectuée par Lolita Roisin (en M2 au laboratoire) a permis de montrer que la capitellacine était active vis-à-vis de *Thiomargarita* (en immunocytochmie et microscopie électronique) à plus forte concentration que contre les bactéries côtières. Ainsi, l'effecteur immunitaire pourrait avoir évolué vers un état plus « permissif » quand à la bactérie sulfo-oxydante tout en se diversifiant pour répondre à la pression de sélection de l'environnement (a)biotique dans lequel les espèces ont évoluées.

CONCLUSION

Cette étude révèle une histoire évolutive complexe au sein d'un complexe d'espèces cryptiques retrouvées en sympatrie le long des côtes de la Manche. L'évolution des clades de la capitellacine pourrait faire l'objet d'une sélection diversifiante qui agirait sur les différents paralogues (clades) qui s'influencent par recombinaison/conversion génique (avec existence de PAMs recombinés) mais aussi/ou à travers de l'introgression adaptative entre les différentes lignées mitochondriales de la côte française. Le maintien d'une telle diversité pourrait s'expliquer par de la sélection balancée dans le temps et l'espace favorisé par l'introgression et dynamique recombinatoire des allèles/duplicats. Le niveau de diversité génétique retrouvé dans la région du PAM traduirait bien cette fois-ci plus un attendu retrouvé dans la littérature à savoir la diversification de l'effecteur immunitaire en tant que tel par duplication/recombinaison et/ou dynamique d'introgression (adaptative). Le scénario

évolutif étudié numéro2 apparait comme étant le plus probable (duplication ancestrale aux deux phylogroupes Cc-Manche1 et Cc-Atlantique) en donnant des tailles de populations et des dates de divergences entre espèces qui paraissent mieux estimées (divergence entre Cc-Manche1 et Cc-Manche2 notamment) et qui expliquerait, entre autre, que tous les individus étudiés possèdent au moins un allèle du clade A et un du clade B (pas de queue d'introgression détectée). Ce scénario s'ajusterait à l'idée d'une espèce qui en recolonisant l'espace après évènement géographique rencontre et s'hybride avec les espèces locales.

Chapitre 4 : Discussion générale et perspectives

Certains invertébrés marins sont régulièrement colonisés par des bactéries ectosymbiotiques/endosymbiotiques, que ce soit au sein des écosystèmes entièrement basés sur la chimiosynthèse (sources hydrothermales, zones de suintements froids, carcasses) ou les écosystèmes côtiers réduits (mangroves, fonds envasés) pour lesquels de telles associations peuvent aussi être documentées (Ott, 1996). En effet, ces associations métazoaire/bactéries dominent les communautés associées aux sources hydrothermales et sont à la base d'une production primaire qui alimente de nombreux consommateurs secondaires : c'est l'un des écosystème le plus productive sur notre planète (Van Dover, 2000). Ces interactions peuvent également être retrouvées au sein d'écosystèmes peu profonds (entre 0 et 200 m) pour lesquelles la production de matière organique est largement dominée par la phototrophie : ces symbioses chimio-synthétiques y sont non dominantes (Bright and Giere, 2005). Ces bactéries chimiolithoautotrophes utilisent des composés comme l'hydrogène, l'hydrogène sulfuré ou encore le méthane ou les nitrates comme source d'énergie et c'est l'ATP apporté par l'oxydation de ces composés réduits qui va permettre de fixer le dioxyde de carbone et/ou le méthane contenu dans les fluides diluées en matière organique qui va ensuite fournir une source de nutrition aux hôtes par transfert de métabolites et/ou digestion des bactériocytes. Au sein des environnements hydrothermaux, ces associations ont pu être décrites, de façon non exhaustive, chez les bivalves, les gastéropodes, et quelques espèces de crevettes, polychètes (notamment les sibloglinidae) et cirripèdes (Dubilier et al., 2008 pour review). Au sein des environnements côtiers, ces associations ont plutôt été recensées chez des bivalves, des nématodes, des annélides polychètes et oligochètes et chez des ciliés coloniaux (Dubilier et al., 2008; Goffredi, 2010). Les annélides sont retrouvés dans tous ces écosystèmes réduits où ont été reportées des associations avec des bactéries chimioautotrophes et chez ces organismes, la relation évolue d'une relation occasionnelle à des cas d'endosymbioses obligatoires avec incorporation des symbiontes dans des cellules/organes spécialisés (Tableau1) en passant par des ectosymbioses obligatoires.

Classe/famille	Hôte	Localisation du/des symbiontes	Habitat	Type de symbiote
Polychaeta Siboglinidae	*Riftia Lamellibrachia Escarpia*	Intracellulaire, Trophosome	Profond (hydrothermal, bois coulés…)	Chimio-lithotrophe capable d'oxyder les sulfures
Polychaeta Siboglinidae	*Sclerolinum*	Intracellulaire, Trophosome	Profond (hydrothermal, bois coulés…)	Chimio-lithotrophe capable d'oxyder les sulfures
Polychaeta Siboglinidae	*Siboglinum Oligobrachia*	Intracellulaire, Trophosome	Profond (suintements, bois coulés…)	Chimio-lithotrophe capable d'oxyder les sulfures et le méthane
Polychaeta Terebellidae	*Alvinella*	Epibiotique	Hydrothermal	Chimio-autotrophe capable d'oxyder les sulfures
Clitellata Phallodrilinae	*Inanidrilus Olavius*	Extracellulaire, Subcuticulaire	Côtier	Chimio-lithotrophe capable d'oxyder les sulfures et les sulfates
Clitellata Tubificinae	*Tubificoides*	Epibiotique	Côtier	Chimio-lithotrophe capable d'oxyder les sulfures et les sulfates

Tableau 1. Liste –non exhaustive– de relations endosymbiotiques/ectosymbiotiques rencontrés chez les annélides inféodés à différents habitats réduits (de Dubilier et al., 2008).

Les ectosymbioses (qui peuvent être décrites comme l'interaction d'un (micro)organisme avec substrat biotique sur lequel il est attaché quel que soit la nature de la relation : mutualisme, commensalisme, parasitisme), sans être majoritaires, sont communément observées dans les écosystèmes marins (Key et al., 1996). Elles diffèrent des endosymbioses à de nombreux égards. Les micro-organismes sont en effet directement exposés aux conditions ambiantes dans lesquelles l'hôte évolue. La spécificité entre les deux partenaires a une origine multifactorielle mais il existe des exigences générales qui peuvent réduire drastiquement le nombre de micro-organismes capable d'établir un lien avec l'hôte. Ainsi, la capacité à se multiplier plus rapidement que les autres micro-organismes aux conditions environnementales (température, potentiel redox, pH, osmolarité …) dans lesquelles vit l'hôte peut représenter une des exigences physiologiques pré-requises à l'interaction. De plus, des substances produites par l'hôte peuvent aussi modifier l'environnement immédiat ainsi que la physiologie des micro-organismes. Au sein de ces substances sécrétées par l'hôte, certaines sont impliquées dans la défense et/ou le contrôle de la prolifération microbienne sur le tégument de l'hôte et définissent l'immunité externe. Celle-ci, "*the*

underappreciate selective force in the evolution of the immune system (Otti et al., 2014) ", peut se décrire par n'importe quel trait héritable qui agit à l'extérieur de l'organisme et augmente la protection de l'organisme face aux pathogènes et/ou module la composition de la communauté microbienne vivant à la surface de l'hôte (Otti et al., 2014). Ainsi peuvent être considérés comme défense externe à la fois les sécrétions ayant une activité antimicrobienne (lysozyme, alkaloïdes, acide lactique, peptides antimicrobiens …) mais aussi les comportements qui affectent la distribution des micro-organismes sur l'hôte (toilettage). Par exemple, il a été observé chez *Nicrophorus vespilloide* (insecte nécrophage) que les adultes déposent des sécrétions contenant du lysozyme sur les carcasses de petits vertébrés dans lesquels ils pondent pour que leur descendance se développe : la survie des larves est alors significativement augmentée (Arce et al., 2012). Chez la termite, il a été montré que deux PAMs ciblent spécifiquement les champignons pathogènes avant qu'ils n'entrent dans la cuticule de l'hôte (Bulmer et al., 2009, 2010). Le rôle protecteur de l'immunité externe chez ces termites a ensuite pu être démontré grâce à l'utilisation d'une molécule qui bloque l'activité d'un des deux PAMs, cette action ayant pour conséquence de rendre ces organismes significativement plus sensibles aux infections. Ces effecteurs immunitaires sont des composants clés du système immunitaire des eucaryotes qui éradiquent rapidement une large gamme d'agents pathogènes (virus, bactéries, champignons : Zasloff, 2002) de l'extérieur. Ceux-ci sont également documentés comme pouvant façonner/contrôler/confiner la microflore symbiotique dans des compartiments anatomiques spécifiques (intestin, trophosome, peau) et contribuent ainsi au maintien/mise en place de symbioses chez les vertébrés et invertébrés (Franzenburg et al., 2013; Login et al., 2011; Tasiemski et al., 2015). L'évolution des peptides antimicrobiens est largement documentée comme étant façonnée par des évènements de duplication avec action de la sélection positive pour créer de la diversité fonctionnelle permettant aux organismes d'affronter de nouveaux pathogènes issus d'environnements/habitats nouveaux (variation spatio-temporelle) et/ou de pathogènes ayant évolués rapidement/ evoluant constamment pour échapper à la réponse immunitaire de l'hôte (Tennessen, 2005; Unckless et al., 2016).

Cette thèse s'est proposée de décrire les patrons de diversité génétique d'une famille de PAM très particulière, présente uniquement chez certaines familles de polychètes et caractérisée par une association entre le peptide anti-microbien lui-même et une protéine

chaperonne : le domaine BRICHOS. Le chapitre 2 a consisté à étudier la diversité génétique du précurseur protéique de l'alvinellacine chez l'annélide hydrothermale *A. pompejana* entre 2 formes cryptiques séparées par une barrière géographique documentée comme étant semi-perméable aux flux de gènes (Plouviez et al., 2010). Dans le cas du chapitre 3, la diversité génétique du précurseur protéique de la capitellacine (orthologue de l'alvinellacine) a été étudiée dans un contexte d'hybridation entre des espèces cryptiques dont deux lignées mitochondriales (appelées ci-après phylogroupe) présentent 2% de divergence au niveau du gène mitochondrial *cox-1*.

Au cours de la thèse, nous nous sommes plus particulièrement intéressés à l'histoire de cette diversification et au rôle de la sélection dans la mise en place et le maintien de cette diversité génétique chez deux précurseurs protéiques de peptides antimicrobiens orthologues. Ces deux effecteurs immunitaires ont été décrits chez deux annélides marins inféodées à des environnements contraignants montrant chacun une association symbiotique obligatoire (*Alvinella pompejana*) versus facultative (*Capitella spp.*). En effet, alors que le polychète hydrothermal *A. pompejana* présente une épibiose très diversifiée et obligatoire laissant présager d'une acquisition relativement ancienne de l'interaction, les annélides du complexe d'espèces *Capitella spp.* sont caractérisées par une épibiose facultative avec une seule souche microbienne (*Thiomargarita* sp.) n'affectant qu'un petit nombre d'individus dans des conditions d'hypoxie très particulières (zones portuaires très envasées). Ces 2 études effectuées en parallèle nous permettent de dresser une première liste de comparaisons et de tirer quelques conclusions générales sur l'évolution de cette famille particulière de PAMs à BRICHOS.

1. Comparaison des principaux résultats sur les processus évolutifs ayant façonné la diversité génétique des deux précurseurs protéiques de peptide antimicrobien.

1.1. Patrons de diversité et pression de sélection sur la région codant le peptide antimicrobien.

Chez *Alvinella pompejana*, l'action d'une sélection purifiante très forte sur la région codant de l'alvinellacine a pu être décrite. En effet, bien que le précurseur protéique soit codé par une famille multigénique ayant une longue histoire évolutive de duplications, le PAM est strictement monomorphe. Cette observation s'oppose à celle réalisée sur la capitellacine qui

présente un très fort niveau de polymorphisme non-synonyme et des mutations non-synonymes fixées entre les lignées mitochondriales décrites. Seule l'espèce Cc-Manche2 possède un PAM monomorphe même si le prépropeptide a dupliqué au moins une fois dans l'espèce. Il apparaît cependant que cette espèce est la moins bien échantillonnée des 3 lignées mitochondriales trouvées et l'on ne peut exclure de trouver de nouveaux variants avec un effort d'échantillonnage plus important. Cette différence importante entre alvinellacine et capitellacine pourrait être due, dans un contexte de contrôle et sélection des épibiontes, à une évolution extrêmement lente du consortium microbien associé au milieu hydrothermal profond. En effet, ces communautés bactériennes hautement spécialisées sont très anciennes et assez peu diversifiées sur les parois des cheminées hydrothermales (3 lignées prédominantes archéoglobales, thermococcales, methanococcales), par rapport aux communautés microbiennes d'autres environnements marins (Schrenk et al., 2003; Kormas et al., 2006). L'orthologue étudié chez l'espèce sœur du point de vue phylogénétique, *A. caudata*, ne diverge de l'alvinellacine que d'un seul acide aminé malgré une histoire évolutive longue d'au moins 20 millions d'années (événement de spéciation des 2 espèces d'*Alvinella* semble pré-dater l'événement de subduction de la plaque Farallon sous la plaque américaine, il y a 25 Ma (Little and Vrijenhoek, 2003). L'hypothèse d'une sélection purifiante extrêmement forte sur cet antibiotique est d'autant plus plausible puisque *A. caudata* évolue dans le même type d'habitat (ce sont deux des espèces syntopiques) avec un consortium microbien dont au moins un phylotype dominant a été montré commun aux deux espèces. On peut donc raisonnablement penser que ce PAM a évolué vers une spécialisation maximale vis-à-vis de son consortium d'épibiontes. Dans le cas des polychètes du genre *Capitella*, cette spécificité est nécessairement faible puisque l'épibiose avec *Thiomargarita* sp. est facultative et n'affecte au maximum qu'environ 20% de la population dans des conditions particulières (vases très anoxiques des ports). Il semble donc que la fonction première du peptide soit toujours une fonction de défense vis-à-vis des pathogènes qui sont eux beaucoup plus diversifiés du point de vue phylogénétique dans le milieu sédimentaire côtier et/ou peuvent évoluer dans le temps et l'espace (Lozupone and Knight, 2007). De plus, les espèces de *Capitella* sont d'une nature très opportuniste et colonisent des habitats meubles relativement variés présentant donc une grande diversité d'espèces microbiennes qui peut représenter une force de sélection sur le PAM entrainant ce niveau de diversité génétique.

1.2. Une diversification génique liée à de la duplication et de la recombinaison inter-génique.

Que ce soit l'alvinellacine ou la capitellacine, on constate que les 2 gènes ont évolué sous l'action de plusieurs événements de duplication, la plupart des copies produites étant sans doute la conséquence d'une duplication en tandem. Ce mode de diversification pourrait être associé à l'acquisition de l'épibiose chez les alvinellidae puisque les événements de duplication apparaissent plus récents que l'événement de spéciation ayant conduit aux 2 espèces *A. pompejana* et *A. caudata*. Néanmoins, une évolution du gène par des duplications successives plutôt récentes semble également prévaloir chez les espèces de *Capitella*, chez qui une interaction avec *Thiomargarita* n'est que peu fréquente, laissant supposer que l'acquisition de cette symbiose n'est pas l'élément déclencheur à cette différenciation génique. Chez les deux lignées mitochondriales les plus proches (Cc-Manche1 et Cc-Atlantique) tous les individus possèdent au moins deux variants du peptide antimicrobien (l'un du clade A et l'autre du clade B) laissant suggérer, que contrairement aux *Alvinella*, cette duplication serait antérieure à l'événement de spéciation datée au minimum du dernier maximum glaciaire, et donc largement antérieure à la mise en place de zones portuaires polluées. Cette diversification génique de la capitellacine et le maintien des duplicats dans les 2 espèces suggèrent donc soit une réponse des vers vis-à-vis d'une diversification des communautés microbiennes rencontrées (e.g. variations saisonnières de la diversité microbienne), soit que l'acquisition de cette épibiose est beaucoup plus ancienne qu'une simple adaptation aux activités polluantes humaines. De plus, plusieurs évènements de recombinaison inter-génique ont été observés entre les locus dupliqués de l'alvinellacine et de la capitellacine. Le fait que la plupart des recombinants ait un nombre de mutations qui leur sont propres permet également de suggérer qu'ils sont apparus relativement tôt après les différentes duplications, et donc que certaines remises en contact des espèces soient plus anciennes qu'actuellement supposées (e.g. recombinants entre les clades A et C). Ces événements de recombinaison sont donc fréquents dans l'évolution du précurseur protéique de ces PAM et un bon moteur pour créer de la diversité génétique en offrant de nouvelles combinaisons qui seront testées par la sélection naturelle (ou pour permettre à certaines mutations avantageuses de se propager dans le polymorphisme).

S'il a pu être détecté sur les gènes de l'alvinellacine qu'un nombre d'allèles qui peut excéder 2 pour certains paralogues chez quelques individus (notamment pour le paralogue 5 le plus divergent), la technique d'acquisition des données n'a pas permis de conclure de façon certaine à l'existence d'un polymorphisme du nombre de copies. Chez le gène de la préprocapitellacine par contre, un nombre variable de copies du gène par individu (jusqu'à 3 allèles pour certains individus au clade B) a pu être observé sans l'ombre d'un doute. Ces observations nous permettent d'émettre l'hypothèse qu'un polymorphisme dans le nombre de copies semble être favorisé chez ce type de peptides antimicrobiens à BRICHOS. Ce phénomène appelé CNV pour *Copy Number Variation* (copie de portion de génome >1000pb) peut avoir des conséquences phénotypiques notamment via un effet dose qui ont été plus amplement décrits pour des fonctions de défense, de réponses à des stress (biotique et abiotique) chez les animaux et les plantes (Lupski and Stankiewicz, 2005; Prunier et al., 2017). Chez les beta-défensines humaines par exemple, un phénomène de CNV a pu être décrit chez l'homme qui peut être lié à la susceptibilité à la maladie de Crohn et au psoriasis (Hollox, 2008). Ce mode de fonctionnement pourrait dès lors suggérer que la fonction première de ce type de PAM à BRICHOS n'est pas le contrôle de la symbiose mais plus un système de réponse immunitaire à la variabilité des environnements biotiques et abiotiques des vers.

1.3. Une diversification non-synonyme des régions BRICHOS et PROREGION dans les 2 PAMs.

Les résultats du chapitre 2 ont mis en évidence l'action d'une sélection diversifiante sur les paralogues du préproalvinellacine au milieu de la région codant la chaperonne (BRICHOS) et la première partie de la prorégion. Ces portions concentrent en effet la plupart des remplacements en acides aminés trouvés au sein et entre les différents paralogues de ce gène. Les résultats du chapitre 3 montrent que cette divergence non-synonyme entre paralogues s'effectue également au sein de la prorégion et sur la région C-terminale du domaine BRICHOS dans un endroit presque similaire à la région sous sélection positive décrite pour l'alvinellacine (bien que plus étalée) lorsque les clades A et B sont comparés. Dans le cas de la lignée mitochondriale la plus divergente (Cc-Manche2) auquel s'associe le clade C (18% de divergence avec les clades A et B), il semble que le gène codant la préprocapitellacine ait évolué à travers un relâchement des pressions de sélection (voir

d'une légère sélection purifiante) avec cependant un pic de divergence non synonyme basé sur quelques acides aminés, localisé au début de la prorégion uniquement. Ces résultats supportent l'hypothèse d'une diversification sous sélection positive des différents paralogues à des positions bien précises du gène, ces positions en début de la prorégion et au milieu du domaine BRICHOS ayant une forte probabilité d'être en déséquilibre de liaison même si cette hypothèse n'a pu être démontrée eu égard aux difficultés que nous avons eu dans la re-capture des allèles et leur génotypage au sein des individus. Il semble donc qu'un mécanisme bien particulier portant sur le fonctionnement optimal de cette défense antimicrobienne ait pu faire l'objet d'une sélection balancée pour générer une telle diversification non-synonyme du prépropeptide. Certains auteurs (Duda et al., 2002; Hollox and Armour, 2008; Hughes and Yeager, 1997a) ont rapporté l'existence d'une complémentarité de charge entre la prorégion et le PAM qui conduirait à une co-évolution sous sélection positive de ces différentes parties du gène. Ce type d'évolution parallèle n'est pas envisageable dans le cas de l'alvinellacine puisque le PAM mature est monomorphe. Une hypothèse alternative sur le rôle de la température a été avancée pour expliquer la diversification du domaine BRICHOS. Elle repose sur l'extrême variabilité thermique de l'environnement hydrothermal et la nécessité d'une bonne conformation du PAM sur une large gamme de températures (le rôle de chaperon moléculaire étant documenté : Willander et al., 2011). Cette hypothèse peut également être proposée dans le cas de la capitellacine car les espèces de *Capitella* sont intertidales et, confrontées à des variations thermiques importantes, qu'elles soient journalières (alternance jour/nuit), tidales (alternance des conditions d'immersion/émersion) ou saisonnières. Néanmoins, la diversité non synonyme rapportée dans le cas du chapitre 3 pour la préprocapitellacine au niveau des trois phylogroupes Cc-Manche1, Cc-Atlantique, et Cc-Manche2 pourrait être également liée à de l'introgression adaptative entre ces lignées après une période d'isolement plus ou moins longue dans des refuges glaciaires ou les habitats auraient été différents. Dans ce cas, les remises en contact secondaire des différentes lignées auraient pu promouvoir le passage de certains allèles entre lignées et leur maintien en fonction de l'hétérogénéité des habitats (lagunes, estuaires, ports) et des microbiomes associés grâce à ce système de gènes dupliqués en tandem.

Dans le cas du peptide signal, les deux études s'accordent globalement sur la présence d'une sélection purifiante sur cette fonction d'adressage puisque celui-ci est monomorphe pour la capitellacine et pour 4 des paralogues de la preproalvinellacine. Ainsi, pour cette famille de PAM à BRICHOS, il semble que modifier le mode d'adressage vers le réticulum endoplasmique (Zhang and Gallo, 2016) ou modifier le site de clivage du PAM (Martoglio and Dobberstein, 1998) soit plutôt délétère pour les 2 taxons étudiés.

1.4. Un gène traversant les barrières semi-perméables aux flux de gène après isolement géographique.

L'analyse populationnelle de la diversité génétique des gènes de la préprocapitellacine et de la préproalvinellacine suggère d'un flux de gènes important entre des entités (lignées) génétiques en phase d'isolement. Que ce soit entre les populations nord et sud d'*A. pompejana* ou entre populations des lignées mitochondriales Cc-Atlantique et Cc-Manche1, les différents allèles du prépropeptide sont capables de traverser la barrière géographique qui sépare ces entités en cours de spéciation. Ces résultats traduisent donc l'existence soit d'une zone de contact secondaire ou les espèces seraient capables de s'hybrider, soit de continuer d'échanger à travers une barrière physique à la dispersion (i.e. barrière équatoriale de l'EPR : Plouviez et al. 2009, 2010, 2013 ou zone du front de Ouessant : Jolly et al. 2006, 2009). Qu'il s'agisse d'introgression d'allèles suite à de l'hybridation locale (barrière génétique), du reflet d'une zone de plus faible migration (barrière physique) ou des 2 réunis, l'absence (ou la faible) différenciation génétique des populations de ces vers aux locus codant les PAMs alors que ces barrières sont quasi-imperméables aux allèles mitochondriaux pourrait s'expliquer par le passage d'allèles avantageux au travers d'une barrière semi perméable permettant la mise en commun d'un pool d'allèles entre des 'espèces' précédemment isolées en allopatrie. Cette idée d'introgression adaptative est d'ailleurs renforcée par le fait que, pour chaque taxon (*A. pompejana* et *Capitella spp*.), la différenciation génétique des populations de part et d'autre de la barrière est beaucoup plus marquée pour l'un des 2 paralogues testés (i.e. paralogue 4 chez *A. pompejana* et clade B chez *Capitella spp.*), laissant supposer qu'un balayage sélectif ai pu avoir lieu pour l'autre paralogue (i.e. paralogue 1 chez *A. pompejana* et clade A chez *Capitella spp.*). Ces résultats, bien que basés sur un petit nombre d'allèles recapturés et un faible nombre d'individus génotypés, sont assez semblables à la structure génétique des PAMs mytiline B et défensine

2 du complexe d'espèces *Mytilus edulis/M. galloprovincialis* rapportés par Boon et al. (2009) de part et d'autre de la barrière génétique séparant ces 2 espèces. Dans ce cas d'étude, l'absence de différenciation et le mélange d'allèles divergents dans les différentes populations géographiques étudiées (avec des écarts à l'équilibre mutation/dérive) renforcent l'hypothèse d'un effet sélectif homogénéisant entre les 2 taxons même si les données ne permettent pas de pouvoir trancher entre un effet balancée ou une sélection directionnelle. Nos travaux supportent donc l'hypothèse que les peptides antimicrobiens des invertébrés marins sont soumis à de fortes pressions sélectives probablement liées aux mécanismes de co-évolution entre l'hôte et ses épibiontes pouvant varier en fonction de l'habitat, tout au moins en ce qui concerne le prépropeptide et plus spécifiquement le domaine BRICHOS.

En conclusion, de nombreuses similitudes peuvent être mises en évidence laissant supposer à un mode de sélection assez semblable entre les espèces qui agirait pour promouvoir et/ou maintenir la diversité génétique dans certaines régions du précurseur protéique de ces PAM à BRICHOS. La principale différence réside en la différence de diversité génétique, entre les deux espèces étudiées *A. pompejana*/Capitella *spp*., retrouvée dans la région codant le PAM. Ceci pourrait être du à un consortium bactérien hydrothermal ancien imposant une sélection purifiante sur la molécule antimicrobienne par sa stabilité dans le temps et l'espace. Ceci diffère par rapport aux environnements côtiers où les communautés microbiennes peuvent évoluer rapidement dans le temps et l'espace. Dans les deux cas, c'est l'action de la sélection balancée qui pourrait générer une telle diversification non-synonyme du prépropeptide.

2. Comparaison des activités et thermostabilité des 2 peptides antimicrobiens : études préliminaires réalisées au laboratoire dans le cadre du stage de M2 de Lolita Roisin.

Pour comparer les performances des PAMs à BRICHOS à différentes température, une étude préliminaire a été effectuée au laboratoire pour étudier la thermostabilité de l'alvinellacine et de la capitellacine (de *Capitella telata*). Une précédente étude sur l'arénicine (PAM de la même famille que l'alvinellacine et la capitellacine) a montré que ce PAM produit par l'annélide *Arenicola marina,* est actif aussi bien à 4°C qu'à 37°C montrant une certaine thermostabilité de la molécule (Andrä et al., 2008). L'étude réalisée au laboratoire (Figure 1, L. Roisin, stage de M2) a permis de tester les activités antimicrobiennes après incubation des PAMs à différentes températures sur plusieurs durées d'incubation. Ces résultats, exprimés en pourcentage d'activité résiduelle (100% représentant une culture bactérienne contrôle sans présence de peptide) ont permis de mettre en évidence que l'alvinellacine est plus thermostable que la capitellacine qui perd au moins 60% de son activité en étant chauffée à 42°C. L'alvinellacine quant à elle, après avoir été chauffée à 90°C, garde 50% de son activité. Ces résultats indiquent, outre le fait que la structure primaire de l'alvinellacine ait évolué pour s'adapter à la nature de la diversité des bactéries hydrothermales et notamment aux epsilon-proteobacteries qui composent les épibiontes, que les différences observées dans la composition du peptide sont la résultante d'une longue adaptation de la molécule hydrothermale à un environnement thermiquement fluctuant puisque cette molécule reste fonctionnelle (bien que moins active) même à 90°C. Dans les 2 cas, le peptide présente un structure en épingle à cheveux maintenue par 2 ponts disulfures, suggérant que la position de ces ponts, légèrement différentes entre les 2 espèces, peut être déterminante dans la stabilité thermique de la molécule.

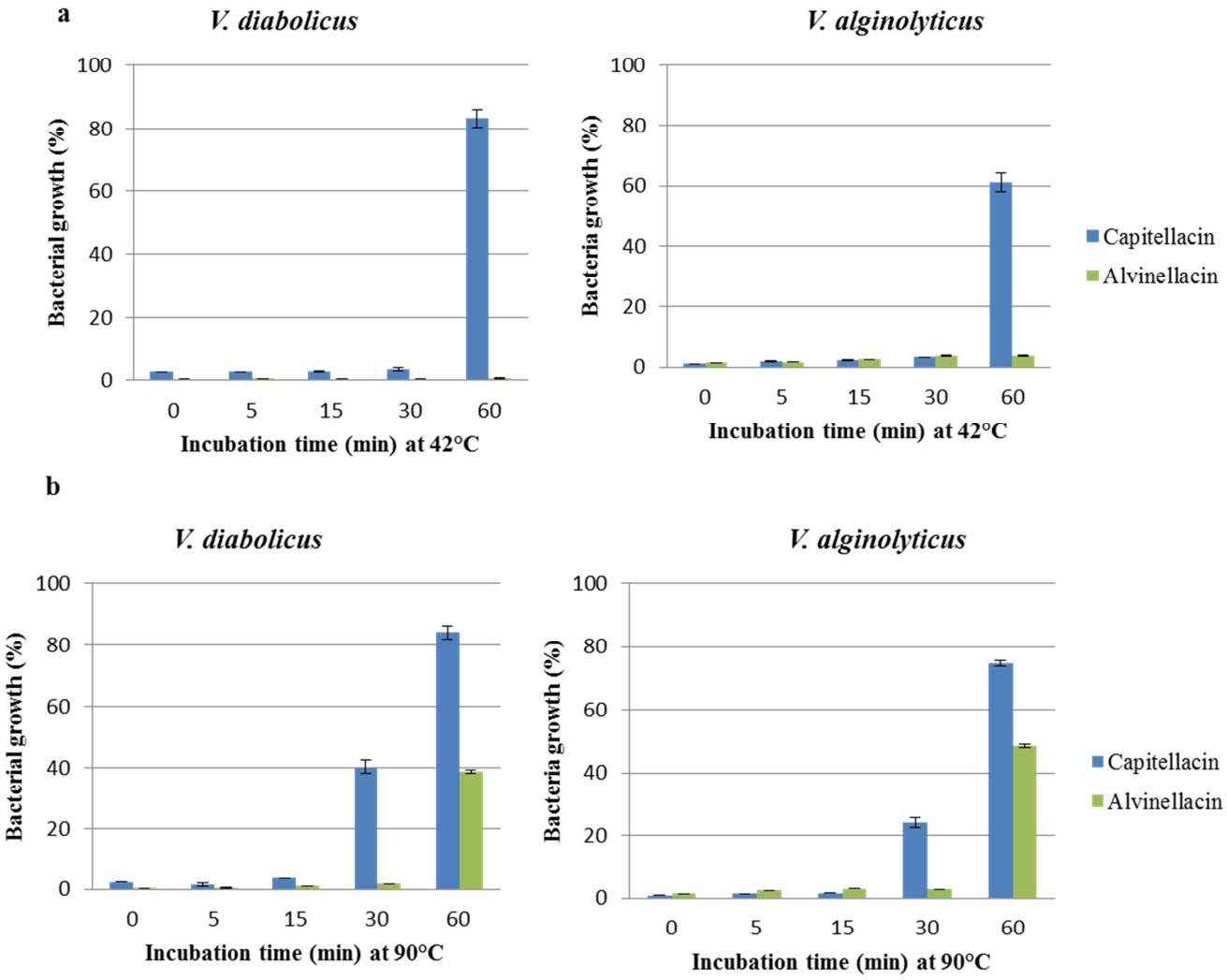


Figure 1. Etude de l'inhibition de la croissance bactérienne par les deux peptides antimicrobiens capitellacine et alvinellacine incubés de 0 à 60 minutes à différentes températures contre la bactérie côtière (*Vibrio alginolyticus*) et la bactérie hydrothermale (*Vibrio diabolicus*). L'inhibition de la croissance bactérienne est observée après 18h à 37°C.

Une deuxième étude (Figure 2) a permis de comparer le mode d'action des deux peptides antimicrobiens alvinellacine et capitellacine (de *Capitella teleta*) sur *Vibrio alginolyticus*. Les modes d'action sont les mêmes et de type « pore forming » (flèches rouges) : le PAM aurait un effet sur l'intégrité membranaire de la bactérie avec relargage du contenu intracellulaire. Cette observation est également vraie dans le cas de l'épibionte de *Capitella spp. Thiomargarita*. L'analyse en effet a montré que capitellacine et alvinellacine tuent également cette bactérie et qui présente un immuno-marquage important avec l'anticorps anti-capitellacine traduisant l'accumulation de peptide à l'intérieur des bactéries tuées. Ces

études suggèrent donc que la structure primaire intrinsèque du PAM n'a que peu d'effet sur le mécanisme d'action du PAM et pose la question du rôle des changements des acides aminés dans la séquence primaire, dès lors que la structure 3D du peptide n'est pas altérée. Si ces changements ont un rôle avéré en termes de thermostabilité pour le cas précis de l'alvinellacine, il est actuellement difficile de savoir si ces différences sont le résultat d'une accumulation neutre de mutations non-synonymes au cours du temps (après spéciation) ou si certains remplacements sont de nature adaptative pour permettre une bonne co-évolution de l'hôte avec les bactéries qui l'entourent.

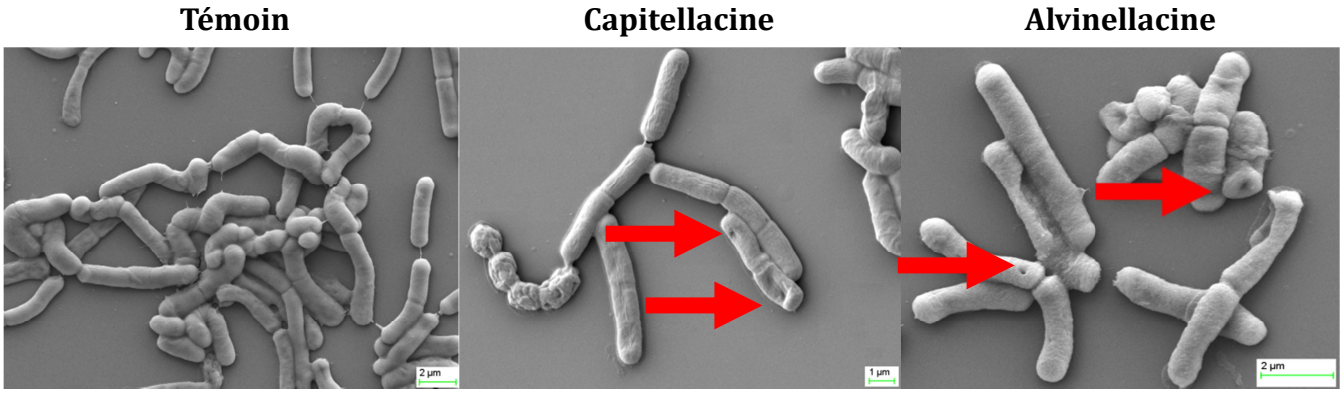


Figure 2. Images en microscopie électronique à balayage de *V. alginolyticus* incubées avec ou sans peptides pendant 4h. Les flèches indiquent la formation de pore dans la paroi bactérienne

3. Perspectives

En étant secrété dans le milieu extracellulaire, les peptides antimicrobiens sont alors fortement contraints par les conditions physico-chimiques de l'habitat de l'hôte. Dans ce contexte, le bon repliement des peptides est un processus primordial pour obtenir une molécule native fonctionnelle. La dénaturation et/ou l'agrégation de ces molécules sont généralement associées à l'altération des hélices et des feuillets et/ou de leurs interactions. La capacité des polypeptides à se replier en une seule et unique conformation (structure tri-dimensionnelle) peut dépendre *in vivo* à la fois de sa structure en acides aminés mais aussi de l'interaction avec diverses protéines et/ou d'enzyme qui permettent un repliement correct de ces protéines (chaperon moléculaire : Dobson, 2003). Les chaperons moléculaires représentent une large variété de protéines ubiquitaires qui possèdent toutes la caractéristique de se lier à des protéines au repliement incomplet et, qui, de ce fait,

augmentent la probabilité pour celles-ci de ne pas *in fine* former d'interactions inter-moléculaires (phénomène d'agrégation) (Clark et Elcock, 2016).

Les PAMs étudiés dans cette thèse sont sécrétées dans l'environnement associé aux annélides étudiés puisque ceux-ci sont retrouvés dans le mucus des organismes et/ou en contact direct avec le consortium épibiotique des espèces étudiées. Ceux-ci sont alors logiquement fortement contraints par les conditions physico-chimiques de l'habitat dans lequel ces organismes évoluent. Les précurseurs protéiques de ces peptides antimicrobiens possèdent la structure typique des protéines à BRICHOS trouvée dans la littérature, ce domaine BRICHOS étant également rapporté comme étant excrété dans le milieu extracellulaire (Hedlund et al., 2009; Sánchez-Pulido et al., 2002). Il est donc très probable que ce domaine chaperon soit également présent dans l'environnement externe des annélides, pour permettre le transport du peptide antimicrobien en feuillet beta, puis clivé lorsque la cible est atteinte. Cette conformation peptide-BRICHOS est largement documentée pour des peptides/protéines associées à diverses maladies lorsque ceux-ci s'auto-agrègent et forment des plaques amyloïdes toxiques pour les cellules (au niveau des neurones dans le cas de la maladie d'Alzheimer (Willander et al., 2011). Dans ce contexte, le domaine BRICHOS prévient l'agrégation en plaques amyloïdes des peptides en feuillet beta auxquels il est associé (Peng et al., 2010; Willander et al., 2012). Même si la production de fibrilles amyloïdes est généralement associée à un signe de maladie, des preuves de plus en plus nombreuses suggèrent que leur synthèse peut également contribuer à la physiologie normale des organismes en accomplissant diverses fonctions importantes par exemple la mélanisation chez les invertébrés (Grimaldi et al., 2012).

Les plaques amyloïdes

Leur formation est donc associée à la conversion d'une protéine (ou d'un peptide) de son état fonctionnel soluble à un état d'agrégation organisé en fibres ou fibrilles qui tuent les cellules ou les empêchent de fonctionner normalement (Sipe and Cohen, 2000). Ces structures sont décrites comme des dépôts localisées dans le milieu extracellulaire et leur formation se fait d'une façon organisée avec : dans un premier temps la formation d'un noyau qui peut ensuite se développer en des structures fibrillaires (protofibrilles ou protofilaments). Celles-ci sont, dans un second temps, élongées et peuvent s'enrouler pour

former une fibrille mature (structure de feuillets beta croisés colinéaire à l'axe de la fibre) (Fändrich, 2007). Les fibres amyloïdes matures sont caractérisées par au moins trois grandes propriétés physico-chimiques facilitant leur étude. Tout d'abord, elles ont une apparence de fibres rectilignes, non ramifiées d'un diamètre d'environ 10 nm en microscopie électronique à transmission (MET) et à force atomique (AFM). Ensuite, elles possèdent également la propriété d'interagir avec certains colorants dont le Rouge Congo ou la thioflavine T (ThT) (Nilsson, 2004). Les protéines qui forment des fibres amyloïdes de caractéristiques physico-chimiques similaires ne présentent cependant pas d'homologie de séquence, ni de structures tridimensionnelles communes. Dobson (2003) propose que la toxicité des agrégats résulte de leurs états non structurés ce qui conduirait à une mauvaise interaction des résidus habituellement enfouis avec certaines protéines qui interviennent dans des processus vitaux pour la cellule.

Lien entre peptide amyloïde et peptide antimicrobien

Les peptides antimicrobiens (PAM) et les peptides amyloïdes (AMY), rapportés sous le nom de peptides cytolytiques, ont été décrits dans un premier temps comme dissemblables du point de vue de la taille (structure primaire courte vs taille variable), de la charge nette (cationique et amphipatique vs principalement hydrophobe), et de l'activité biologique (activités antimicrobiennes majoritaires vs activités anti-neuronales) (Butterfield and Lashuel, 2010; Zasloff, 2002). Cependant leur mode d'action est similaire en se liant aux membranes des cellules permettant pour l'un de compromettre l'intégrité membranaire des bactéries et pour l'autre, l'intégrité des cellules neuronales tout en étant peu toxiques pour d'autres types cellulaires (Bucciantini et al., 2002; Last and Miranker, 2013). Des études récentes ont permis de montrer que AMYs et PAMs possèdent de nombreuses similitudes : les PAMs peuvent s'auto-assemblés en une structure semblable aux fibrilles amyloïdes (« cross-beta-sheet structures ») et les AMYs peuvent quant à eux montrer des activités antimicrobiennes au moins similaires voir supérieures à certains PAMs (Chu et al., 2012; Kagan et al., 2012; Soscia et al., 2010). Il a été montré par exemple que le peptide amyloide beta (A-beta), médiateur clé de la maladie d'Alzheimer est un peptide antimicrobien en tant que tel ayant des activités antimicrobiennes similaires voire supérieures au peptide antimicrobien LL-37 (de la famille des cathélicidines humaines) (Soscia et al., 2010). Ceci permet aux auteurs de suggérer qu'une activité antimicrobienne pourrait être sa fonction *in*

vivo première. Dans le cas de la protégrine-1 (peptide en feuillet beta), il a été montré que son mode d'action de type « channel forming » est très similaire au mode d'action du peptide A-beta (Jang et al., 2008), ce qui permet de suggérer que c'est la structure en feuillet beta qui joue un rôle critique dans la prédisposition des peptides à interagir avec les membranes et à former des canaux dans les membranes. Jang et al. (2011) ont ensuite montré que la protégrine-1 pouvait également former des structures similaires aux fibrilles du peptide amyloïde leur permettant, à eux aussi, de conclure à une fonction antimicrobienne possible chez les AMYs. Ainsi, des similitudes à la fois fonctionnelles et structurales pourraient suggérer un lien potentiel entre peptides amyloïdes et peptides antimicrobiens avec des activités antimicrobiennes et des propriétés amyloïdes communes pour ces types de peptides (review in Zhang et al., 2014).

Ainsi, deux perspectives principales peuvent être définies :

- Etudier la stabilité conformationnelle des peptides antimicrobiens avec et sans domaineS BRICHOS en fonction des conditions abiotiques
- Déterminer l'effet de la variation génétique sur les activités antimicrobiennes et sur la formation des agrégats selon les conditions du milieu

3.1. Conditions abiotiques et agrégation/stabilité du PAM

L'agrégation des peptides antimicrobiens pour former des pores dans les parois bactériennes a largement été documentée notamment puisque c'est cette propriété qui permet leur fonctionnalité antimicrobienne : par exemple dans le modèle « barrel-stave ». L'agrégation des peptides précède en effet l'insertion de ceux-ci dans la bicouche membranaire (Brogden, 2005). Dans le cas des PAMs sécrétés vers l'extérieur, et notamment l'alvinellacine, les conditions physico-chimiques (température, pH, composition chimique du milieu) peuvent avoir un impact direct sur l'agrégation des protéines/peptides.

Par exemple, il a été montré que la présence de certains métaux favorise l'agrégation des molécules en feuillets beta sous certaines conditions physico-chimiques comme cela a pu être démontré pour le Zn(II) et Cu(II) avec le peptide A-beta responsable de la maladie d'Alzheimer (Miura et al., 2000). L'action du Fe(III), Al(III) sur l'agrégation de ce même peptide A-beta a été également montré selon la concentration de ces métaux (Kawahara et al., 2001; Mantyh et al., 1993). Au contraire, certains composés tels que le dimethylsulfoxide

(DMSO) permettent la dissolution complète des fibrilles amyloïdes formées par le beta2-microglobuline (bien qu'il faille de très fortes concentrations). L'impact de la température et de la pression hydrostatique sur la stabilité des fibrilles amyloïdes a été étudié et a permis de mettre en évidence que les fibrilles amyloïdes matures sont hautement stables à forte pression mais que des températures élevées peuvent par contre facilement casser la plupart des agrégats (Dirix et al., 2005; Meersman and Dobson, 2006). La température et la pression sont reconnus comme étant parmi les variables les plus importantes pour la conformation des peptides (importants pour la dynamique conformationnelle) et que changer par exemple la pression dans laquelle évolue une protéine pouvait favoriser l'apparition d'agrégats (Meersman et al., 2006). Il a été montré que des fibrilles amyloïdes formées par un peptide similaire au peptide A-beta (responsable de la maladie d'Alzheimer) pouvaient être dissociées par des températures autour de 100°C (373K) (Sasahara et al., 2005).

Ainsi, la formation de fibrilles amyloïdes se fait selon des conditions *in vitro* ou *in vivo* diverses et variées qui varient selon les organismes étudiés et/ou les conditions physico-chimiques rencontrées par l'organisme. Il est donc particulièrement intéressant d'étudier la capacité d'auto-agrégation du/des peptide(s) dans des conditions environnementales proches du milieu hydrothermal et/ou côtier. Ces études pourront donc être menées sur les PAMs alvinellacine et capitellacine à l'aide des techniques telles la coloration des agrégats au rouge Congo ou le suivi de la formation de plaques par ThT. Ces analyses seront couplées à des mesures d'activités antimicrobiennes pour déterminer dans quelle mesure ces facteurs environnementaux inhibent/favorisent la lyse de la membrane bactérienne selon leurs effets sur la dynamique d'agrégation.

Une étude préliminaire réalisée au laboratoire semblerait montrer qu'en conditions acides, l'alvinellacine s'auto-agrège (Aurélie Tasiemski, pers comm). Le rôle des variants des domaines BRICHOS dans ce phénomène reste à investiguer.

3.2. Polymorphisme et activité/agrégation

Le polymorphisme non synonyme retrouvé dans la région codante d'un PAM peut être maintenu par un impact direct sur la valeur sélective de l'individu porteur et/ou la résultante d'adaptation à des environnements fluctuants dans le temps et l'espace. L'évolution des peptides antimicrobiens est globalement documentée comme étant soumise à une évolution positive inter-espèces mais aussi entre les différents paralogues trouvés au sein d'une espèce lorsque les PAMs sont codés par des familles multigéniques chez les vertébrés et les invertébrés. L'évolution parallèle de plusieurs gènes permet de diversifier l'activité antimicrobienne sans coût additionnel en termes de mortalité différentielle lorsque l'environnement fluctue, d'acquérir une spécificité de la cible antimicrobienne ou promouvoir de la néofonctionnalité. Aussi, l'augmentation en fréquence de certains variants pourrait se produire principalement lorsque les hôtes pénètrent dans de nouvelles niches et sont forcés de s'adapter à de nouvelles espèces pathogènes non rencontrées auparavant (Tennessen, 2005).

Par exemple, Unckless et al., 2016 ont permis de montrer l'impact d'un seul changement en acide aminé dans la séquence de la diptericine sur la résistance de drosophiles (*Drosophila melanogaster* et *Drosophila simulans*) à des infections bactériennes. La seule modification d'un unique acide aminé dans la séquence codante de peptide antimicrobien a également été montré comme impactant les activités antifongique de 6 isoformes de la drosomycine (Yang et al., 2006;). De plus, Kagan et al. (2012) expliquent que "*A number of different factors can contribute to the formation of amyloid β-sheet structures including proteolysis, amino acid mutation, high concentration, acidic pH, binding to metals and interaction with lipid membranes*".

Dans ce contexte, les données de polymorphisme obtenues au sein du chapitre 3 pourront permettre de tester si fonctionnellement le polymorphisme décrit est donc le résultat d'une accumulation neutre de mutations non-synonymes au cours du temps ou si certains remplacements sont effectivement de nature adaptative (modification du mode d'action/ d'agrégation optimale en fonction des conditions physico-chimique ?). Ainsi, il pourra être possible d'étudier si chaque variant du PAM de la capitellacine possède une activité antimicrobienne qui lui est propre et des différences d'agrégation en fonction des conditions biotique/abiotique du milieu. La synergie entre peptides antimicrobiens pourra également être testée. Dans le cadre du chapitre 2, les deux variants de l'alvinellacine : celle d'*A.*

pompejana et *A. caudata*, pourront également être testés pour tester si cette variation de séquence est donc neutre du point de vue des activités ce qui confirmerait que l'action de la sélection purifiante qui agirait sur ces PAMs seraient pour garder une fonctionnalité fixée tôt dans l'évolution de ces espèces sœurs de vers hydrothermales

3.3. Rôle de la chaperonne.

Le rôle inhibiteur du domaine BRICHOS dans l'agrégation des peptides amyloïdes en feuillet beta a été montré a de nombreuses reprises (Peng et al., 2010; Sánchez-Pulido et al., 2002; Willander et al., 2011). Le domaine BRICHOS peut à ce titre être considéré comme un chaperon moléculaire en permettant une bonne conformation du peptide auquel il est associé mais cette fonction n'a jamais été démontrée dans le cas des peptides antimicrobiens puisque, seuls les PAMs d'annélide possèdent jusqu'à présent cette caractéristique (Tasiemski et al. 2014).

Une étude préliminaire sur l'impact de la température et la pression hydrostatique suggèrerait que ces facteurs affectent la conformation en feuillet beta de l'alvinellacine (en favorisant son auto-agrégation à hautes températures/pressions : Aurélie Tasiemski *personnal comm*) et, en conséquence, sa capacité à former des agrégats et *in fine* à tuer/sélectionner le consortium bactérien. De plus, une première étude structurale suggère également que le domaine BRICHOS changerait de conformation et ce de façon réversible lorsqu'il est soumis à des températures et/ou à des pressions croissantes (Willander et al., 2012). Dans ce contexte, la diversification du domaine BRICHOS en de nombreux variants alléliques chez l'alvinellacine et la capitellacine pourrait être sous sélection positive en raison de son rôle anti-amyloïde qui favoriserait, ou non, l'agrégation du PAM en fonction des conditions physico-chimiques rencontrées. Ceci pourrait permettre une fonctionnalité optimale du peptide en feuillet beta auquel il est rattaché dans une large gamme de conditions environnementales telles que celles rencontrées au niveau des sources hydrothermales ou dans le domaine intertidal.

Ainsi, il serait donc intéressant d'initier des études sur l'agrégation du peptide selon les différents variants du BRICHOS en changeant les conditions physico chimiques et de visualiser les relations entre conformation/agrégation des peptides antimicrobiens (et leurs

activités antimicrobiennes) pour tester l'hypothèse d'un rôle de chaperon moléculaire. Pour réaliser ce type d'étude fonctionnelle de relation entre peptide et molécule chaperonne, la production de protéines recombinantes du domaine BRICHOS sur des variants prédéfinis représente une étape clé.

Cette production a pu être réalisée lors de ma thèse sur trois variants de l'alvinellacine et l'Annexe 4 présente les différentes étapes de surexpression du domaine BRICHOS réalisée au cours de cette thèse.

3.3.1. Surexpression des différents variants du domaine BRICHOS

3.3.1.1 Choix des trois variants du domaine BRICHOS de la preproalvinellacine

Les variants ont été choisis car ils contiennent les mutations trouvées en chapitre 2 comme étant diagnostiques des différents clades et comme étant potentiellement soumis à la sélection diversifiante (Figures 3 et 4).

Seul le domaine BRICHOS a été produit puisqu'il est reporté que pour les protéines chez lesquels ce domaine a été étudié, il est clivé à partir du précurseur protéique de part et d'autres (notamment par l'ADAM10, métalloprotéase) pour être libéré du domaine transmembranaire et du peptide en feuillet beta auquel il est largement rapporté comme étant associé. Les quelques études qui ont produit ce domaine BRICHOS ont également fait le choix de ne pas produire, dans la mesure du possible, le reste du précurseur protéique pour n'étudier que le lien entre ce domaine chaperon et le peptide (Peng et al., 2010; Willander et al., 2011) .

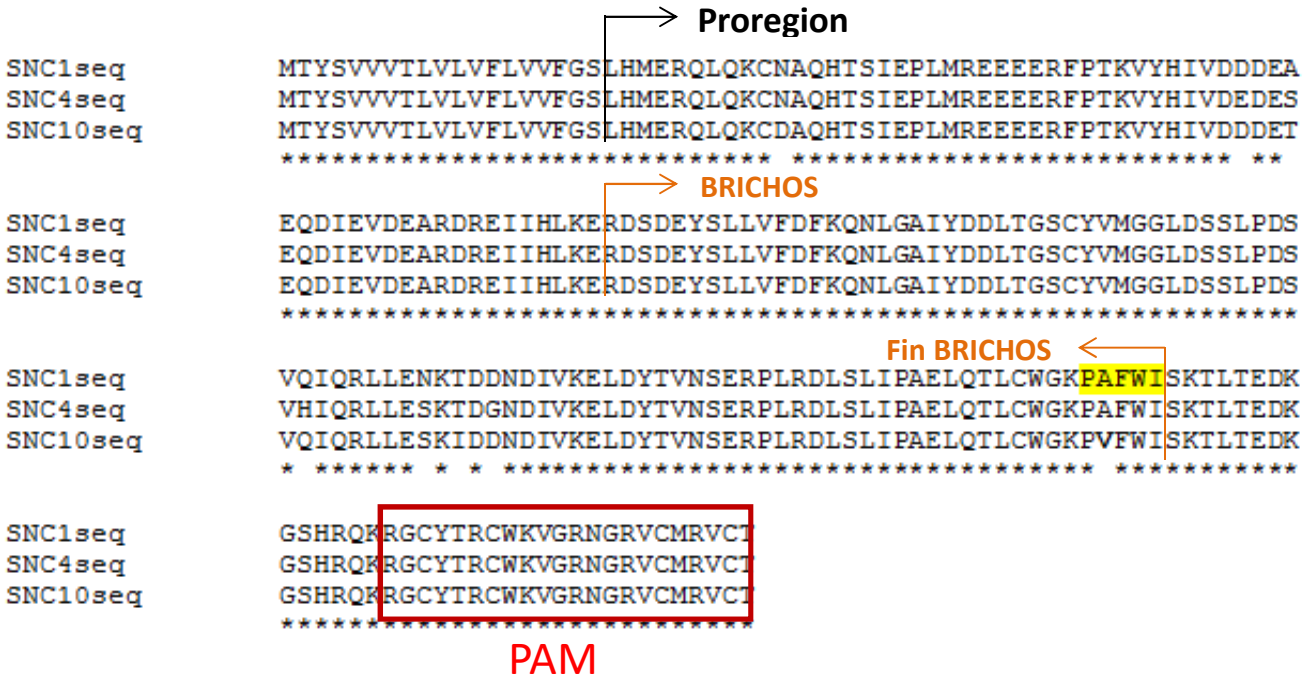


Figure 3. Alignement protéique du précurseur entier des trois variants choisis. Rouge : le PAM, orange : domaine BRICHOS, en noir la prorégion et en amont : le peptide signal. La présence d'étoile sous l'alignement montre un alignement parfait des acides aminés, l'absence montre un remplacement en termes d'acides aminés dans un des variants.

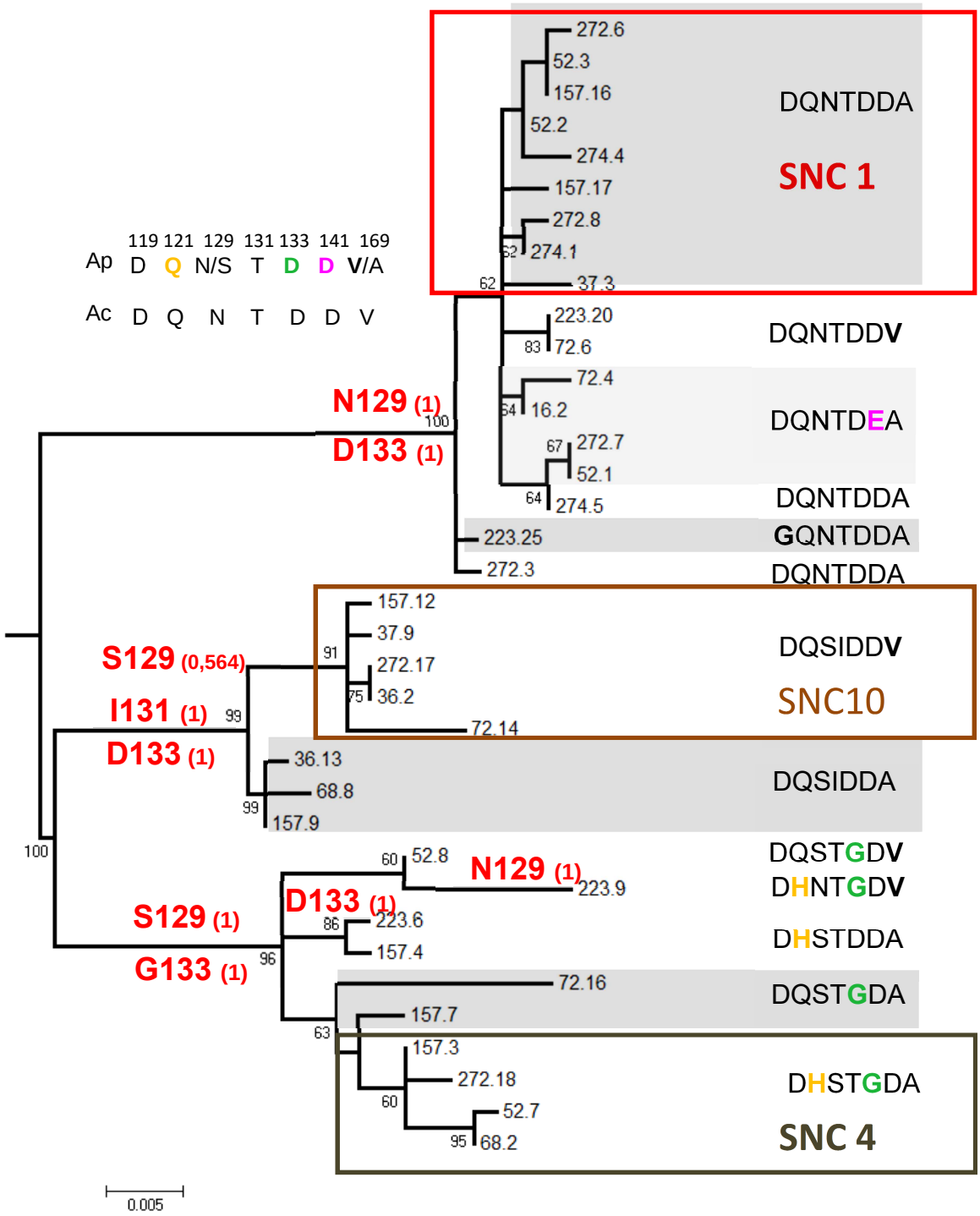


Figure 4. Issue du chapitre 2. Rappel des trois variants du BRICHOS choisis dans l'arbre de divergence des allèles de la région 3 ' où ont été cartographiées les différentes mutations du domaine BRICHOS. Les trois variants (appelés SNC1, SNC10, SNC4) choisis possèdent en conséquence les mutations diagnostique de chaque clade

3.3.1.2. Choix du vecteur d'expression et de la souche bactérienne.

Les bactéries E. coli Origami(DE3)pLysS de Novagen ont été choisies car elles permettent la formation de pont disulfure dans la molécule produite. En effet, il s'agit de souches commerciales de *E. coli* mutantes *qui* ont été modifiées dans le but de permettre un meilleur repliement des protéines et éviter la formation de corps d'inclusion. De plus, des mutations dans les gènes codant pour la thiorédoxine réductase *trxB* et glutathion réductase *gor* permettent une meilleure formation des ponts disulfures dans le cytoplasme et ainsi un meilleur repliement des protéines contenant des ponts disulfures. La souche pLysS a été choisie car elles expriment le lysozyme T7, qui inhibe l'activité basale de l'ARN polymérase T7, optimisant la régulation du niveau d'expression.

Le vecteur pet32c (Novagen, Madison, WI) a été choisi car il permettait une construction qui ne rajoute que très peu d'acides aminés au domaine BRICHOS tout en possédant un tag histidine (purification aisée par chromotagraphie d'affinité sur colonne de nickel). La construction (Figure 5) a donc été optimisée pour ajouter le moins de nucléotides possible en prenant en compte que les deux derniers acides aminés WI du domaine BRICHOS correspondent au site de coupure de l'enzyme BAMHI (avec décalage T+GGATTC-> WI) et le choix des trois premiers acides aminés correspondent au site de coupure de l'enzyme NcoI (enzyme dont le site de restriction est le plus proche de la coupure à l'enterokinase). Au final, seuls quatre acides aminés ont étés ajouté aux différents variants.

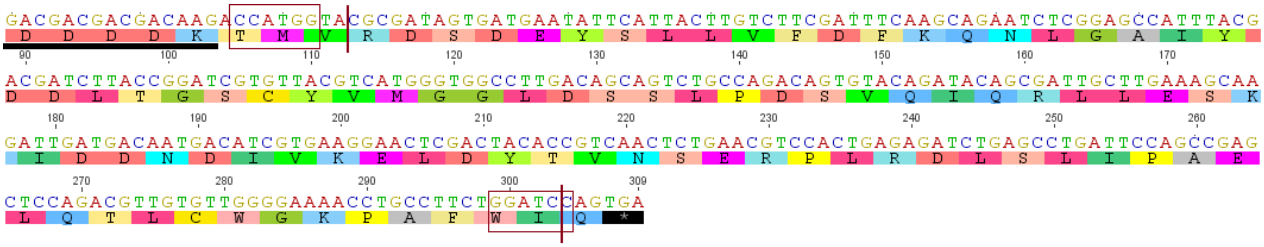


Figure 5. Construction permettant la production du domaine BRICHOS. Souligné en noir : site de coupure de l'enterokinase, entouré en rouge : site de coupure de NcoI puis de BAMHI, les deux barres verticales rouges correspondent au début et à la fin du domaine BRICHOS.

3.3.1.3. Résultats obtenus

La première étape a donc consisté à induire l'expression des protéines après avoir inséré les variants dans le vecteur d'expression selon le protocole de l'Annexe 4. Cette induction s'est effectuée grâce à une concentration de 0.25mM d'IPTG et a pu être visualisée par un Western-Blot. La Figure 6 montre le SDS-PAGE effectué à partir de la culture bactérienne après 6h d'induction puis à l'aide d'un anticorps anti-BRICHOS afin de visualiser si ce domaine pouvait réellement être détecté dans les différentes bandes et surtout dans celles de forte intensité en dessous de 26kDa puisque la taille attendue de la protéine produite à ce stade est de 25kDa.

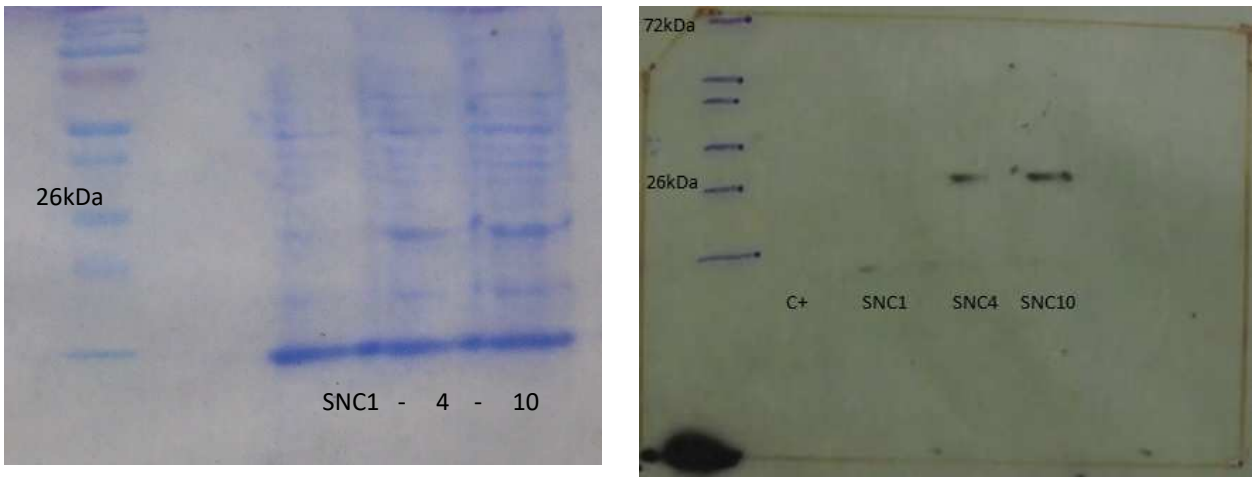


Figure 6. SDS-PAGE (12% acrylamide) réalisé après 6h d'induction pour les trois variants SNC1-SNC4-SNC10. Une bande d'intensité moyenne à forte à la taille attendue (25kDa) est visible révélant la présence de la protéine BRICHOS (10kDa) + vecteur d'expression (15kDa).

Deuxième SDS-PAGE

La deuxième étape, après induction et vérification de la sur-expression des protéines par les *E. coli*, a ensuite consisté à lyser les cellules bactériennes et à purifier la protéine d'intérêt à l'aide de son tag histidine sur colonne de nickel. Cette chromatographie possède la délicate étape d'éluer la protéine à une forte concentration (250mM) en retenant la protéine dans la colonne tout en éluant les autres protéines lors de premier lavage de la colonne. Ainsi, un SDS-PAGE permet de tester toutes les fractions récoltées après élution de la protéine pour déterminer si cette étape d'élution a permis de récupérer la protéine tagguée. Ceci est

récapitulé à la Figure 7 et permet de comparer les différentes fractions d'élution à une fraction « pure » après lyse des cellules sans passage sur colonne His-Trap. Ici, les concentrations sont bien décroissantes au fur et à mesure des élutions et il apparait que la fraction « pure » semble bien posséder plus de protéines que les fractions 1 à 7 passées sur HisTrap.

Pour la suite de l'analyse, les fractions 2 à 4 ont été concentrées puis digérées à l'enterokinase pour détacher la protéine du vecteur d'expression et obtenir la protéine seule (selon le protocole établis en Annexe 4).

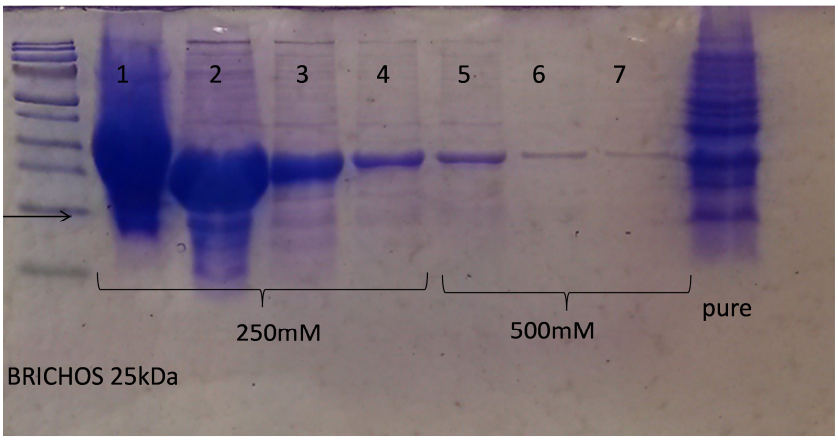


Figure 7. SDS-PAGE après passage sur colonne His-Trap. Deux élutions sont montrées pour un seul variant (ici SNC1) permettant de montrer les différentes concentrations en notre protéine d'intérêt au cours deux étapes majeures d'élutions (250mM imidazole et 500mM).

Cette expérimentation réalisée au cours de cette thèse m'a donc permis de surexprimer trois variants du BRICHOS qui sont désormais disponible au laboratoire pour les futurs tests fonctionnels. Ceci permettra de déterminer le rôle de ce domaine notamment pour déterminer si celui-ci aura dans ce système un rôle de chaperon moléculaire qui aidera le peptide en feuillet beta auquel il est associé à se replier correctement sous différentes conditions abiotiques.

Bibliographie

Adema, C.M., Hertel, L.A., Miller, R.D., and Loker, E.S. (1997). A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. Proc. Natl. Acad. Sci. *94*, 8691–8696.

Aguilar, A., Roemer, G., Debenham, S., Binns, M., Garcelon, D., and Wayne, R.K. (2004). High MHC diversity maintained by balancing selection in an otherwise genetically monomorphic mammal. Proc. Natl. Acad. Sci. *101*, 3490–3494.

Alayse-Danet, A.M., Desbruyeres, D., and Gaill, F. (1987). The possible nutritional or detoxification role of the epibiotic bacteria of Alvinellid polychaetes: review of current data. Symbiosis *4*, 51–61.

Andrä, J., Jakovkin, I., Grötzinger, J., Hecht, O., Krasnosdembskaya, A.D., Goldmann, T., Gutsmann, T., and Leippe, M. (2008). Structure and mode of action of the antimicrobial peptide arenicin. Biochem. J. *410*, 113–122.

Arce, A.N., Johnston, P.R., Smiseth, P.T., and Rozen, D.E. (2012). Mechanisms and fitness effects of antibacterial defences in a carrion beetle. J. Evol. Biol. *25*, 930–937.

Audzijonyte, A., Ovcarenko, I., Bastrop, R., and Väinölä, R. (2008). Two cryptic species of the *Hediste diversicolor* group (Polychaeta, Nereididae) in the Baltic Sea, with mitochondrial signatures of different population histories. Mar. Biol. *155*, 599–612.

Autem, M., Salvidio, S., Pasteur, N., Desbruyères, D., and 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 Académie Sci. Sér. 3 Sci. Vie *301*, 131–135.

Awadalla, P. (2003). The evolutionary genomics of pathogen recombination. Nat. Rev. Genet. *4*, 50–60.

Balandin, S.V., and Ovchinnikova, T.V. (2016). Antimicrobial peptides of invertebrates. Part 1. structure, biosynthesis, and evolution. Russ. J. Bioorganic Chem. *42*, 229–248.

Ballard, J.W., and Kreitman, M. (1994). Unraveling selection in the mitochondrial genome of Drosophila. Genetics *138*, 757–772.

Balseiro, P., Falcó, A., Romero, A., Dios, S., Martínez-López, A., Figueras, A., Estepa, A., and Novoa, B. (2011). *Mytilus galloprovincialis* myticin C: a chemotactic molecule with antiviral activity and immunoregulatory properties. PLoS One *6*, e23140.

Bandelt, H.-J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. *16*, 37–48.

Baross, J.A., and Holden, J.F. (1996). Overview of hyperthermophiles and their heat-shock proteins. Adv. Protein Chem. *48*, 1–34.

Bartels-Hardege, H.D., and Zeeck, E. (1990). Reproductive behaviour of *Nereis diversicolor* (Annelida: Polychaeta). Mar. Biol. *106*, 409–412.

Bartlett, T.C., Cuthbertson, B.J., Shepard, E.F., Chapman, R.W., Gross, P.S., and Warr, G.W. Crustins, Homologues of an 11.5-kDa antibacterial peptide, from two species of penaeid shrimp, *Litopenaeus vannamei* and *Litopenaeus setiferus*. Mar. Biotechnol. *4*, 278–293.

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

Barton, N.H., and Hewitt, G.M. (1985). Analysis of hybrid zones. Annu. Rev. Ecol. Syst. *16*, 113–148.

Barton, N.H., and Hewitt, G.M. (1989). Adaptation, speciation and hybrid zones. Nature *341*, 497–503.

Bazin, E., Glémin, S., and Galtier, N. (2006). Population size does not influence mitochondrial genetic diversity in animals. Science *312*, 570–572.

Belich, M.P., Madrigal, J.A., Hildebrand, W.H., Zemmour, J., Williams, R.C., Luz, R., Petzl-Erler, M.L., and Parham, P. (1992). Unusual HLA-B alleles in two tribes of Brazilian Indians. Nature *357*, 326–329.

Bellan, G., Desrosiers, G., and Willsie, A. (1988). Use of an annelid pollution index for monitoring a moderately polluted littoral zone. Mar. Pollut. Bull. *19*, 662–665.

Berlov, M.N., and Maltseva, A.L. (2016). Immunity of the lugworm *Arenicola marina*: cells and molecules. ISJ *13*, 247–256.

Bernhard, J.M., Buck, K.R., Farmer, M.A., and Bowser, S.S. (2000). The Santa Barbara Basin is a symbiosis oasis. Nature *403*, 77.

Berry, O.F. (2006). Mitochondrial DNA and population size. Science *314*, 1388–1390.

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

Bierne, N., Tanguy, A., Faure, M., Faure, B., David, E., Boutet, I., Boon, E., Quere, N., Plouviez, S., Kemppainen, P., et al. (2007). Mark–recapture cloning: a straightforward and cost-effective cloning method for population genetics of single-copy nuclear DNA sequences in diploids. Mol. Ecol. Notes *7*, 562–566.

Birky, C.W., Maruyama, T., and Fuerst, P. (1983). An approach to population and evolutionary genetic theory for genes in mitochondria and chloroplasts, and some results. Genetics *103*, 513–527.

Blake, J.A., Grassle, J.P., and Eckelbarger, K.J. (2009). *Capitella teleta*, a new species designation for the opportunistic and experimental *Capitella sp. I*, with a review of the literature for confirmed records. Zoosymposia *2*, 25–53.

Boman, H.G. (1995). Peptide antibiotics and their role in innate immunity. Annu. Rev. Immunol. *13*, 61–92.

Boman, H.G. (2003). Antibacterial peptides: basic facts and emerging concepts. J. Intern. Med. *254*, 197–215.

Boman, H.G., and Steiner, H. (1981). Humoral immunity in *Cecropia* pupae. In Current Topics in Microbiology and Immunology, W. Henle, P.H. Hofschneider, H. Koprowski, O. Maaløe, F. Melchers, R. Rott, H.G. Schweiger, and P.K. Vogt, eds. (Springer Berlin Heidelberg), pp. 75–91.

Boman, H.C., Boman, I.A., Andreu, D., Li, Z.Q., Merrifield, R.B., Schlenstedt, G., and Zimmermann, R. (1989). Chemical synthesis and enzymic processing of precursor forms of cecropins A and B. J. Biol. Chem. *264*, 5852–5860.

Boon, E., Faure, M.F., and Bierne, N. (2009). The flow of antimicrobial peptide genes through a genetic barrier between *Mytilus edulis* and *M. galloprovincialis*. J. Mol. Evol. *68*, 461–474.

Borghans, J.A., Beltman, J.B., and De Boer, R.J. (2004). MHC polymorphism under host-pathogen coevolution. Immunogenetics *55*, 732–739.

Bosch, T.C., Augustin, R., Anton-Erxleben, F., Fraune, S., Hemmrich, G., Zill, H., Rosenstiel, P., Jacobs, G., Schreiber, S., Leippe, M., et al. (2009). Uncovering the evolutionary history of innate immunity: the simple metazoan *Hydra* uses epithelial cells for host defence. Dev. Comp. Immunol. *33*, 559–569.

Brandt, A., De Broyer, C., Ebbe, B., Ellingsen, K.E., Gooday, A.J., Janussen, D., Kaiser, S., Linse, K., Schueller, M., Thomson, M.R.A., et al. (2012). Southern Ocean deep benthic biodiversity. Rogers AD Johnston NM Murphy EJ et al. Antarct. Ecosyst. Extreme Environ. Chang. World. Chichester: Blackwell Publ. Ltd., 291–334.

Brault, N., Bourquin, S., Guillocheau, F., Dabard, M.-P., Bonnet, S., Courville, P., Estéoule-Choux, J., and Stepanoff, F. (2004). Mio–Pliocene to Pleistocene paleotopographic evolution of Brittany (France) from a sequence stratigraphic analysis: relative influence of tectonics and climate. Sediment. Geol. *163*, 175–210.

Brey, P.T., Lee, W.-J., Yamakawa, M., Koizumi, Y., Perrot, S., Francois, M., and Ashida, M. (1993). Role of the integument in insect immunity: epicuticular abrasion and induction of cecropin synthesis in cuticular epithelial cells. Proc. Natl. Acad. Sci. *90*, 6275–6279.

Bright, M., and Giere, O. (2005). Microbial symbiosis in Annelida. Symbiosis *38*, 1–45.

Brogden, K.A. (2005). Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? Nat. Rev. Microbiol. *3*, 238–250.

Brown, W.M., George, M., and Wilson, A.C. (1979). Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. *76*, 1967–1971.

Bucciantini, M., Giannoni, E., Chiti, F., Baroni, F., Formigli, L., Zurdo, J., Taddei, N., Ramponi, G., Dobson, C.M., and Stefani, M. (2002). Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. Nature *416*, 507–511.

Buchon, N., Silverman, N., and Cherry, S. (2014). Immunity in *Drosophila melanogaster* from microbial recognition to whole-organism physiology. Nat. Rev. Immunol. *14*, 796–810.

Bulet, P., Hetru, C., Dimarcq, J.-L., and Hoffmann, D. (1999). Antimicrobial peptides in insects; structure and function. Dev. Comp. Immunol. *23*, 329–344.

Bulet, P., Stöcklin, R., and Menin, L. (2004). Antimicrobial peptides: from invertebrates to vertebrates. Immunol. Rev. *198*, 169–184.

Bulmer, M.S., and Crozier, R.H. (2004). Duplication and diversifying selection among termite antifungal peptides. Mol. Biol. Evol. *21*, 2256–2264.

Bulmer, M.S., and Crozier, R.H. (2006). Variation in positive selection in termite GNBPs and Relish. Mol. Biol. Evol. *23*, 317–326.

Bulmer, M.S., Bachelet, I., Raman, R., Rosengaus, R.B., and Sasisekharan, R. (2009). Targeting an antimicrobial effector function in insect immunity as a pest control strategy. Proc. Natl. Acad. Sci. *106*, 12652–12657.

Bulmer, M.S., Lay, F., and Hamilton, C. (2010). Adaptive evolution in subterranean termite antifungal peptides. Insect Mol. Biol. *19*, 669–674.

Butterfield, S.M., and Lashuel, H.A. (2010). Amyloidogenic protein–membrane interactions: mechanistic insight from model systems. Angew. Chem. Int. Ed. *49*, 5628–5654.

Carr, C.M., Hardy, S.M., Brown, T.M., Macdonald, T.A., and Hebert, P.D. (2011). A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes. PLoS One *6*, e22232.

Cary, S.C., Cottrell, M.T., Stein, J.L., Camacho, F., and Desbruyeres, D. (1997). Molecular identification and localization of filamentous symbiotic bacteria associated with the hydrothermal vent annelid *Alvinella pompejana*. Appl. Environ. Microbiol. *63*, 1124–1130.

Cary, S.C., Shank, T., and Stein, J. (1998). Worms bask in extreme temperatures. Nature *391*, 545.

Castro, M.S., and Fontes, W. (2005). Plant defense and antimicrobial peptides. Protein Pept. Lett. *12*, 11–16.

Cavanaugh, C.M., McKiness, Z.P., Newton, I.L., and Stewart, F.J. (2006). Marine chemosynthetic symbioses. In The Prokaryotes, (Springer), pp. 475–507.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet *2*, e64.

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

Charlet, M., Chernysh, S., Philippe, H., Hetru, C., Hoffmann, J.A., and Bulet, P. (1996). Innate Immunity isolation of several cysteine-rich antimicrobial peptides from the blood of a mollusc, *Mytilus edulis*. J. Biol. Chem. *271*, 21808–21813.

Cherry, R., Desbruyeres, D., Heyraud, M., and Nolan, C. (1992). High levels of natural radioactivity in hydrothermal vent polychaetes. Comptes Rendus Académie Sci. Sér. 3 Sci. Vie *315*, 21–26.

Chevaldonne, P., Jollivet, D., Vangriesheim, A., and Desbruyères, D. (1997). Hydrothermal-vent Alvinellid polychaete dispersal in the eastern Pacific. 1. Influence of vent site distribution, bottom currents, and biological patterns. Limnol. Oceanogr. *42*, 67–80.

Chevaldonné, P., Fisher, C.R., Childress, J.J., Desbruyères, D., Jollivet, D., Zal, F., and Toulmond, A. (2000). Thermotolerance and the “Pompeii worms.” Mar. Ecol. Prog. Ser. *208*, 293–295.

Cho, J.H., Park, C.B., Yoon, Y.G., and Kim, S.C. (1998). Lumbricin I, a novel proline-rich antimicrobial peptide from the earthworm: purification, cDNA cloning and molecular characterization. Biochim. Biophys. Acta BBA-Mol. Basis Dis. *1408*, 67–76.

Choi, K.-Y., Chow, L.N., and Mookherjee, N. (2012). Cationic host defence peptides: multifaceted role in immune modulation and inflammation. J. Innate Immun. *4*, 361–370.

Chu, H., Pazgier, M., Jung, G., Nuccio, S.-P., Castillo, P.A., De Jong, M.F., Winter, M.G., Winter, S.E., Wehkamp, J., Shen, B., et al. (2012). Human α-defensin 6 promotes mucosal innate immunity through self-assembled peptide nanonets. Science *337*, 477–481.

Clark, A.G., and Wang, L. (1997). Molecular population genetics of *Drosophila* immune system genes. Genetics *147*, 713–724.

Clark, P.L., and Elcock, A.H. (2016). Molecular chaperones: providing a safe place to weather a midlife protein-folding crisis. Nat. Struct. Mol. Biol. *23*, 621–623.

Coen, E., Strachan, T., and Dover, G. (1982a). Dynamics of concerted evolution of ribosomal DNA and histone gene families in the melanogaster species subgroup of *Drosophila*. J. Mol. Biol. *158*, 17–35.

Coen, E.S., Thoday, J.M., and Dover, G. (1982b). Rate of turnover of structural variants in the rDNA gene family of *Drosophila melanogaster*. Nature *295*, 564–568.

Cottin, D., Shillito, B., Chertemps, T., Thatje, S., Léger, N., and Ravaux, J. (2010). Comparison of heat-shock responses between the hydrothermal vent shrimp *Rimicaris exoculata* and the related coastal shrimp *Palaemonetes varians*. J. Exp. Mar. Biol. Ecol. *393*, 9–16.

Crovella, S., Antcheva, N., Zelezetsky, I., Boniotto, M., Pacor, S., Falzacappa, M.V., and Tossi, A. (2005). Primate β-defensins-structure, function and evolution. Curr. Protein Pept. Sci. *6*, 7–21.

Cuthbertson, B.J., Shepard, E.F., Chapman, R.W., and Gross, P.S. (2002). Diversity of the penaeidin antimicrobial peptides in two shrimp species. Immunogenetics *54*, 442–445.

Cuthbertson, B.J., Shepard, E.F., Chapman, R.W., and Gross, P.S. Diversity of the penaeidin antimicrobial peptides in two shrimp species. Immunogenetics *54*, 442–445.

Cuvillier-Hot, V., Boidin-Wichlacz, C., and Tasiemski, A. (2014). Polychaetes as annelid models to study ecoimmunology of marine organisms. J. Mar. Sci. Technol. *22*, 9–14.

Cuvillier-Hot, V., Gaudron, S.M., Massol, F., Boidin-Wichlacz, C., Pennel, T., Lesven, L., Net, S., Papot, C., Ravaux, J., and Vekemans, X. (2017). Immune failure reveals vulnerability of populations exposed to pollution in the bioindicator species *Hediste diversicolor*. Sci. Total Environ. 16p.

Dai, Y.-J., Wang, Y.-Q., Zhang, Y.-H., Liu, Y., Li, J., Wei, S., Zhao, L.-J., Zhou, Y., Lin, L., and Lan, J.-F. (2017). The role of ficolin-like protein (PcFLP1) in the antibacterial immunity of red swamp crayfish (*Procambarus clarkii*). Mol. Immunol. *81*, 26–34.

Danovaro, R., Snelgrove, P.V.R., and Tyler, P. (2014). Challenging the paradigms of deep-sea ecology. Trends Ecol. Evol. *29*, 465–475.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods *9*, 772–772.

Dawkins, R., and Krebs, J.R. (1979). Arms races between and within species. Proc. R. Soc. Lond. B Biol. Sci. *205*, 489–511.

De Bakker, P.I., McVean, G., Sabeti, P.C., Miretti, M.M., Green, T., Marchini, J., Ke, X., Monsuur, A.J., Whittaker, P., Delgado, M., et al. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. Nat. Genet. *38*, 1166–1172.

Decaestecker, E., Gaba, S., Raeymaekers, J.A., Stoks, R., Van Kerckhoven, L., Ebert, D., and De Meester, L. (2007). Host-parasite 'Red Queen' dynamics archived in pond sediment. Nature *450*, 870.

DelValls, T.Á., Forja, J.M., and Gómez-Parra, A. (2002). Seasonality of contamination, toxicity, and quality values in sediments from littoral ecosystems in the Gulf of Cádiz (SW Spain). Chemosphere *46*, 1033–1043.

Demuynck, S., and Dhainaut-Courtois, N. (1993). Identification of extracellular haemoglobin as the major high molecular weight cadmium-binding protein of the polychaete annelid *Nereis diversicolor*. Comp. Biochem. Physiol. C Pharmacol. Toxicol. Endocrinol. *106*, 467–472.

Demuynck, S., and Dhainaut-Courtois, N. (1994). Metal-protein binding patterns in the polychaete worm *Nereis diversicolor* during short-term acute cadmium stress. Comp. Biochem. Physiol. C Pharmacol. Toxicol. Endocrinol. *108*, 59–64.

Demuynck, S., Li, K.W., Schors, R., and Dhainaut-Courtois, N. (1993). Amino acid sequence of the small cadmium-binding protein (MP II) from *Nereis diversicolor* (annelida, polychaeta). Eur. J. Biochem. *217*, 151–156.

Deng, C., Cheng, C.-H.C., Ye, H., He, X., and Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. Proc. Natl. Acad. Sci. *107*, 21593–21598.

Derycke, S., De Meester, N., Rigaux, A., Creer, S., Bik, H., Thomas, W.K., and Moens, T. (2016). Coexisting cryptic species of the *Litoditis marina* complex (Nematoda) show differential resource use and have distinct microbiomes with high intraspecific variability. Mol. Ecol. *25*,2093–110.

Desbruyeres, D., and Laubier, L. (1980). *Alvinella pompejana* gen. sp. nov., Ampharetidae aberrant des sources hydrothermales de la ride Est-Pacifique. Oceanol. Acta *3*, 267–274.

Desbruyeres, D., and Laubier, L. (1991). Systematics, phylogeny, ecology and distribution of the Alvinellidae (Polychaeta) from deep-sea hydrothermal vents. Ophelia *5*, 31–45.

Desbruyeres, D., Chevaldonné, P., Alayse, A.-M., Jollivet, D., Lallier, F.H., Jouin-Toulmond, C., Zal, F., Sarradin, P.-M., Cosson, R., Caprais, J.-C., et al. (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 Res. Part II Top. Stud. Oceanogr. *45*, 383–422.

Desbruyères, D., Hashimoto, J., and Fabri, M.-C. (2006). Composition and biogeography of hydrothermal vent communities in Western Pacific back-arc basins. Back-Arc Spreading Syst. Geol. Biol. Chem. Phys. Interact. 215–234.

Dhainaut, A., and Scaps, P. (2001). Immune defense and biological responses induced by toxics in Annelida. Can. J. Zool. *79*, 233–253.

Diamond, G., Beckloff, N., Weinberg, A., and Kisich, K.O. (2009). The Roles of antimicrobial peptides in innate host defense. Curr. Pharm. Des. *15*, 2377–2392.

Dirix, C., Meersman, F., MacPhee, C.E., Dobson, C.M., and Heremans, K. (2005). High hydrostatic pressure dissociates early aggregates of TTR 105–115, but not the mature amyloid fibrils. J. Mol. Biol. *347*, 903–909.

Dobson, C.M. (2003). Protein folding and misfolding. Nature *426*, 884–890.

Dorschner, R.A., Pestonjamasp, V.K., Tamakuwala, S., Ohtake, T., Rudisill, J., Nizet, V., Agerberth, B., Gudmundsson, G.H., and Gallo, R.L. (2001). Cutaneous injury induces the release of

cathelicidin anti-microbial peptides active against group A Streptococcus. J. Invest. Dermatol. *117*, 91–97.

Doyle, J., and Doyle, J.L. (1987). Genomic plant DNA preparation from fresh tissue-CTAB method. Phytochem Bull *19*, 11–15.

Drummond, A.J., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Heled, J., Kearse, M., Moir, R., Stones-Havas, S., Sturrock, S., et al. (2010). Geneious v6. 1.6. Website Http://www.geneious. .com.

Du Pasquier, L. (2006). Germline and somatic diversification of immune recognition elements in Metazoa. Immunol. Lett. *104*, 2–17.

Dubilier, N., Bergin, C., and Lott, C. (2008). Symbiotic diversity in marine animals: the art of harnessing chemosynthesis. Nat. Rev. Microbiol. *6*, 725–740.

Duda, T.F., Vanhoye, D., and Nicolas, P. (2002). Roles of diversifying selection and coordinated evolution in the evolution of amphibian antimicrobial peptides. Mol. Biol. Evol. *19*, 858–864.

Duhig, N.C., Stolz, J., Davidson, G.J., and Large, R.R. (1992). Cambrian microbial and silica gel textures in silica iron exhalites from the Mount Windsor volcanic belt, Australia; their petrography, chemistry, and origin. Econ. Geol. *87*, 764–784.

Dybdahl, M.F., and Lively, C.M. (1998). Host-parasite coevolution: evidence for rare advantage and time-lagged selection in a natural population. Evolution *52*, 1057–1066.

Edmonds, H.N., Michael, P.J., Baker, E.T., Connelly, D.P., Snow, J.E., Langmuir, C.H., Dick, H.J.B., Mühe, R., German, C.R., and Graham, D.W. (2003). Discovery of abundant hydrothermal venting on the ultraslow-spreading Gakkel ridge in the Arctic Ocean. Nature *421*, 252–256.

Ehlers, J., and Gibbard, P.L. (2007). The extent and chronology of Cenozoic global glaciation. Quat. Int. *164*, 6–20.

Einfeldt, A.L., Doucet, J.R., and Addison, J.A. (2014). Phylogeography and cryptic introduction of the ragworm *Hediste diversicolor* (Annelida, Nereididae) in the Northwest Atlantic. Invertebr. Biol. *133*, 232–241.

Ejsmond, M.J., and Radwan, J. (2015). Red Queen processes drive positive selection on major histocompatibility complex (MHC) genes. PLoS Comput Biol *11*, e1004627.

Elder Jr, J.F., and Turner, B.J. (1995). Concerted evolution of repetitive DNA sequences in eukaryotes. Q. Rev. Biol. *70*, 297–320.

Eldon, B., Riquet, F., Yearsley, J., Jollivet, D., and Broquet, T. (2016). Current hypotheses to explain genetic chaos under the sea. Curr. Zool. *62*, 551–566.

Ellis, R.J. (2006). Molecular chaperones: assisting assembly in addition to folding. Trends Biochem. Sci. *31*, 395–401.

Ellis, R.J., and Van der Vies, S.M. (1991). Molecular chaperones. Annu. Rev. Biochem. *60*, 321–347.

Erler, S., Lhomme, P., Rasmont, P., and Lattorff, H.M.G. (2014). Rapid evolution of antimicrobial peptide genes in an insect host–social parasite system. Infect. Genet. Evol. *23*, 129–137.

Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol. Bioinforma. Online *1*, 47.

Fändrich, M. (2007). On the structural definition of amyloid fibrils and other polypeptide aggregates. Cell. Mol. Life Sci. *64*, 2066–2078.

Fawcett, J.A., and Innan, H. (2011). Neutral and Non-Neutral Evolution of Duplicated Genes with Gene Conversion. Genes *2*, 191–209.

Feder, M.E., and Hofmann, G.E. (1999). Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. Annu. Rev. Physiol. *61*, 243–282.

Fisher, C., and Skibinski, D.O.F. (1990). Sex-biased mitochondrial DNA heteroplasmy in the marine mussel *Mytilus*. Proc. R. Soc. Lond. B Biol. Sci. *242*, 149–156.

Fisher, C.R., Takai, K., and Le Bris, N. (2007). Hydrothermal vent ecosystems. Oceanography *20*, 14–23.

Fontanillas, E., Galzitskaya, O.V., Lecompte, O., Lobanov, M.Y., Tanguy, A., Mary, J., Girguis, P.R., Hourdez, S., and 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 Biol. Evol. *9*, 279–296.

Franzenburg, S., Walter, J., Künzel, S., Wang, J., Baines, J.F., Bosch, T.C.G., and Fraune, S. (2013). Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proc. Natl. Acad. Sci. *110*, E3730–E3738.

Fraune, S., and Bosch, T.C. (2010). Why bacteria matter in animal development and evolution. Bioessays *32*, 571–580.

Frohm, M., Agerberth, B., Ahangari, G., St\a ahle-Bäckdahl, M., Lidén, S., Wigzell, H., and Gudmundsson, G.H. (1997). The expression of the gene coding for the antibacterial peptide LL-37 is induced in human keratinocytes during inflammatory disorders. J. Biol. Chem. *272*, 15258–15263.

Froy, O., and Gurevitz, M. (2003). Arthropod and mollusk defensins–evolution by exon-shuffling. TRENDS Genet. *19*, 684–687.

Fu, Y.-X., and Li, W.-H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.

Fugère, N., Brousseau, P., Krzystyniak, K., Coderre, D., and Fournier, M. (1996). Heavy metal-specific inhibition of phagocytosis and different in vitro sensitivity of heterogeneous coelomocytes from Lumbricusterrestris (Oligochaeta). Toxicology *109*, 157–166.

Fuller, C.A., Postava-Davignon, M.A., West, A., and Rosengaus, R.B. (2011). Environmental conditions and their impact on immunocompetence and pathogen susceptibility of the Caribbean termite *Nasutitermes acajutlae*. Ecol. Entomol. *36*, 459–470.

Funk, D.J., and Omland, K.E. (2003). Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 397–423.

Galinier, R., Roger, E., Sautiere, P.-E., Aumelas, A., Banaigs, B., and Mitta, G. (2009). Halocyntin and papillosin, two new antimicrobial peptides isolated from hemocytes of the solitary tunicate, *Halocynthia papillosa*. J. Pept. Sci. *15*, 48–55.

Gamenick, I., Abbiati, M., and Giere, O. (1998). Field distribution and sulphide tolerance of *Capitella capitata* (Annelida: Polychaeta) around shallow water hydrothermal vents off Milos (Aegean Sea). A new sibling species? Mar. Biol. *130*, 447–453.

Ganz, T., Selsted, M.E., Szklarek, D., Harwig, S.S., Daher, K., Bainton, D.F., and Lehrer, R.I. (1985). Defensins. Natural peptide antibiotics of human neutrophils. J. Clin. Invest. *76*, 1427–1435.

Gestal, C., Costa, M., Figueras, A., and Novoa, B. (2007). Analysis of differentially expressed genes in response to bacterial stimulation in hemocytes of the carpet-shell clam *Ruditapes decussatus*: identification of new antimicrobial peptides. Gene *406*, 134–143.

Ghosh, J., Lun, C.M., Majeske, A.J., Sacchi, S., Schrankel, C.S., and Smith, L.C. (2011). Invertebrate immune diversity. Dev. Comp. Immunol. *35*, 959–974.

Giangrande, A., Licciano, M., and Musco, L. (2005). Polychaetes as environmental indicators revisited. Mar. Pollut. Bull. *50*, 1153–1162.

Goffredi, S.K. (2010). Indigenous ectosymbiotic bacteria associated with diverse hydrothermal vent invertebrates. Environ. Microbiol. Rep. *2*, 479–488.

Goldson, A.J., Hughes, R.N., and Gliddon, C.J. (2001). Population genetic consequences of larval dispersal mode and hydrography: a case study with bryozoans. Mar. Biol. *138*, 1037–1042.

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

Goudet, J. (1995). FSTAT (version 1.2): a computer program to calculate F-statistics. J. Hered. *86*, 485–486.

Gouy, M., Guindon, S., and Gascuel, O. (2009). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. *27*, 221–224.

Grassle, J. (1980). Polychaete sibling species. In Aquatic Oligochaete Biology, R.O. Brinkhurst, and D.G. Cook, eds. (Springer US), pp. 25–32.

Grassle, J.P., and Grassle, J.F. (1976). Sibling species in the marine pollution indicator *Capitella* (Polychaeta). Science *192*, 567–569.

Grimaldi, A., Girardello, R., Malagoli, D., Falabella, P., Tettamanti, G., Valvassori, R., Ottaviani, E., and De Eguileor, M. (2012). Amyloid/Melanin distinctive mark in invertebrate immunity. ISJ *9*, 153–162.

Groenendijk, D., Lücker, S.M., Plans, M., Kraak, M.H., and Admiraal, W. (2002). Dynamics of metal adaptation in riverine chironomids. Environ. Pollut. *117*, 101–109.

Grzymski, J.J., Murray, A.E., Campbell, B.J., Kaplarevic, M., Gao, G.R., Lee, C., Daniel, R., Ghadiri, A., Feldman, R.A., and Cary, S.C. (2008). Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. Proc. Natl. Acad. Sci. *105*, 17516–17521.

Guo, Y., Shen, Y.-H., Sun, W., Kishino, H., Xiang, Z.-H., and Zhang, Z. (2011). Nucleotide diversity and selection signature in the domesticated silkworm, *Bombyx mori*, and wild silkworm, *Bombyx mandarina*. J. Insect Sci. *11*.

Halldórsdóttir, K., and Árnason, E. (2015). Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species. PeerJ *3*, e976.

Hanington, P.C., and Zhang, S.-M. (2011). The primary role of fibrinogen-related proteins in invertebrates is defense, not coagulation. J. Innate Immun. *3*, 17–27.

Harris, F., Dennison, S.R., and Phoenix, D.A. (2009). Anionic antimicrobial peptides from eukaryotic organisms. Curr. Protein Pept. Sci. *10*, 585–606.

Harrison, R.G. (1990). Hybrid zones: windows on evolutionary process. Oxf. Surv. Evol. Biol. *7*, 69–128.

Harrison, R.G. (1993). Hybrid zones and the evolutionary process (Oxford University Press on Demand).

Harrison, R.G., and Larson, E.L. (2014). Hybridization, introgression, and the nature of species boundaries. J. Hered. *105*, 795–809.

Hauton, C., Brockton, V., and Smith, V.J. (2006). Cloning of a crustin-like, single whey-acidic-domain, antibacterial peptide from the haemocytes of the European lobster, *Homarus gammarus*, and its response to infection with bacteria. Mol. Immunol. *43*, 1490–1496.

Hebert, P.D., Ratnasingham, S., and de Waard, J.R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. Lond. B Biol. Sci. *270*, S96–S99.

Hedlund, J., Johansson, J., and Persson, B. (2009). BRICHOS-a superfamily of multidomain proteins with diverse functions. BMC Res. Notes *2*, 180.

Hedrick, P.W. (2013). Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. *22*, 4606–4618.

Hellberg, M.E. (2009). Gene flow and isolation among populations of marine animals. Annu. Rev. Ecol. Evol. Syst. *40*.

Hellgren, O., and Sheldon, B.C. (2011). Locus-specific protocol for nine different innate immune genes (antimicrobial peptides: β-defensins) across passerine bird species reveals within-species coding variation and a case of trans-species polymorphisms. Mol. Ecol. Resour. *11*, 686–692.

Hewitt, G.M. (2004). Genetic consequences of climatic oscillations in the Quaternary. Philos. Trans. R. Soc. Lond. B Biol. Sci. *359*, 183–195.

Hilliard, J., Hajduk, M., and Schulze, A. (2016). Species delineation in the *Capitella* species complex (Annelida: Capitellidae): geographic and genetic variation in the northern Gulf of Mexico. Invertebr. Biol. *135*, 415–422.

Hollox, E.J., and Armour, J.A. (2008). Directional and balancing selection in human beta-defensins. BMC Evol. Biol. *8*, 113.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, C.C., Wright, M.W., et al. (2004). Gene map of the extended human MHC. Nat. Rev. Genet. *5*, 889–899.

Hou, F., Gao, T., Liu, T., Jia, Z., Liu, Y., Sun, C., and Liu, X. (2016). Identification of 10 transcripts of FREP in penaeid shrimp *Litopenaeus vannamei*. Fish Shellfish Immunol. *58*, 436–441.

Hourdez, S., and Lallier, F.H. (2007). Adaptations to hypoxia in hydrothermal-vent and cold-seep invertebrates. Rev. Environ. Sci. Biotechnol. *6*, 143–159.

Huang, B., Zhang, L., Li, L., Tang, X., and Zhang, G. (2015). Highly diverse fibrinogen-related proteins in the Pacific oyster *Crassostrea gigas*. Fish Shellfish Immunol. *43*, 485–490.

Hudson, R.R., and Kaplan, N.L. (1988). The coalescent process in models with selection and recombination. Genetics *120*, 831–840.

Hudson, R.R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. Genetics *116*, 153–159.

Hughes, A.L., and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature *335*, 167–170.

Hughes, A.L., and Yeager, M. (1997a). Coordinated amino acid changes in the evolution of mammalian defensins. J. Mol. Evol. *44*, 675–682.

Hughes, A.L., and Yeager, M. (1997b). Molecular evolution of the vertebrate immune system. BioEssays *19*, 777–786.

Hughes, A.L., and Yeager, M. (1998). Natural selection at Major Histocompatibility Complex loci of vertebrates. Annu. Rev. Genet. *32*, 415–435.

Hughes, A.L., Hughes, M.K., and Watkins, D.I. (1993). Contrasting roles of interallelic recombination at the HLA-A and HLA-B loci. Genetics *133*, 669–680.

Hung, C.-W., Jung, S., Grötzinger, J., Gelhaus, C., Leippe, M., and Tholey, A. (2014). Determination of disulfide linkages in antimicrobial peptides of the macin family by combination of top-down and bottom-up proteomics. J. Proteomics *103*, 216–226.

Hurtado, L.A., Lutz, R.A., and Vrijenhoek, R.C. (2004). Distinct patterns of genetic differentiation among annelids of eastern Pacific hydrothermal vents. Mol. Ecol. *13*, 2603–2615.

Huson, D.H. (1998). SplitsTree: analyzing and visualizing evolutionary data. Bioinforma. Oxf. Engl. *14*, 68–73.

Hutchings, P. (1998). Biodiversity and functioning of polychaetes in benthic sediments. Biodivers. Conserv. *7*, 1133–1145.

Hymon, R.M., Koski, R.A., and Sinclair, C. (1984). Fossils of hydrothermal vent worms from Cretaceous sulfide ores of the Samail Ophiolite, Oman. Science *223*, 1407–1410.

Jacobs, D.K., and Lindberg, D.R. (1998). Oxygen and evolutionary patterns in the sea: onshore/offshore trends and recent recruitment of deep-sea faunas. Proc. Natl. Acad. Sci. *95*, 9396–9401.

Jang, H., Ma, B., Lal, R., and Nussinov, R. (2008). Models of toxic β-sheet channels of protegrin-1 suggest a common subunit organization motif shared with toxic Alzheimer β-amyloid ion channels. Biophys. J. *95*, 4631–4642.

Jang, H., Arce, F.T., Mustata, M., Ramachandran, S., Capone, R., Nussinov, R., and Lal, R. (2011). Antimicrobial protegrin-1 forms amyloid-like fibrils with rapid kinetics suggesting a functional link. Biophys. J. *100*, 1775–1783.

Jang, S.-J., Park, E., Lee, W.-K., Johnson, S.B., Vrijenhoek, R.C., and Won, Y.-J. (2016). Population subdivision of hydrothermal vent polychaete *Alvinella pompejana* across equatorial and Easter Microplate boundaries. BMC Evol. Biol. *16*, 235.

Jang, W.S., Kim, K.N., Lee, Y.S., Nam, M.H., and Lee, I.H. (2002). Halocidin: a new antimicrobial peptide from hemocytes of the solitary tunicate, *Halocynthia aurantium*. FEBS Lett. *521*, 81–86.

Jiggins, F.M., and Kim, K.-W. (2005). The Evolution of Antifungal Peptides in Drosophila. Genetics *171*, 1847–1859.

Jollivet, D., Mary, J., Gagnière, N., Tanguy, A., Fontanillas, E., Boutet, I., Hourdez, S., Segurens, B., Weissenbach, J., Poch, O., et al. (2012). Proteome adaptation to high temperatures in the ectothermic hydrothermal vent Pompeii worm. PLOS ONE *7*, e31150.

Jolly, M.T., Viard, F., Gentil, F., Thiébaut, E., and Jollivet, D. (2006). Comparative phylogeography of two coastal polychaete tubeworms in the Northeast Atlantic supports shared history and vicariant events. Mol. Ecol. *15*, 1841–1855.

Jouin-Toulmond, C., Zal, F., and Hourdez, S. (1997). Genital apparatus and ultrastructure of the spermatozoa in *Alvinella pompejana* (Annelida: Polychaeta). Cah. Biol. Mar. *38*, 128-129

Jung, S., Dingley, A.J., Augustin, R., Anton-Erxleben, F., Stanisak, M., Gelhaus, C., Gutsmann, T., Hammer, M.U., Podschun, R., Bonvin, A.M., et al. (2009). Hydramacin-1, structure and antibacterial activity of a protein from the basal metazoan *Hydra*. J. Biol. Chem. *284*, 1896–1905.

Kagan, B.L., Jang, H., Capone, R., Arce, F.T., Ramachandran, S., Lal, R., and Nussinov, R. (2012). Antimicrobial properties of amyloid peptides. Mol. Pharm. *9*, 708.

Kato, Y., Aizawa, T., Hoshino, H., Kawano, K., Nitta, K., and Zhang, H. (2002). abf-1 and abf-2, ASABF-type antimicrobial peptide genes in *Caenorhabditis elegans*. Biochem. J. *361*, 221–230.

Kawahara, M., Kato, M., and Kuroda, Y. (2001). Effects of aluminum on the neurotoxicity of primary cultured neurons and on the aggregation of β-amyloid protein. Brain Res. Bull. *55*, 211–217.

Kelley, J., Walter, L., and Trowsdale, J. (2005). Comparative genomics of major histocompatibility complexes. Immunogenetics *56*, 683–695.

Kemppainen, P., Panova, M., Hollander, J., and Johannesson, K. (2009). Complete lack of mitochondrial divergence between two species of NE Atlantic marine intertidal gastropods. J. Evol. Biol. *22*, 2000–2011.

Key Jr, M.M., Jeffries, W.B., Voris, H.K., and Yang, C.M. (1996). Epizoic bryozoans, horseshoe crabs, and other mobile benthic substrates. Bull. Mar. Sci. *58*, 368–384.

Kiel, S., and Tyler, P.A. (2010). Chemosynthetically-driven ecosystems in the Deep Sea. In The Vent and Seep Biota, (Springer), pp. 1–14.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide-sequences. J. Mol. Evol. *16*, 111–120.

Knowlton, N. (1993). Sibling species in the sea. Annu. Rev. Ecol. Syst. 189–216.

Koechlin, N., and Grasset, M. (1988). Silver contamination in the marine polychaete annelid *Sabella pavonina* S.: A cytological and analytical study. Mar. Environ. Res. *26*, 249–263.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B Biol. Sci.

Kondrashov, F.A., Rogozin, I.B., Wolf, Y.I., and Koonin, E.V. (2002). Selection in the evolution of gene duplications. Genome Biol. *3*, 0008.1–0008.9.

Kormas, K.A., Tivey, M.K., Von Damm, K., and Teske, A. (2006). Bacterial and archaeal phylotypes associated with distinct mineralogical layers of a white smoker spire from a deep-sea hydrothermal vent site (9° N, East Pacific Rise). Environ. Microbiol. *8*, 909–920.

Kreitman, M., and Hudson, R.R. (1991). Inferring the evolutionary histories of the Adh and Adh-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics *127*, 565–582.

Lagabrielle, Y. (2005). La dorsale est-Pacifique entre 10 et 20 S. Alternance du volcanisme et de la tectonique le long de la zone active axiale. Géomorphologie Relief Process. Environ. *11*, 105–120.

Lai, Y., and Gallo, R.L. (2009). AMPed up immunity: how antimicrobial peptides have multiple roles in immune defense. Trends Immunol. *30*, 131–141.

Lamberty, M., Zachary, D., Lanot, R., Bordereau, C., Robert, A., Hoffmann, J.A., and Bulet, P. (2001). Insect immunity constitutive expression of a cysteine-rich antifungal and a linear antibacterial peptide in a termite insect. J. Biol. Chem. *276*, 4085–4092.

Lardicci, C., and Ceccherelli, G. (1994). Dinamica di popolazione di una specie del complesso Capitella capitata in un piccolo bacino salmastro dell'Isola dell'Elba (Società italiana di biologia marina).

Lassègues, M., Roch, P., and Valembois, P. (1989). Antibacterial activity of *Eisenia fetida* andrei coelomic fluid: Evidence, induction, and animal protection. J. Invertebr. Pathol. *53*, 1–6.

Last, N.B., and Miranker, A.D. (2013). Common mechanism unites membrane poration by amyloid and antimicrobial peptides. Proc. Natl. Acad. Sci. *110*, 6382–6387.

Lazzaro, B.P. (2008). Natural selection on the Drosophila antimicrobial immune system. Curr. Opin. Microbiol. *11*, 284–289.

Lazzaro, B.P., and Clark, A.G. (2001). Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the attacin genes of *Drosophila melanogaster*. Genetics *159*, 659–671.

Lazzaro, B.P., and Clark, A.G. (2003). Molecular population genetics of inducible antibacterial peptide genes in *Drosophila melanogaster*. Mol. Biol. Evol. *20*, 914–923.

Le Bris, N., and Gaill, F. (2006). How does the annelid *Alvinella pompejana* deal with an extreme hydrothermal environment? In Life in Extreme Environments, (Springer), pp. 315–339.

Lehrer, R.I., Selsted, M.E., Szklarek, D., and Fleischmann, J. (1983). Antibacterial activity of microbicidal cationic proteins 1 and 2, natural peptide antibiotics of rabbit lung macrophages. Infect. Immun. *42*, 10–14.

Lehrer, R.I., Tincu, J.A., Taylor, S.W., Menzel, L.P., and Waring, A.J. (2003). Natural peptide antibiotics from tunicates: structures, functions and potential uses. Integr. Comp. Biol. *43*, 313–322.

Leigh, J.W., and Bryant, D. (2015). Popart: full-feature software for haplotype network construction. Methods Ecol. Evol. *6*, 1110–1116.

Lenormand, T., Guillemaud, T., Bourguet, D., and Raymond, M. (1998). Appearance and sweep of a gene duplication: adaptive response and potential for new functions in the mosquito *Culex pipiens*. Evolution *52*, 1705–1712.

Leulier, F., Parquet, C., Pili-Floury, S., Ryu, J.-H., Caroff, M., Lee, W.-J., Mengin-Lecreulx, D., and Lemaitre, B. (2003). The Drosophila immune system detects bacteria through specific peptidoglycan recognition. Nat. Immunol. *4*, 478–484.

Librado, P., and Rozas, J. (2009). DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics *25*, 1451–1452.

Little, C.T., and Vrijenhoek, R.C. (2003). Are hydrothermal vent animals living fossils? Trends Ecol. Evol. *18*, 582–588.

Little, T.J., and Cobbe, N. (2005). The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan-and a thioester-recognizing protein. Insect Mol. Biol. *14*, 599–605.

Little, C.T., Danelian, T., Herrington, R.J., and Haymon, R.M. (2004). Early Jurassic hydrothermal vent community from the Franciscan Complex, California. J. Paleontol. *78*, 542–559.

Login, F.H., and Heddi, A. (2013). Insect immune system maintains long-term resident bacteria through a local response. J. Insect Physiol. *59*, 232–239.

Login, F.H., Balmand, S., Vallier, A., Vincent-Monegat, C., Vigneron, A., Weiss-Gayet, M., Rochat, D., and Heddi, A. (2011). Antimicrobial peptides keep insect endosymbionts under control. Science *334*, 362–365.

López-García, B., Lee, P.H.A., Yamasaki, K., and Gallo, R.L. (2005). Anti-fungal activity of cathelicidins and their potential role in *Candida albicans* skin infection. J. Invest. Dermatol. *125*, 108–115.

Lozupone, C.A., and Knight, R. (2007). Global patterns in bacterial diversity. Proc. Natl. Acad. Sci. *104*, 11436–11440.

Lupski, J.R., and Stankiewicz, P. (2005). Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. PLoS Genet. *1*, e49.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155.

Lynn, D.J., Lloyd, A.T., Fares, M.A., and O'Farrelly, C. (2004a). Evidence of positively selected sites in mammalian α-defensins. Mol. Biol. Evol. *21*, 819–827.

Lynn, D.J., Higgs, R., Gaines, S., Tierney, J., James, T., Lloyd, A.T., Fares, M.A., Mulcahy, G., and O'Farrelly, C. (2004b). Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. Immunogenetics *56*, 170–177.

Maggs, C.A., Castilho, R., Foltz, D., Henzler, C., Jolly, M.T., Kelly, J., Olsen, J., Perez, K.E., Stam, W., Väinölä, R., et al. (2008). Evaluating signatures of glacial refugia for North Atlantic benthic marine taxa. Ecology *89*, S108–S122.

Mallet, J. (2005). Hybridization as an invasion of the genome. Trends Ecol. Evol. *20*, 229–237.

Mallet, J. (2007). Hybrid speciation. Nature *446*, 279–283.

Maltseva, A.L., Kotenko, O.N., Kokryakov, V.N., Starunov, V.V., and Krasnodembskaya, A.D. (2014). Expression pattern of arenicins—the antimicrobial peptides of polychaete *Arenicola marina*. Front. Physiol. *5*.

Mantyh, P.W., Ghilardi, J.R., Rogers, S., DeMaster, E., Allen, C.J., Stimson, E.R., and Maggio, J.E. (1993). Aluminum, iron, and zinc ions promote aggregation of physiological concentrations of β-amyloid peptide. J. Neurochem. *61*, 1171–1174.

Marchand, J., Leignel, V., Moreau, B., and Chénais, B. (2009). Characterization and sequence analysis of manganese superoxide dismutases from Brachyura (Crustacea: Decapoda): hydrothermal bythograeidae versus littoral crabs. Comp. Biochem. Physiol. B Biochem. Mol. Biol. *153*, 191–199.

Martin, D., and Rybicki, E. (2000). RDP: detection of recombination amongst aligned sequences. Bioinformatics *16*, 562–563.

Martoglio, B., and Dobberstein, B. (1998). Signal sequences: more than just greasy peptides. Trends Cell Biol. *8*, 410–415.

Masson, F., Zaidman-Rémy, A., and Heddi, A. (2016). Antimicrobial peptides and cell processes tracking endosymbiont dynamics. Phil Trans R Soc B *371*, 20150298.

Maxwell, A.I., Morrison, G.M., and Dorin, J.R. (2003). Rapid sequence divergence in mammalian β-defensins by adaptive evolution. Mol. Immunol. *40*, 413–421.

McDonald, J.H., Kreitman, M., and others (1991). Adaptive protein evolution at the Adh locus in Drosophila. Nature *351*, 652–654.

Medzhitov, R., and Janeway Jr, C. (2000). Innate immunity. N. Engl. J. Med. *343*, 338–344.

Meersman, F., and Dobson, C.M. (2006). Probing the pressure–temperature stability of amyloid fibrils provides new insights into their molecular properties. Biochim. Biophys. Acta BBA-Proteins Proteomics *1764*, 452–460.

Meersman, F., Smeller, L., and Heremans, K. (2006). Protein stability and dynamics in the pressure–temperature plane. Biochim. Biophys. Acta BBA-Proteins Proteomics *1764*, 346–354.

Meiklejohn, C.D., Montooth, K.L., and Rand, D.M. (2007). Positive and negative selection on the mitochondrial genome. Trends Genet. *23*, 259–263.

Méndez, N. (2006). Life cycle of *Capitella sp. Y* (Polychaeta: Capitellidae) from Estero del Yugo, Mazatlan, Mexico. J. Mar. Biol. Assoc. U. K. *86*, 263–269.

Méndez, N., Romero, J., and Flos, J. (1997). Population dynamics and production of the polychaete *Capitella capitata* in the littoral zone of Barcelona (Spain, NW Mediterranean). J. Exp. Mar. Biol. Ecol. *218*, 263–284.

Méndez, N., Linke-Gamenick, I., and Forbes, V.E. (2000). Variability in reproductive mode and larval development within the *Capitella capitata* species complex. Invertebr. Reprod. Dev. *38*, 131–142.

Michaelson, D., Rayner, J., Couto, M., and Ganz, T. (1992). Cationic defensins arise from charge-neutralized propeptides: a mechanism for avoiding leukocyte autocytotoxicity? J. Leukoc. Biol. *51*, 634–639.

Mitta, G., Vandenbulcke, F., Hubert, F., and Roch, P. (1999). Mussel defensins are synthesised and processed in granulocytes then released into the plasma after bacterial challenge. J Cell Sci *112*, 4233–4242.

Mitta, G., Vandenbulcke, F., Hubert, F., Salzet, M., and Roch, P. (2000a). Involvement of Mytilins in Mussel antimicrobial defense. J. Biol. Chem. *275*, 12954–12962.

Mitta, G., Hubert, F., Dyrynda, E.A., Boudry, P., and Roch, P. (2000b). Mytilin B and MGD2, two antimicrobial peptides of marine mussels: gene structure and expression analysis. Dev. Comp. Immunol. *24*, 381–393.

Mitta, G., Vandenbulcke, F., and Roch, P. (2000c). Original involvement of antimicrobial peptides in mussel innate immunity. FEBS Lett. *486*, 185–190.

Miura, T., Suzuki, K., Kohata, N., and Takeuchi, H. (2000). Metal binding modes of Alzheimer's amyloid β-peptide in insoluble aggregates and soluble complexes. Biochemistry (Mosc.) *39*, 7024–7031.

Moalic, Y., Desbruyères, D., Duarte, C.M., Rozenfeld, A.F., Bachraty, C., and Arnaud-Haond, S. (2011). Biogeography revisited with network theory: retracing the history of hydrothermal vent communities. Syst. Biol. *61*, 127–137.

Muths, D., Davoult, D., Gentil, F., and Jollivet, D. (2006). Incomplete cryptic speciation between intertidal and subtidal morphs of *Acrocnida brachiata* (Echinodermata: Ophiuroidea) in the Northeast Atlantic. Mol. Ecol. *15*, 3303–3318.

Muths, D., Davoult, D., Jolly, M.T., Gentil, F., and 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*, 667–679.

Mylonakis, E., Podsiadlowski, L., Muhammed, M., and Vilcinskas, A. (2016). Diversity, evolution and medical applications of insect antimicrobial peptides. Phil Trans R Soc B *371*, 20150290.

Nei, M. (1987). Molecular evolutionary genetics (Columbia university press).

Nguyen, L.T., Haney, E.F., and Vogel, H.J. (2011). The expanding scope of antimicrobial peptide structures and their modes of action. Trends Biotechnol. *29*, 464–472.

Nielsen, R., and Wakeley, J. (2001). Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics *158*, 885–896.

Nilsson, M.R. (2004). Techniques to study amyloid fibril formation in vitro. Methods *34*, 151–160.

Nizet, V., Ohtake, T., Lauth, X., Trowbridge, J., Rudisill, J., Dorschner, R.A., Pestonjamasp, V., Piraino, J., Huttner, K., and Gallo, R.L. (2001). Innate antimicrobial peptide protects the skin from invasive bacterial infection. Nature *414*, 454–457.

Nygren, A. (2014). Cryptic polychaete diversity: a review. Zool. Scr. *43*, 172–183.

Oard, S.V., and Enright, F.M. (2006). Expression of the antimicrobial peptides in plants to control phytopathogenic bacteria and fungi. Plant Cell Rep. *25*, 561–572.

Obbard, D.J., Gordon, K.H.J., Buck, A.H., and Jiggins, F.M. (2009). The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. B Biol. Sci. *364*, 99–115.

Ohno, S. (1970). Evolution by gene duplication. xv-+ 160 pp.

Oliveira, D.C., Raychoudhury, R., Lavrov, D.V., and Werren, J.H. (2008). Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp Nasonia (Hymenoptera: Pteromalidae). Mol. Biol. Evol. *25*, 2167–2180.

Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiol. Mol. Biol. Rev. *75*, 361–422.

Oren, Z., and Shai, Y. (1998). Mode of action of linear amphipathic α-helical antimicrobial peptides. Pept. Sci. *47*, 451–463.

Osaki, T., Omotezako, M., Nagayama, R., Hirata, M., Iwanaga, S., Kasahara, J., Hattori, J., Ito, I., Sugiyama, H., and Kawabata, S. (1999). Horseshoe Crab hemocyte-derived antimicrobial polypeptides, tachystatins, with sequence similarity to spider neurotoxins. J. Biol. Chem. *274*, 26172–26178.

Ota, T., Sitnikova, T., and Nei, M. (2000). Evolution of vertebrate immunoglobulin variable gene segments. In Origin and Evolution of the Vertebrate Immune System, P.D.L.D. Pasquier, and P.G.W.L. M.D, eds. (Springer Berlin Heidelberg), pp. 221–245.

Ott, J. (1996). Sulphide ectosymbioses in shallow marine habitats Mar Ecol Prog Ser 11: 369–382

Otti, O., Tragust, S., and Feldhaar, H. (2014). Unifying external and internal immune defences. Trends Ecol. Evol. *29*, 625–634.

Oudhoff, M.J., Bolscher, J.G., Nazmi, K., Kalay, H., van't Hof, W., Amerongen, A.V.N., and Veerman, E.C. (2008). Histatins are the major wound-closure stimulating factors in human saliva as identified in a cell culture assay. FASEB J. *22*, 3805–3812.

Ovchinnikova, T.V., Aleshina, G.M., Balandin, S.V., Krasnosdembskaya, A.D., Markelov, M.L., Frolova, E.I., Leonova, Y.F., Tagaev, A.A., Krasnodembsky, E.G., and Kokryakov, V.N. (2004). Purification and primary structure of two isoforms of arenicin, a novel antimicrobial peptide from marine polychaeta *Arenicola marina*. FEBS Lett. *577*, 209–214.

Padhi, A., and Verghese, B. (2008). Molecular diversity and evolution of myticin-C antimicrobial peptide variants in the Mediterranean mussel, *Mytilus galloprovincialis*. Peptides *29*, 1094–1101.

Pan, W., Liu, X., Ge, F., and Zheng, T. (2003). Reconfirmation of antimicrobial activity in the coelomic fluid of the earthworm *Eisenia fetida andrei* by colorimetric assay. J. Biosci. *28*, 723–731.

Pan, W., Liu, X., Ge, F., Han, J., and Zheng, T. (2004). Perinerin, a novel antimicrobial peptide purified from the clamworm *Perinereis aibuhitensis* Grube and its partial characterization. J. Biochem. (Tokyo) *135*, 297–304.

Papot, C., Cascella, K., Toullec, J.-Y., and 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. Ecol. Evol. *6*, 1555–1575.

Papot, C., Massol, F., Jollivet, D., and Tasiemski, A. (2017). Antagonistic evolution of an antibiotic and its molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat. Sci. Rep. *7.* 1454-1460.

Parsons, T.J., Olson, S.L., and Braun, M.J. (1993). Unidirectional spread of secondary sexual plumage traits across an avian hybrid zone. Sciences. *260*, 1643–1643.

Pees, B., Yang, W., Zárate-Potes, A., Schulenburg, H., and Dierking, K. (2015). High innate immune specificity through diversified C-type lectin-like domain proteins in invertebrates. J. Innate Immun. *8*, 129–142.

Peng, S., Fitzen, M., Jörnvall, H., and Johansson, J. (2010). The extracellular domain of Bri2 (ITM2B) binds the ABri peptide (1–23) and amyloid β-peptide (Aβ1–40): Implications for Bri2 effects on processing of amyloid precursor protein and Aβ aggregation. Biochem. Biophys. Res. Commun. *393*, 356–361.

Pérez-Portela, R., Arranz, V., Rius, M., and Turon, X. (2013). Cryptic speciation or global spread? The case of a cosmopolitan marine invertebrate with limited dispersal capabilities. Sci. Rep. *3*, 3197.

Piccino, P., Viard, F., Sarradin, P., Le Bris, N., Le Guen, D., and Jollivet, D. (2004). Thermal selection of PGM allozymes in newly founded populations of the thermotolerant vent polychaete *Alvinella pompejana*. Proc. R. Soc. Lond. B Biol. Sci. *271*, 2351–2359.

Pinho, C., and Hey, J. (2010). Divergence with gene flow: models and data. Annu. Rev. Ecol. Evol. Syst. *41*, 215–230.

Plouviez, S., Shank, T.M., Faure, B., Daguin-Thiebaut, C., Viard, F., Lallier, F.H., and Jollivet, D. (2009). Comparative phylogeography among hydrothermal vent species along the East Pacific Rise reveals vicariant processes and population expansion in the South. Mol. Ecol. *18*, 3903–3917.

Plouviez, S., Le Guen, D., Lecompte, O., Lallier, F.H., and 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 Evol. Biol. *10*, 220.

Posada, D., and Crandall, K.A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. *98*, 13757–13762.

Pouny, Y., Rapaport, D., Mor, A., Nicolas, P., and Shai, Y. (1992). Interaction of antimicrobial dermaseptin and its fluorescently labeled analogs with phospholipid membranes. Biochemistry (Mosc.) *31*, 12416–12423.

Powell, M.A., and Somero, G.N. (1986). Adaptations to sulfide by hydrothermal vent animals: sites and mechanisms of detoxification and metabolism. Biol. Bull. *171*, 274–290.

Pradillon, F., and Gaill, F. (2003). Oogenesis characteristics in the hydrothermal vent polychaete *Alvinella pompejana*. Invertebr. Reprod. Dev. *43*, 223–235.

Pradillon, F., Zbinden, M., Mullineaux, L.S., and Gaill, F. (2005). Colonisation of newly-opened habitat by a pioneer species, *Alvinella pompejana* (Polychaeta: Alvinellidae), at East Pacific Rise vent sites. Mar. Ecol. Prog. Ser. *302*, 147–157.

Provan, J., and Bennett, K.D. (2008). Phylogeographic insights into cryptic glacial refugia. Trends Ecol. Evol. *23*, 564–571.

Prunier, J., Caron, S., and MacKay, J. (2017). CNVs into the wild: screening the genomes of conifer trees (Picea spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. BMC Genomics *18*, 97.

Puillandre, N., Lambert, A., Brouillet, S., and Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol. Ecol. *21*, 1864–1877.

Radding, C.M. (1978). Genetic recombination: strand transfer and mismatch repair. Annu. Rev. Biochem. *47*, 847–880.

Radford, S.E. (2000). Protein folding: progress made and promises ahead. Trends Biochem. Sci. *25*, 611–618.

Rahnamaeian, M., Cytryńska, M., Zdybicka-Barabas, A., Dobslaff, K., Wiesner, J., Twyman, R.M., Zuchner, T., Sadd, B.M., Regoes, R.R., Schmid-Hempel, P., et al. (2015). Insect antimicrobial peptides show potentiating functional interactions against Gram-negative bacteria. Proc R Soc B *282*, 20150293.

Raj, P.A., Antonyraj, K.J., and Karunakaran, T. (2000). Large-scale synthesis and functional elements for the antimicrobial activity of defensins. Biochem. J. *347*, 633–641.

Ramos-Onsins, S., and Aguadé, M. (1998). Molecular evolution of the Cecropin multigene family in Drosophila: functional genes vs. pseudogenes. Genetics *150*, 157–171.

Ramos-Onsins, S.E., and Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. Mol. Biol. Evol. *19*, 2092–2100.

Ravaux, J., Hamel, G., Zbinden, M., Tasiemski, A.A., Boutet, I., Léger, N., Tanguy, A., Jollivet, D., and Shillito, B. (2013). Thermal limit for metazoan life in question: in vivo heat tolerance of the Pompeii worm. PLoS One *8*, e64074.

Relf, J.M., Chisholm, J.R.S., Kemp, G.D., and Smith, V.J. (1999). Purification and characterization of a cysteine-rich 11.5-kDa antibacterial protein from the granular haemocytes of the shore crab, *Carcinus maenas*. Eur. J. Biochem. *264*, 350–357.

Ribardière, A., Daguin-Thiébaut, C., Houbin, C., Coudret, J., Broudin, C., Timsit, O., and Broquet, T. (2017). Geographically distinct patterns of reproductive isolation and hybridization in two sympatric species of the *Jaera albifrons* complex (marine isopods). Ecol. Evol.

Rogers, A.D., Johnston, N.M., Murphy, E.J., and Clarke, A. (2012). Antarctic ecosystems: an extreme environment in a changing world (John Wiley & Sons).

Rokitskaya, T.I., Kolodkin, N.I., Kotova, E.A., and Antonenko, Y.N. (2011). Indolicidin action on membrane permeability: carrier mechanism versus pore formation. Biochim. Biophys. Acta *1808*, 91–97.

Rolff, J., and Schmid-Hempel, P. (2016). Perspectives on the evolutionary ecology of arthropod antimicrobial peptides. Phil Trans R Soc B *371*, 20150297.

Rousset, V., Pleijel, F., Rouse, G.W., Erséus, C., and Siddall, M.E. (2007). A molecular phylogeny of annelids. Cladistics *23*, 41–63.

Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., and Clark, A.G. (2007a). Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. *39*, 1461–1468.

Sackton, T.B., Lazzaro, B.P., Schlenke, T.A., Evans, J.D., Hultmark, D., and Clark, A.G. (2007b). Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. *39*, 1461–1468.

Salman, V., Amann, R., Girnth, A.-C., Polerecky, L., Bailey, J.V., Høgslund, S., Jessen, G., Pantoja, S., and Schulz-Vogt, H.N. (2011). A single-cell sequencing approach to the classification of large, vacuolated sulfur bacteria. Syst. Appl. Microbiol. *34*, 243–259.

Salzman, N.H., Hung, K., Haribhai, D., Chu, H., Karlsson-Sjöberg, J., Amir, E., Teggatz, P., Barman, M., Hayward, M., Eastwood, D., et al. (2010). Enteric defensins are essential regulators of intestinal microbial ecology. Nat. Immunol. *11*, 76–83.

Samakovlis, C., Kylsten, P., Kimbrell, D.A., Engström, A., and Hultmark, D. (1991). The andropin gene and its product, a male-specific antibacterial peptide in *Drosophila melanogaster*. EMBO J. *10*, 163–169.

Sánchez-Pulido, L., Devos, D., and Valencia, A. (2002). BRICHOS: a conserved domain in proteins associated with dementia, respiratory distress and cancer. Trends Biochem. Sci. *27*, 329–332.

Sarradin, P.-M., Caprais, J.-C., Riso, R., Kerouel, R., and Aminot, A. (1999). Chemical environment of the hydrothermal mussel communities in the Lucky Strike and Menez Gwen vent fields, Mid Atlantic ridge. Cah. Biol. Mar. *40*, 93–104.

Sasahara, K., Naiki, H., and Goto, Y. (2005). Kinetically controlled thermal response of β 2-microglobulin amyloid fibrils. J. Mol. Biol. *352*, 700–711.

Schauber, J., and Gallo, R.L. (2008). Antimicrobial peptides and the skin immune defense system. J. Allergy Clin. Immunol. *122*, 261–266.

Schikorski, D., Cuvillier-Hot, V., Leippe, M., Boidin-Wichlacz, C., Slomianny, C., Macagno, E., Salzet, M., and Tasiemski, A. (2008). Microbial challenge promotes the regenerative process of the injured central nervous system of the medicinal leech by inducing the synthesis of antimicrobial peptides in neurons and microglia. J. Immunol. *181*, 1083–1095.

Schlenke, T.A., and Begun, D.J. (2003). Natural selection drives Drosophila immune system evolution. Genetics *164*, 1471–1480.

Schmitt, P., Gueguen, Y., Desmarais, E., Bachère, E., and De Lorgeril, J. (2010). Molecular diversity of antimicrobial effectors in the oyster *Crassostrea gigas*. BMC Evol. Biol. *10*, 23.

Schrenk, M.O., Kelley, D.S., Delaney, J.R., and Baross, J.A. (2003). Incidence and diversity of microorganisms within the walls of an active deep-sea sulfide chimney. Appl. Environ. Microbiol. *69*, 3580–3592.

Schulenburg, H., Boehnisch, C., and Michiels, N.K. (2007). How do invertebrates generate a highly specific innate immune response? Mol. Immunol. *44*, 3338–3344.

Schulz, H.N. (2006). The genus Thiomargarita. In The Prokaryotes, (Springer), pp. 1156–1163.

Schutte, B.C., Mitros, J.P., Bartlett, J.A., Walters, J.D., Jia, H.P., Welsh, M.J., Casavant, T.L., and McCray, P.B. (2002). Discovery of five conserved β-defensin gene clusters using a computational search strategy. Proc. Natl. Acad. Sci. *99*, 2129–2133.

Scocchi, M., Wang, S., Gennaro, R., and Zanetti, M. (1998). Cloning and analysis of a transcript derived from two contiguous genes of the cathelicidin family. Biochim. Biophys. Acta BBA - Gene Struct. Expr. *1398*, 393–396.

Scriven, J.J., Whitehorn, P.R., Goulson, D., Tinsley, M., and others (2016). Niche partitioning in a sympatric cryptic species complex. Ecol. Evol.

Seaver, E.C. (2016). Annelid models I: *Capitella teleta*. Curr. Opin. Genet. Dev. *39*, 35–41.

Selsted, M.E., and Ouellette, A.J. (2005). Mammalian defensins in the antimicrobial immune response. Nat. Immunol. *6*, 551–557.

Semple, C.A., Rolfe, M., and Dorin, J.R. (2003). Duplication and selection in the evolution of primate b-defensin genes. Genome Biol *4*, R31.

Senderovich, Y., and Halpern, M. (2013). The protective role of endogenous bacterial communities in chironomid egg masses and larvae. ISME J. *7*, 2147–2158.

Seo, J.-K., Nam, B.-H., Go, H.-J., Jeong, M., Lee, K.-Y., Cho, S.-M., Lee, I.-A., and Park, N.G. (2016). Hemerythrin-related antimicrobial peptide, msHemerycin, purified from the body of the Lugworm, *Marphysa sanguinea*. Fish Shellfish Immunol. *57*, 49–59.

Siebenaller, J., and Somero, G.N. (1978). Pressure-adaptive differences in lactate dehydrogenases of congeneric fishes living at different depths. Science *201*, 255–257.

Silva, C.F., Shimabukuro, M., Alfaro-Lucas, J.M., Fujiwara, Y., Sumida, P.Y., and Amaral, A.C. (2016). A new *Capitella* polychaete worm (Annelida: Capitellidae) living inside whale bones in the abyssal South Atlantic. Deep Sea Res. Part Oceanogr. Res. Pap. *108*, 23–31.

Silva, C.F., Seixas, V.C., Barroso, R., Domenico, M.D., Amaral, A.C.Z., and Paiva, P.C. (2017). Demystifying the *Capitella capitata* complex (Annelida, Capitellidae) diversity by morphological and molecular data along the Brazilian coast. PLOS ONE *12*, e0177760.

Simmaco, M., Mignogna, G., and Barra, D. (1998). Antimicrobial peptides from amphibian skin: What do they tell us? Pept. Sci. *47*, 435–450.

Sipe, J.D., and Cohen, A.S. (2000). Review: history of the amyloid fibril. J. Struct. Biol. *130*, 88–98.

Smith, V.J., Fernandes, J.M., Kemp, G.D., and Hauton, C. (2008). Crustins: enigmatic WAP domain-containing antibacterial proteins from crustaceans. Dev. Comp. Immunol. *32*, 758–772.

Soehnlein, O. (2009). Direct and alternative antimicrobial mechanisms of neutrophil-derived granule proteins. J. Mol. Med. *87*, 1157–1164.

Song, H., Buhay, J.E., Whiting, M.F., and Crandall, K.A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. *105*, 13486–13491.

Soscia, S.J., Kirby, J.E., Washicosky, K.J., Tucker, S.M., Ingelsson, M., Hyman, B., Burton, M.A., Goldstein, L.E., Duong, S., Tanzi, R.E., et al. (2010). The Alzheimer's disease-associated amyloid β-protein is an antimicrobial peptide. PloS One *5*, e9505.

Southward, A.J., and Southward, E.C. (1978). Recolonization of rocky shores in Cornwall after use of toxic dispersants to clean up the Torrey Canyon spill. J. Fish. Board Can. *35*, 682–706.

Sperstad, S.V., Haug, T., Blencke, H.-M., Styrvold, O.B., Li, C., and Stensvåg, K. (2011). Antimicrobial peptides from marine invertebrates: Challenges and perspectives in marine antimicrobial peptide discovery. Biotechnol. Adv. *29*, 519–530.

Spurgin, L.G., and Richardson, D.S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. R. Soc. Lond. B Biol. Sci. rspb20092084.

Su, C., and Nei, M. (1999). Fifty-million-year-old polymorphism at an immunoglobulin variable region gene locus in the rabbit evolutionary lineage. Proc. Natl. Acad. Sci. *96*, 9710–9715.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics *105*, 437–460.

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

Takahata, N., and Nei, M. (1990). Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics *124*, 967–978.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. *24*, 1596–1599.

Tang, M., and Hong, M. (2009). Structure and mechanism of β-hairpin antimicrobial peptides in lipid bilayers from solid-state NMR spectroscopy. Mol. Biosyst. *5*, 317–322.

Tasiemski, A., Vandenbulcke, F., Mitta, G., Lemoine, J., Lefebvre, C., Sautiere, P.-E., and Salzet, M. (2004). Molecular characterization of two novel antibacterial peptides inducible upon bacterial challenge in an annelid, the leech *Theromyzon tessulatum*. J. Biol. Chem. *279*, 30973–30982.

Tasiemski, A., Schikorski, D., Le Marrec-Croq, F., Pontoire-Van Camp, C., Boidin-Wichlacz, C., and Sautière, P.-E. (2007). Hedistin: A novel antimicrobial peptide containing bromotryptophan constitutively expressed in the NK cells-like of the marine annelid, *Nereis diversicolor*. Dev. Comp. Immunol. *31*, 749–762.

Tasiemski, A., Jung, S., Boidin-Wichlacz, C., Jollivet, D., Cuvillier-Hot, V., Pradillon, F., Vetriani, C., Hecht, O., Sönnichsen, F.D., Gelhaus, C., et al. (2014). Characterization and Function of the First Antibiotic Isolated from a Vent Organism: The Extremophile Metazoan *Alvinella pompejana*. PLOS ONE *9*, e95737.

Tasiemski, A., Massol, F., Cuvillier-Hot, V., Boidin-Wichlacz, C., Roger, E., Rodet, F., Fournier, I., Thomas, F., and Salzet, M. (2015). Reciprocal immune benefit based on complementary production of antibiotics by the leech *Hirudo verbana* and its gut symbiont *Aeromonas veronii*. Sci. Rep. *5*.

Tassanakajon, A., Somboonwiwat, K., and Amparyup, P. (2015). Sequence diversity and evolution of antimicrobial peptides in invertebrates. Dev. Comp. Immunol. *48*, 324–341.

Taylor, S.W., Craig, A.G., Fischer, W.H., Park, M., and Lehrer, R.I. (2000). Styelin D, an Extensively Modified Antimicrobial Peptide from Ascidian Hemocytes. J. Biol. Chem. *275*, 38417–38426.

Tennessen, J.A. (2005). Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. J. Evol. Biol. *18*, 1387–1394.

Tennessen, J.A., and Blouin, M.S. (2007). Selection for antimicrobial peptide diversity in frogs leads to gene duplication and low allelic variation. J. Mol. Evol. *65*, 605–615.

Tennessen, J.A., and Blouin, M.S. (2008). Balancing selection at a frog antimicrobial peptide locus: fluctuating immune effector alleles? Mol. Biol. Evol. *25*, 2669–2680.

Thiermann, F., Akoumianaki, I., Hughes, J.A., and Giere, O. (1997). Benthic fauna of a shallow-water gaseohydrothermal vent area in the Aegean Sea (Milos, Greece). Mar. Biol. *128*, 149–159.

Tomioka, S., Kondoh, T., Sato-Okoshi, W., Ito, K., Kakui, K., and Kajihara, H. (2016). Cosmopolitan or cryptic species? A case study of *Capitella teleta* (Annelida: Capitellidae). Zoolog. Sci. *33*, 545–554.

Trowsdale, J., and Parham, P. (2004). Mini-review: Defense strategies and immunity-related genes. Eur. J. Immunol. *34*, 7–17.

Tsutsumi, H. (1987). Population dynamics of *Capitella capitata* (Polychaeta; Capitellidae) in an organically polluted cove. Mar. Ecol. Prog. Ser. 139–149.

Tsutsumi, H., and Kikuchi, T. (1984). Study of the life history of *Capitella capitata* (Polychaeta: Capitellidae) in Amakusa, South Japan including a comparison with other geographical regions. Mar. Biol. *80*, 315–321.

Tsutsumi, H., Fukunaga, S., Fujita, N., and Sumida, M. (1990). Relationship between growth of *Capitella sp*. Org. Enrich. Sediment Mar. Ecol. Prog. Ser. *63*, 157–162.

Tsutsumi, H., Wainright, S., Montani, S., Saga, M., Ichihara, S., and Kogure, K. (2001). Exploitation of a chemosynthetic food resource by the polychaete *Capitella sp. I*. Mar. Ecol. Prog. Ser. *216*, 119–127.

Tunnicliffe, V., McArthur, A.G., and McHugh, D. (1998). A biogeographical perspective of the deep-sea hydrothermal vent fauna. Adv. Mar. Biol. *34*, 353–442.

Tyler, P.A., German, C.R., Ramirez-Llodra, E., and Van Dover, C.L. (2002). Understanding the biogeography of chemosynthetic ecosystems. Oceanol. Acta *25*, 227–241.

Tzou, P., Ohresser, S., Ferrandon, D., Capovilla, M., Reichhart, J.-M., Lemaitre, B., Hoffmann, J.A., and Imler, J.-L. (2000). Tissue-specific inducible expression of antimicrobial peptide genes in *Drosophila* surface epithelia. Immunity *13*, 737–748.

Ueno, S., Kusaka, K., Tamada, Y., Minaba, M., Zhang, H., Wang, P.-C., and Kato, Y. (2008). Anionic C-terminal proregion of nematode antimicrobial peptide cecropin P4 precursor inhibits antimicrobial activity of the mature peptide. Biosci. Biotechnol. Biochem. *72*, 3281–3284.

Unckless, R.L., and Lazzaro, B.P. (2016). The potential for adaptive maintenance of diversity in insect antimicrobial peptides. Phil Trans R Soc B *371*, 20150291.

Unckless, R.L., Howick, V.M., and Lazzaro, B.P. (2016). Convergent balancing selection on an antimicrobial peptide in *Drosophila*. Curr. Biol. *26*, 257–262.

Van Dover, C.L. (2000). The Ecology of deep-sea hydrothermal vents (Princeton University Press, New Jersey).

Van Dover, C.L., German, C.R., Speer, K.G., Parson, L.M., and Vrijenhoek, R.C. (2002). Evolution and biogeography of deep-Sea vent and seep invertebrates. Science *295*, 1253–1257.

Vanhoye, D., Bruston, F., Nicolas, P., and Amiche, M. (2003). Antimicrobial peptides from hylid and ranin frogs originated from a 150-million-year-old ancestral precursor with a conserved signal peptide but a hypermutable antimicrobial domain. FEBS J. *270*, 2068–2081.

Vera, M., Martínez, P., Poisa-Beiro, L., Figueras, A., and Novoa, B. (2011). Genomic organization, molecular diversification, and evolution of antimicrobial peptide myticin-C genes in the mussel (*Mytilus galloprovincialis*). PloS One *6*, e24041.

Vermeij, G.J. (1991). When biotas meet: understanding biotic interchange. Science *253*, 1099–1104.

Viarengo, A., and Nott, J.A. (1993). Mechanisms of heavy metal cation homeostasis in marine invertebrates. Comp. Biochem. Physiol. Part C Comp. Pharmacol. *104*, 355–372.

Ville, P., Roch, P., Cooper, E.L., Masson, P., and Narbonne, J.-F. (1995). PCBs increase molecular-related activities (lysozyme, antibacterial, hemolysis, proteases) but inhibit macrophage-related functions (phagocytosis, wound healing) in earthworms. J. Invertebr. Pathol. *65*, 217–224.

Virgilio, M., and Abbiati, M. (2004). Habitat discontinuity and genetic structure in populations of the estuarine species *Hediste diversicolor* (Polychaeta: Nereididae). Estuar. Coast. Shelf Sci. *61*, 361–367.

Virgilio, M., Backeljau, T., and Abbiati, M. (2006). Mitochondrial DNA and allozyme patterns of *Hediste diversicolor* (Polychaeta: Nereididae): the importance of small scale genetic structuring. Mar. Ecol. Prog. Ser. *326*, 157–165.

Vogel, T.U., Evans, D.T., Urvater, J.A., O'Connori, D.H., Hughes, A.L., and Watkins, D.I. (1999). Major histocompatibility complex class I genes in primates: co-evolution with pathogens. Immunol. Rev. *167*, 327–337.

Völkel, S., and Grieshaber, M.K. (1996). Mitochondrial sulfide oxidation in *Arenicola marina*. FEBS J. *235*, 231–237.

Wang, X., Wang, X., Zhang, Y., Qu, X., and Yang, S. (2003). An antimicrobial peptide of the earthworm *Pheretima tschiliensis*: cDNA cloning, expression and immunolocalization. Biotechnol. Lett. *25*, 1317–1323.

Warren, L.M. (1976). A population study of the polychaete *Capitella capitata* at plymouth. Mar. Biol. *38*, 209–216.

Wassing, G.M., Bergman, P., Lindbom, L., and van der Does, A.M. (2015). Complexity of antimicrobial peptide regulation during pathogen–host interactions. Int. J. Antimicrob. Agents *45*, 447–454.

Watkins, D.I., McAdam, S.N., Liu, X., Strang, C.R., Milford, E.L., Levine, C.G., Garber, T.L., Dogon, A.L., Lord, C.I., Ghim, S.H., et al. (1992). New recombinant HLA-B alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. Nature *357*, 329–333.

Watson, F.L., Püttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., Rebel, V.I., and Schmucker, D. (2005). Extensive diversity of Ig-superfamily proteins in the immune system of insects. Science *309*, 1874–1878.

Watterson, G.A. (1975). On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. *7*, 256–276.

Willander, H., Hermansson, E., Johansson, J., and Presto, J. (2011). BRICHOS domain associated with lung fibrosis, dementia and cancer–a chaperone that prevents amyloid fibril formation? FEBS J. *278*, 3893–3904.

Willander, H., Askarieh, G., Landreh, M., Westermark, P., Nordling, K., Keränen, H., Hermansson, E., Hamvas, A., Nogee, L.M., Bergman, T., et al. (2012). High-resolution structure of a BRICHOS domain and its implications for anti-amyloid chaperone activity on lung surfactant protein C. Proc. Natl. Acad. Sci. *109*, 2325–2329.

Woerner, A.E., Cox, M.P., and Hammer, M.F. (2007). Recombination-filtered genomic datasets by information maximization. Bioinformatics *23*, 1851–1853.

Worobey, M., and Holmes, E.C. (1999). Evolutionary aspects of recombination in RNA viruses. J. Gen. Virol. *80*, 2535–2543.

Wright, S. (1951). The genetical structure of populations. Ann. Eugen. *15*, 323–354.

Wutzler, R., Foerster, K., and Kempenaers, B. (2012). MHC class I variation in a natural blue tit population (Cyanistes caeruleus). Genetica *140*, 349–364.

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, D., Biragyn, A., Hoover, D.M., Lubkowski, J., and Oppenheim, J.J. (2004). Multiple roles of antimicrobial defensins, cathelicidins, and eosinophil-derived neurotoxin in host defense. Annu Rev Immunol *22*, 181–215.

Yang, W.-Y., Wen, S.-Y., Huang, Y.-D., Ye, M.-Q., Deng, X.-J., Han, D., Xia, Q.-Y., and Cao, Y. (2006). Functional divergence of six isoforms of antifungal peptide Drosomycin in *Drosophila melanogaster*. Gene *379*, 26–32.

Yeung, A.T.Y., Gellatly, S.L., and Hancock, R.E.W. (2011). Multifunctional cationic host defence peptides and their clinical applications. Cell. Mol. Life Sci. *68*, 2161.

Zasloff, M. (1987). Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. Proc. Natl. Acad. Sci. *84*, 5449–5453.

Zasloff, M. (2002). Antimicrobial peptides of multicellular organisms. Nature *415*, 389–395.

Zeng, Q.-Q., He, K., Sun, D.-D., Ma, M.-Y., Ge, Y.-F., Fang, S.-G., and Wan, Q.-H. (2016). Balancing selection and recombination as evolutionary forces caused population genetic variations in golden pheasant MHC class I genes. BMC Evol. Biol. *16*. 144–149.

Zhang, L., and Gallo, R.L. (2016). Antimicrobial peptides. Curr. Biol. *26*, R14–R19.

Zhang, S.-M., and Loker, E.S. (2003). The FREP gene family in the snail *Biomphalaria glabrata*: additional members, and evidence consistent with alternative splicing and FREP retrosequences. Dev. Comp. Immunol. *27*, 175–187.

Zhang, S.-M., and Loker, E.S. (2004). Representation of an immune responsive gene family encoding fibrinogen-related proteins in the freshwater mollusc *Biomphalaria glabrata*, an intermediate host for *Schistosoma mansoni*. Gene *341*, 255–266.

Zhang, J., Sun, Q., Luan, Z., Lian, C., and Sun, L. (2017). Comparative transcriptome analysis of Rimicaris sp. reveals novel molecular features associated with survival in deep-sea hydrothermal vent. Sci. Rep. *7*, 2000.

Zhang, M., Zhao, J., and Zheng, J. (2014). Molecular understanding of a potential functional link between antimicrobial and amyloid peptides. Soft Matter *10*, 7425–7451.

Zhang, S.-M., Adema, C.M., Kepler, T.B., and Loker, E.S. (2004). Diversification of Ig superfamily genes in an invertebrate. Science *305*, 251–254.

Zhang, X., Sun, Z., Zhuo, R., Hou, Q., and Lin, G. (2001). Purification and characterization of two antibacterial peptides from *Eisenia fetida*. Sheng Wu Hua Xue Yu Sheng Wu Wu Li Jin Zhan *29*, 955–960.

Zhao, M., Wang, Y., Shen, H., Li, C., Chen, C., Luo, Z., and Wu, H. (2013). Evolution by selection, recombination, and gene duplication in MHC class I genes of two Rhacophoridae species. BMC Evol. Biol. *13*, 113.

Zierenberg, R.A., Adams, M.W., and Arp, A.J. (2000). Life in extreme environments: Hydrothermal vents. Proc. Natl. Acad. Sci. *97*, 12961–12962.

ANNEXES

Annexe 1. Caractérisation des transcrits de l'alvinellacine et mise en évidence de transcrits tronqués.

Annexe 2. Article publié dans Scientific report.

Annexe 3. Valeurs de F_{st} calculées pour les clades de la capitellacine au sein des deux régions 5' et 3' par paires de populations et de façon globale.

Annexe 4. Protocole de productions des variants du BRICHOS.

Annexe 5. Liste des articles et valorisations scientifiques de la thèse.

Annexe 1. Caractérisation des transcrits de la preproalvinellacine et mise en évidence de transcrits tronqués

Matériel et Méthode.

- **Dissection des tissus/organes d'*Alvinella pompejana***

Les animaux fraichement récoltés ont été disséqués à bord de l'Atalante sous loupe binoculaire dans du RNALater lors de la Campagne MESCAL 2 de 2012. Les organes et tissus prélevés ont été placés dans du Trizol, congelés dans de l'azote liquide puis conservés à -80°C jusqu'à l'extraction des ARN totaux au laboratoire.

- **Extraction d'ARN et synthèse d'ADNc**

Les tissus ont été broyés avec des billes de céramique puis les ARN totaux ont été extraits grâce au tampon Trizol selon le protocole du fabriquant (Trizol Reagent Invitrogen). Les ADNc ont été synthétisés en utilisant une amorce oligodT afin d'éviter les amplifications de l'ADN génomique.

Les transcrits ont été amplifiés à l'aide la Taq Uptitherm^{TM} selon le protocole décrit au chapitre 2 à l'aide des amorces 5'F et 3'R (5'F : ATGACGTATTCTGTAGTTGTGACGCTGGTC ; 3'R : CTCAGTGAAATGAAGCAGGTGAGTTATG) et visualisés sous lampe UV. Les bandes obtenues de la taille attendue ont été découpées et utilisées pour le clonage en TA cloning (TA cloning kit, INVITROGEN). Pour chaque tissu, les inserts de 32 clones blancs (positifs) ont été amplifiés à l'aide des amorces M13F et M13R et séquencés dans les 2 sens à l'aide d'un séquenceur ABI3100.

Les séquences forward et reverse de chaque clone ont été assemblées à l'aide du module « De novo Assemble » du logiciel Geneious et une séquence consensus a été produite pour chaque clone. Toutes les séquences ont ensuite été alignées et un alignement global a été réalisé et comparé à l'alignement des séquences génomiques obtenues sur le gène du précurseur protéique présenté au chapitre 2. Cette comparaison avait pour but d'assigner chaque transcrit à un paralogue précédemment caractérisé au niveau du génome d'*Alvinella pompejana*.

Résultats

La Figure 1 récapitule les relations phylogénétiques entre transcrits en ajoutant, pour la région 5', une séquence référence de chaque paralogue défini par l'étude sur l'ADN génomique. Seules des séquences diagnostiques des clades 1, 3a, 3b et 4 sont retrouvées au sein du 'pool' de transcrits amplifiés sur l'ensemble des tissus. Les séquences des paralogues 2 et 5 sont notablement absentes des séquences obtenues. Au niveau génomique, le paralogue 5 a été décrit précédemment comme étant le plus divergent, avec un intron 1 particulier, et un premier exon contenant un microsatellite dans sa région codante qui ne change pas le cadre de lecture de l'ADNc. Le paralogue 2 quant-à-lui est le seul qui possède deux délétions importantes de codons qui en font l'un des transcrits les plus courts sans affecter le cadre de lecture. Ce paralogue montre en effet une délétion de 10 codons et une délétion de 22 codons dans le deuxième exon de la région 5' sans que le cadre de lecture soit affecté (cf chapitre précédent). Les quatre autres paralogues ne possèdent ni délétion ni insertion dans les régions exoniques et montrent des niveaux de divergence moindre que ceux trouvés dans l'ADN génomique.

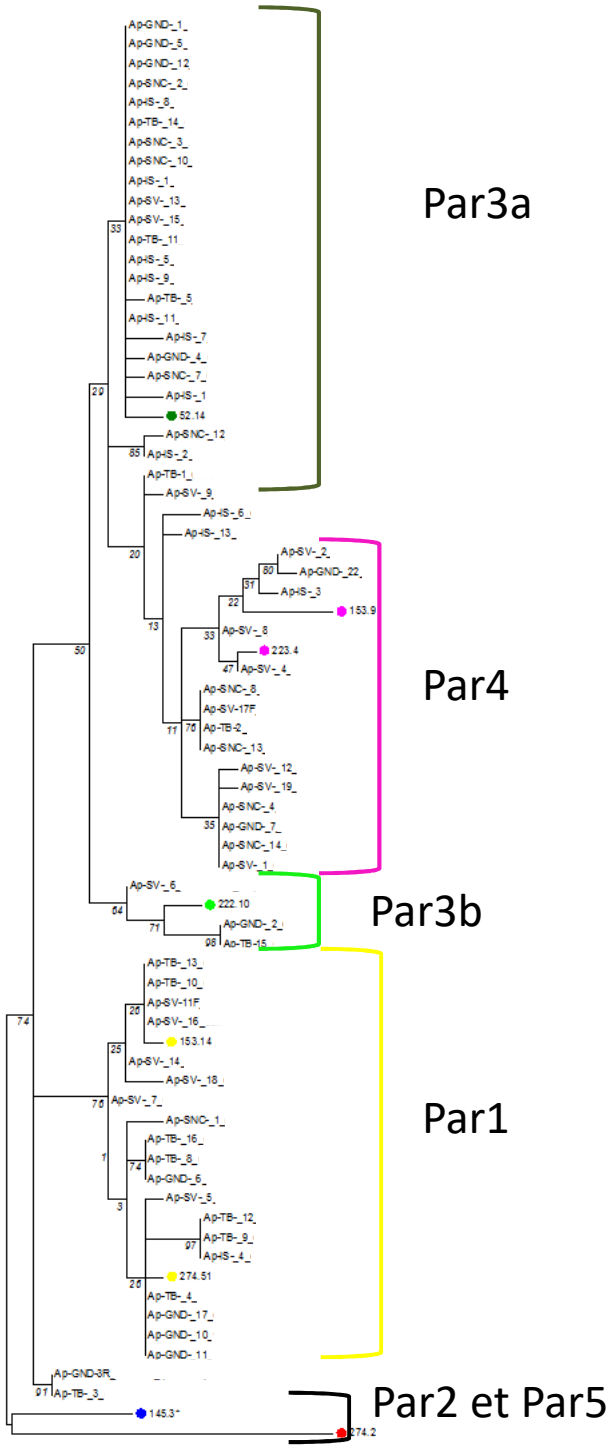


Figure 1 : Relation phylogénétique entre les transcrits de la préproalvinellacine. Les paralogues caractérisés au chapitre 2 (en région génomique 5') sont indiqués en couleur par1 : jaune, par2 et par5 : noir, 3a : vert foncé, 3b : vert clair, 4 : violet).

Identification d'autres transcrits : existence de transcrits tronqués

En plus d'identifier les transcrits dont la séquence correspond aux paralogues, cette approche a de façon inattendue, révélé d'autres messagers d'une taille plus petite que l'attendue. En effet, au cours du criblage des clones, il est apparu qu'un transcrit plus court (250 pb) avait été systématiquement amplifié. Ces amplicons de 250 pb ont donc été séquencés pour caractériser leur nature.

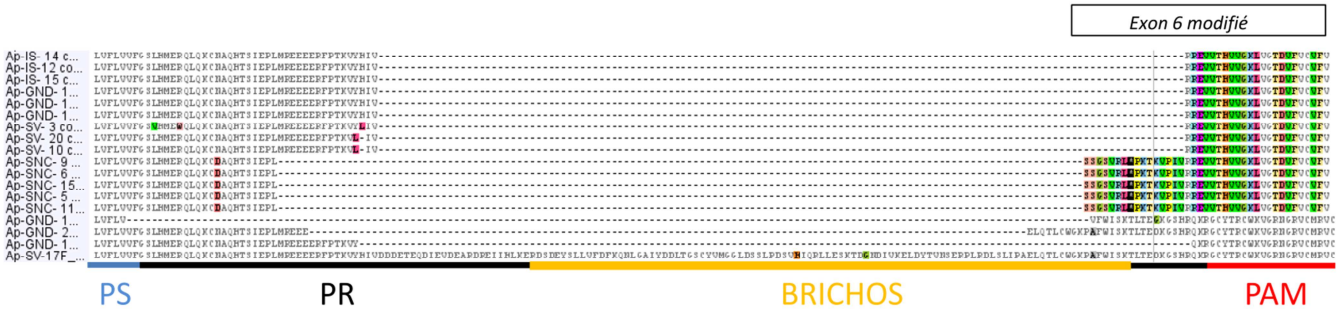


Figure 2. Alignement des transcrits tronqués le long des différentes régions du précurseur protéique et comparaison avec le transcrit originel (la dernière séquence protéique : SV-17F). En bleu : peptide signal ; en noir: prorégion, jaune: BRICHOS, rouge: PAM. La présence d'acides aminés colorés indique un changement du cadre de lecture par rapport au transcrit originel ajouté pour comparaison. La présence d'un carré noir avec une étoile indique l'apparition d'un codon stop due à un décalage du cadre de lecture. Différents tissus ont été étudiés : intestin supérieur (IS), gonade (GND), système vasculaire (SV), système nerveux central (SNC).

La Figure 2 récapitule la structure des transcrits tronqués. Les transcrits tronqués amplifiés indépendamment ont globalement le même décalage du cadre de lecture et la même délétion qui commence au niveau de l'exon 3 et se termine avant le PAM. Ceci pourrait plaider en faveur d'une contamination mais différents tissus ont été utilisés dans cette étude et au sein de chaque tissu les transcrits montrent une signature de clade (définis au chapitre 2) qui diffère entre tissu. Les transcrits tronqués amplifiés ne sont donc pas les mêmes au sein de chaque tissus bien qu'ils montrent le même patron de troncature ce qui ne va pas dans le sens d'une contamination. Dans tous les cas, c'est la région du domaine BRICHOS qui dans tous les cas est délétée avec absence ou non de la région du PAM que ce soit par décalage du cadre de lecture ou bien par apparition d'un codon stop.

Le mécanisme d'épissage alternatif bien décrit pour d'autres molécules de l'immunité chez les invertébrés ne semble ainsi pas être un mécanisme omniprésent pour créer de la diversité chez les peptides antimicrobiens. Dans notre cas, la recherche de ces transcrits sur le génome par Blastn n'a donné aucun résultat, indiquant la présence de la même séquence sur aucun scaffold du génome. Il faut néanmoins nuancer cette absence par le fait que la zone de duplications en tandem de l'alvinellacine n'a pas pu être assemblée correctement (probablement dû à la redondance des séquences répétées) ce qui pourrait être le signe d'une région avec une forte prévalence en gène dupliqués et qui pourrait *in fine* aller dans le sens de ces auteurs : ces transcrits courts pourraient être le produit de l'expression de gènes qui existent physiquement dans le génome.

Ces résultats pourraient suggérer un même patron d'épissage alternatif (délétions) au moins pour certains tissus (IS-GND-SV). D'autres transcrits avec des patrons d'indels différents existent ce qui pourrait suggérer l'existence de mécanismes d'épissage plus fréquents. Aussi, les différents transcrits tronqués montrent des mutations diagnostiques de différents clades par tissu indiquant qu'il pourrait exister un type de transcrits tronqués par tissus.

Annexe 2. Article publié dans Scientific Reports.

Antagonistic evolution of an antibiotic and its molecular chaperone : how to maintain a vital ectosymbiosis in a highly fluctuating habitats.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Antagonistic evolution of an antibiotic and its molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat

Received: 10 October 2016
Accepted: 30 March 2017
Published online: 03 May 2017

Claire Papot[1], François Massol[1], Didier Jollivet[2] & Aurélie Tasiemski[1]

Evolution of antimicrobial peptides (AMPs) has been shown to be driven by recurrent duplications and balancing/positive selection in response to new or altered bacterial pathogens. We use *Alvinella pompejana*, the most eurythermal animal known on Earth, to decipher the selection patterns acting on AMP in an ecological rather than controlled infection approach. The preproalvinellacin multigenic family presents the uniqueness to encode a molecular chaperone (BRICHOS) together with an AMP (alvinellacin) that controls the vital ectosymbiosis of *Alvinella*. In stark contrast to what is observed in the context of the Red queen paradigm, we demonstrate that exhibiting a vital and highly conserved ecto-symbiosis in the face of thermal fluctuations has led to a peculiar selective trend promoting the adaptive diversification of the molecular chaperone of the AMP, but not of the AMP itself. Because BRICHOS stabilizes beta-stranded peptides, this polymorphism likely represents an eurythermal adaptation to stabilize the structure of alvinellacin, thus hinting at its efficiency to select and control the epibiosis across the range of temperatures experienced by the worm; Our results fill some knowledge gaps concerning the function of BRICHOS in invertebrates and offer perspectives for studying immune genes in an evolutionary ecological framework.

Antimicrobial peptides (AMPs) constitute key components of the innate immune system that rapidly eradicate or incapacitate pathogenic agents such as viruses, bacteria or fungi attempting to invade and proliferate multicellular eukaryotes[1–3]. In the last decade, they have also been shown to control and confine the symbiotic microflora in specific anatomical compartments (e.g. gut, bacteriomes, skin), thus contributing to the symbiostasis of both invertebrates and vertebrates[4–8]. In metazoans, the evolution of AMPs has been shown to be driven by recurrent duplications (i.e. creation of paralogs) and balancing/positive selection in response to new and/or altered bacterial pathogens that can be encountered in a novel habitat and/or that have rapidly evolved to escape the immune response[9–12]. In terms of co-evolutionary dynamics, patterns of evolution of AMPs thus seem to generally follow a hybrid route between the matching-allele (balancing selection at a given locus) and the gene-for-gene (arms race with pathogens through gene duplications with positive diversifying selection between paralogs) paradigms of Red Queen dynamics[13]. Most empirical evidence behind this assertion comes from experimentally challenged model organisms subjected to specific controlled conditions in the laboratory and/or from data focused on the well-protected inner part of the multicellular host *i.e.* the internal immunity (inside the body *sensu lato*). Because the body acts as a wall buffering external abiotic and biotic variations, selection processes driven by environmental constraints on innate immunity can be considered to fluctuate more outside the organism than inside.

Multiple data demonstrate that AMPs not only act internally but can also be secreted into the environment surrounding an organism where they participate in external immunity, referred to as "any heritable trait acting outside of an organism improving protection from pathogens or manipulating the composition of the microbial

[1]University Lille, CNRS, UMR 8198 - Evo-Eco-Paleo, SPICI group, F-59000, Lille, France. [2]AD2M, ABICE team, Université Pierre et Marie Curie-CNRS, UMR7144, Station Biologique de Roscoff, 29682, Roscoff, France. Didier Jollivet and Aurélie Tasiemski contributed equally to this work. Correspondence and requests for materials should be addressed to A.T. (email: aurelie.tasiemski@univ-lille1.fr)

community in favor of an organism"[14]. In the case of extreme, frequently disturbed and stressful environments, the external immunity of an organism will depend on its ability to control the functioning of its externally secreted immune products under very variable conditions. In a sense, the coevolution of both the host immune system and the microbial communities in extreme environments adds another constraint to the usual Red Queen model, namely coevolution of two partners submitted to harsh selection for local adaptation to fluctuating environmental conditions, and this scenario has yet to be fully understood. Annelids are particularly suited to study the adaptation of external immunity to changing and harsh environmental conditions because (i) they are amongst the rare metazoans able to thrive in extreme and highly fluctuating habitats (e.g. hydrothermal vents, highly polluted anoxic sediments, polar environments), and (ii) they do not have barriers (i.e. exoskeleton or shell) to physically protect their skin from direct biotic/abiotic interactions. Rather than physical protection, they have developed a strong external immunity based on production of mucus and AMPs by the epidermic cells that respectively trap and kill/select pathogenic/symbiotic bacteria. In a sense, annelid defense is more comparable to that observed in amphibians or in mammals than that observed in cuticulates (i.e. arthropods and nematodes), the two most studied invertebrate phyla[8,15–17]. Polychaeta (marine worms considered as the primitive annelids) produce original AMPs, some of which are restricted to just one worm family (e.g. preproalvinellacin) or even a single species (e.g. preprohedistin). This suggests that a high AMP selection at the interspecific level has probably occurred in relation to the ecology of these organisms[18,19].

In this study, we took advantage of the peculiar microbial and physico-chemical ecology of the extremophile annelid *Alvinella pompejana*, the most eurythermal and amongst the most thermo-tolerant animals known on Earth, to decipher the selection patterns acting on an AMP, namely alvinellacin, in an evolutionary ecological framework. By being part of the external immunity of *A. pompejana*, alvinellacin is at the direct interface with abiotic and biotic constraints imposed by life in the hottest part of the deep-sea hydrothermal ecosystem. Once secreted by the epidermal cells, alvinellacin accumulates on the surface of the worm and inside its tube, thus contributing to the external immunity of the worm[19]. Through its specific bactericidal activities, it participates in the control and selection of the environmental bacteria forming the typical complex symbiotic microflora that covers the dorsal tegument of this thermophilic annelid endemic of hydrothermal chimneys along the East Pacific Rise[19]. Epibionts have been shown to supply *A. pompejana* with nutrients and to detoxify heavy metals from their habitat[20]. The combination of this epibiosis with the particular microbial environment created inside the tube allows the worm to thrive under 'hot' conditions[21,22]. In its tube, *A. pompejana* actively pumps the surrounding cold seawater to bathe in a diluted mixed fluid, which is slightly less acidic and less concentrated in free hydrogen sulfides[23–26]. This behaviour exposes the consortium of Epsilon-proteobacteria making up the epibiota of *A. pompejana* to less extreme, but still fluctuating (ranging from 7° to 84 °C), environmental conditions[27]. According to an environmental genomic study, this peculiar habitat has led to the selection of a limited number of specialized bacterial strains with greater eurythermal adaptation and metabolic flexibility[28]. One intriguing point is how natural selection has operated on the worm's external immunity to maintain this intimate and highly specific partnership present in all worms collected throughout its known geographic range (6,000 km)[28].

In this context, the main goal of this study was to determine how external immune effectors, such as the *preproalvinellacin* gene, have been selected to maintain their efficiency at selecting and controlling the eurythermal epibiotic community in an extreme and fluctuating habitat. In order to understand the functioning of this immune gene as a controller of the worm's epibiotic community, we examined (i) the levels of non-synonymous and synonymous genetic diversities over the different domains of the preproalvinellacin gene using two well-separated geographic populations of *A. pompejana*, and (ii) the divergence from its syntopic and phylogenetically close sister species *A. caudata* bearing the same epibiota. The originality of the present study lies in the search for the signature of adaptive evolution in an AMP (here alvinellacin) not in the context of pathogenicity, but rather in the context of the evolutionary constraints imposed by the obligatory maintenance of a specific, complex and vital ectosymbiosis in the face of eurythermality.

Methods

Specimen sampling. *Alvinella* spp. specimens were collected during the oceanographic cruises BIOSPEEDO (2004) and MESCAL (2012) at two geographically well separated sites (18°25.93S, 113°23.32W, 2640 m; 9°50.32N, 104°17.52W, 2550 m). Animals were collected with the manned submersible Nautile and directly flash-frozen on board (see Extended Experimental Procedures). DNA extractions were then performed using a CTAB/PVP protocol modified from[29] and previously described in ref. 30.

PCR amplification, cloning and sequencing. The whole gene encoding the preproalvinellacin (1949 bp) was previously amplified by PCR using primers specifically designed from the 5′ and 3′ UTR regions[19]. Because of its length, the gene was then sub-divided into two regions for further amplification at the population level. A detailed description of the gene and the primer design is given in the Extended Experimental Procedures (Table S1). Allelic sequences were obtained from a series of individuals of the two *Alvinella* species using the mark-recapture cloning method[31]. A detailed description of the procedures together with the primer sequences are given in the extended methods (Table S1). Sequencing was run on an ABI 3100 using BigDye© v3.0 terminator chemistry and the retrieved sequences were proofread using the Geneious De Novo Assemble module. Alignments were then performed and adjusted using the PairWise/Multiple Alignment module of the Geneious software. *In vitro* recombinants due to cloning were manually checked by searching for any abnormal combination of tag ends and removed from the dataset. Sequence datasets were then cleaned for PCR-induced allelic chimeras and artifactual singletons following a complex procedure of recombinant removal using RDP4.0 (Recombination Detecting Program) software[32]. This procedure is described in detail in the extended experimental procedures.

Paralog identification and individual genotyping. Because of duplications, a series of paralog-specific primers were designed from the cleaned sequence dataset without 'natural' recombinants (see Table S1). Allele genotyping within each locus was then performed on a subset of 16 individuals by direct sequencing of the 5′ region of the gene, in which there was enough diagnostic mutations to discriminate between the suspected paralogs (see extended experimental procedures).

Genetic diversity and neutral tests. Standard molecular diversity indices (S, $\theta\pi$, θ_w, and H_d) were estimated in *A. pompejana* for the 5′ region of the preproalvinellacin gene using the DnaSP v5.0 software[33] globally and for each putative paralog, respectively. The estimators θ_w and $\theta\pi$ (both estimating the population parameter $4N_e.\mu$ under neutral assumptions) were compared to each other using Tajima's D test and other neutrality tests such as Fu & Li's D and Fu & Li's F, which are more sensitive for detecting past demographic changes than Tajima's D test[34]. The average number of nucleotide substitutions per site (D_{xy}) was also computed between pairs of putative paralogs. Evolution of the multigenic and the within-duplicate genetic diversity ($\theta\pi$) were also estimated along the gene for the two *Alvinella* species for both the exonic and intronic regions using a sliding window (size = 50 bp; step = 10 bp) with the software DNAsp 5.0[33].

Networks and coalescence trees. A phylogenetic reconstruction of duplications was performed on sequences obtained for the 3′ exonic region of the gene for two individuals of each species and for the 5′ region of the most-recaptured individuals of *A. pompejana*. The latter tree topology was then used to map amino-acid polymorphisms on the BRICHOS domain and to reconstruct the emergence of 'natural' recombinants. In both cases, the software jModelTest 2.1.7[35] was used to select the best model of substitutions, and tree reconstructions were performed using the Maximum Likelihood method implemented in the software MEGA6[36] and PhyML 3.0[37] for the sake of comparison. The generalized time-reversible GTR+I+G model of substitutions[38–40] was then tested against the selected best models, according to Akaike (AIC) and Bayesian (BIC) information criteria, using backward hierarchical likelihood ratio tests (hLRT), and subsequently used for the tree reconstruction (see extended experimental procedures). Allelic relationships (natural recombinants included) were examined in *A. pompejana* using the neighborNet method implemented in the program SplitsTree4 software package[41] for the overall most divergent 5′ region only. This method was used as it uses reticulation to account for intergenic recombination.

Search for positive selection along the preproalvinellacin gene. To detect selection imprints on each specific domain, a majority rule consensus sequence was built from each paralog domain to perform pairwise d_N/d_S comparisons using the yn00 package of the PaML software[42]. Evolution of d_N/d_S along the gene was also assessed by calculating both the average ratio of the fixed non-synonymous to synonymous substitutions, K_a/K_s, between paralogs and the average ratio of the polymorphic non-synonymous to synonymous substitutions, π_a/π_s, within each paralog using a sliding window (size = 50 bp, step = 10 bp) with the software DNAsp 5.0[33].

Mutation mapping on phylogenetic tree. Amino-acid replacements associated with the 'hot spot' of diversity in the BRICHOS domain were mapped onto the paralog ML tree topology obtained with the software MEGA 6.0. The ML tree was subsequently used as a reference to perform a likelihood (Empirical Bayes) reconstruction of ancestral amino-acid sequences using the aaML package of the PAML software with an empirical Dayhoff matrix of amino-acid replacements.

Search for positive selection in the Propiece and BRICHOS regions. Alignments of the consensus coding sequences of paralogs of both the propiece region and its BRICHOS domain were used together with the outlier sequences of *A. caudata* to detect putative codons under positive selection using the Maximum Likelihood method implemented in the CodeML package of PaML, with the Goldman & Yang's model of codon substitutions[43]. For both alignments, the software jModelTest 2.1.7 was used to select the best model of substitutions according to the AIC and BIC. This model was then used to reconstruct the reference tree topology using PhyML 3.0. The site models M1a, M2a, M7 and M8 were subsequently compared under the assumption of no variation in the mutation rate between duplicates over time (see extended experimental procedures)[43]. In addition, Bayesian methods (NEB and BEB) of codon classification into different classes of omega were also used with a p-value threshold of 0.95 to identify positively-selected sites. Only the BEB method is robust enough to separate positively-selected from selectively-relaxed sites without uncertainty.

MacDonald-Kreitman test across paralogous genes. The MacDonald-Kreitman (MK) test was also used to detect signs of positive selection between pairs of paralogs for both the Propiece and BRICHOS domains taking advantage of their sequencing in several *A. pompejana* individuals using the module implemented in DNAsp vs 5.0. This test usually compares the ratios of non-synonymous and synonymous substitutions in the divergence (d_N/d_S) and in the polymorphism (p_N/p_S) of two closely related species but was also used to detect positive selection between pairs of paralogs[44–46]. The MK test was performed on a series of no recombining alleles to avoid excesses of non-synonymous polymorphic changes by recombination and loss of power in detecting positive selection in the paralog divergence. We also checked whether one of the assumptions of the MK test may be violated by a possible selective relaxation prior to the duplication event (see ref. 47) by performing a branch-model comparative analysis of duplicates (i.e. the "one ω ratio" M_0 vs. the "free ω ratio" M_1) using the package CodeML of the software PaML[42].

Preproalvinellacin gene induction in animals exposed to different thermal and pressure regimes. Experiments were performed onboard during the MESCAL cruise. For the pressure experiments, *Alvinella* individuals (n = 10) were transferred immediately after raising from 2500 m into the DESEARES aquarium and maintained at 250 bars for 12 h to recover from depressurization, at constant temperature. A thermal

shock experiment was also performed with a new set of isobaric BALIST equipment to retrieve and conduct experiments on worms at a constant *in situ* pressure[48]. Briefly, after recovery from the PERISCOP sampling device, *A. pompejana* specimens (n = 9) were subjected in the BALIST aquarium to three distinct thermal shocks (20, 42 and 54 °C) for 2 hours. Expression of the inducible hsp 70 and preproalvinellacin genes was quantified by RTqPCR from total RNA extracted from the experimental specimens according to the procedures detailed in ref. 48. Gene expression was normalized to expression of RPS26.

Results

Gene diversification of preproalvinellacin in the genus *Alvinella*. A first allelic screening revealed that the number of alleles per individual greatly exceeded two for both *Alvinella* species and ranged from 4 to 12 alleles according to the effort of recapture, even after correcting for singleton excesses. After removing PCR-artifactual recombinants between heterozygous gene copies, a phylogenetic reconstruction of alleles was performed using the four most recaptured individuals of the two *Alvinella* species (Fig. 1A and Fig. S1). The resulting tree displayed a reciprocal monophyly between the two species with a rather flattened shape of the coalescent, suggesting a recent and independent diversification of the gene after speciation (Fig. 1A). In both species, alleles were grouped in more than two clades for each of the four individuals tested. Tandem duplication was then supported by the size of the PCR products obtained on genomic DNA amplified with the forward and reverse preproalvinellacin primers (Fig. 1B).

Global genetic diversities were similar for the two species, but strikingly differed in intensity along the precursor gene (Fig. 2A,B). In *A. pompejana*, the gene displayed an astonishingly high nucleotidic diversity ($\theta\pi = 0.25$) in the first intron, whereas it only reached a maximum (around 0.15) in the last intron preceding the AMP coding region in *A. caudata*. Both species displayed lower genetic diversity in exons than in introns with similar levels of variation ($\theta\pi = 0.025$). The last exon, containing the AMP, was unexpectedly monomorphic in *A. pompejana* and weakly polymorphic (with three distinct variants) in *A. caudata* (Fig. S1). Finally, the two *Alvinella* AMPs differed by only one fixed replacement (Ser - > Asp) at position 198 (Fig. S1).

Identification and characterization of paralogous genes in *A. pompejana*. The genotyping of individuals with paralog-specific primers allowed us to distinguish four paralogs (1, 2, 4 and 5) with either a homozygous or heterozygous state for each individual. Paralog 3 displayed more than 3 alleles per individual and was subsequently sub-divided into par3a and par3b. Once this genotyping procedure had been performed, the sequence dataset was assembled as a reticulated network of alleles (Fig. 3A) and led to the same exact ML tree topologies (Fig. 3B) using either the GTR+I+G or the selected best models of substitutions obtained from the Bayesian informative criterion (BIC) of jModelTest (Tables S2 and S3, Fig. S2). This network, together with the distribution of alleles within each individual, indicated that preproalvinellacin is encoded by a multigenic family of at least six genes. Five distinct recombination events were robustly characterized (Fig. 3B), each of these displaying a different series of linked sites (mainly in the intronic parts of the gene) in several individuals. Three of these recombinants displayed their own set of specific mutations and represent 'old' events in the history of diversification of the gene. The number of nucleotide substitutions per site (D_{xy}) was computed between pairs of paralogs, leading to an average divergence of 0.158 (Table S4). Without the most divergent paralog par5 ($D_{xy} = 0.314$), the remaining divergence for the other pairwise comparisons was four times smaller ($D_{xy} = 0.079$). Both nucleotide ($\theta\pi$) and haplotype (H_d) diversities were high and quite variable between paralogs but did not significantly differ from neutral expectations (Table 1).

Strength of selection between domains and along the gene. Domains exhibited striking differences in terms of K_a/K_s ratios between paralogs (Fig. 2, Table S5). In the signal peptide, nearly all paralogs exhibited the same amino-acid signature with the exception of par1. Values of K_a/K_s were more heterogeneous in the propiece region and the BRICHOS domain with ratios close to or exceeding one between many pairs of paralogs, including the most divergent paralog 5/E in the BRICHOS domain. By contrast, the AMP itself displayed no K_a/K_s signal because of its lack of genetic variation.

A sharp peak of K_a/K_s between positions 360 and 410 was observed in the BRICHOS domain with a maximal value of up to 12 when looking at the polymorphism-to-divergence variation along the gene using a sliding window (Fig. 2C). In contrast, the averaged π_a/π_s across paralogs exhibited several peaks in both the peptide signal and the BRICHOS domain, but with maxima well below one, suggesting diversifying selection between the duplicated genes (Fig. 2C). McDonald-Kreitman tests failed to detect significant positive selection between pairs of paralogs either in the 5′ or the 3′ regions of the gene. Such a failure was possibly due to the high recombining rate between paralogs even if the alignments tested were devoid of recombining alleles.

Ancestral reconstruction of the BRICHOS domain and mutation mapping. Both *Alvinella* species exhibit a high number of amino-acid replacements over a small portion of the BRICHOS domain, but on distinct sites. The ancestral reconstruction of 36 BRICHOS sequences allowed us to infer whether the 'hot spot' of non-synonymous diversity was likely due to the retention of an ancestral polymorphism or to the positive fixation of some specific mutations between duplicates. At least seven polymorphic amino-acid replacements were found in the BRICHOS domain (Fig. 4). Reconstructions at ancestral nodes leading to duplicates were robust ($p > 0.99$) and indicated that three of them (N129S, T131I and D133G) were paralog-diagnostic. Reconstruction of ancestral states was similar when applying the same method on a smaller set of consensus coding sequences of duplicates using the selected best model in jModelTest (K80+I model, Fig. S3). Other replacements were randomly distributed between paralogs in terminal positions, suggesting either that these mutations reflect an ancestral polymorphism or have recently been exchanged by recombination. In the Propiece region, we evidenced

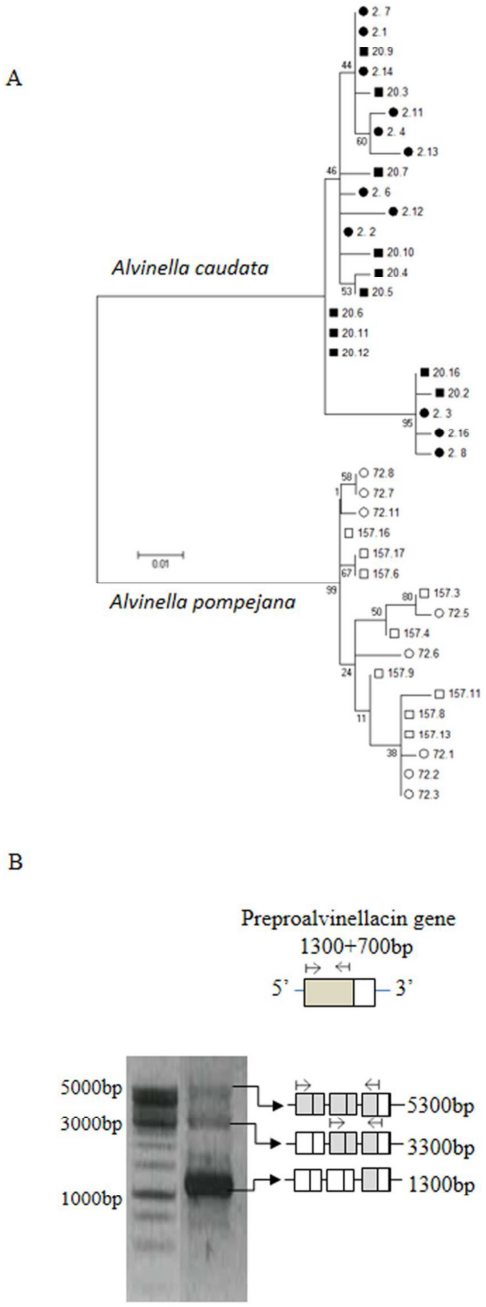


Figure 1. Gene diversification of preproalvinellacin. (**A**) Coalescence tree of alleles found in two well-recaptured individuals of *Alvinella pompejan*a (white) and its sister species *Alvinella caudata* (black), and (**B**) molecular evidence that it comes from tandemly repeated gene duplications in *A. caudata*. (**A**) Reconstruction of lineages without recombination was performed on the 3′ region on all nucleotide sites by Maximum Likelihood using the K2P model in MEGA 5.0. Allele coverage: A. *caudata*: two individuals with 15 clone recaptures each, *A. pompejana*: two individuals with 50 clone recaptures each. (**B**) Agarose gel electrophoresis with a 1 kb DNA ladder showing the amplification of the 1300 bp 5′ region of the *A. caudata* preproalvinellacin (complete gene: 2000 bp) using the 5′ primers (arrows). Longer extra bands indicate the co-amplification of two and three linked genes as summarized by the boxes representing the tandemly-duplicated gene and the position of the PCR products.

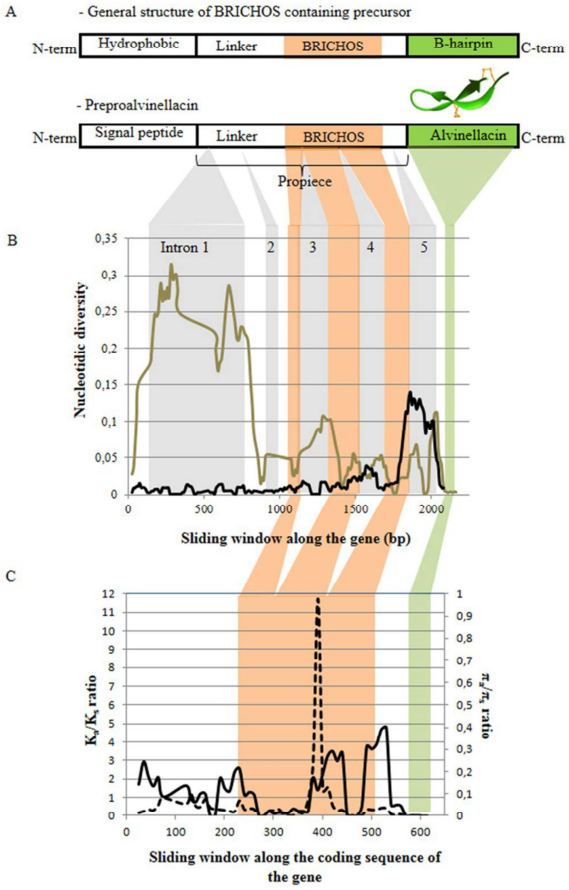


Figure 2. Evolution of genetic diversity and K_a/K_s along the preproalvinellacin gene and its corresponding coding sequence: (**A**) active alvinellacin is cleaved from a larger proteic precursor (*i.e.* preproalvinellacin). In contrast to all described AMPs, the preproalvinellacin family harbors the pattern of a BRICHOS containing protein: a hydrophobic domain (the signal peptide), a propiece with a linker and a BRICHOS domain and a C-terminal region with β-sheet propensities (alvinellacin). (**B**) Sliding window of the overall nucleotidic diversity ($\theta\pi$) along the intronic and exonic regions of the gene in *A. pompejana* (black) and *A. caudata* (grey), (**C**) between-paralog K_a/K_s mean (dashed line/left) and the average within-paralog π_a/π_s (solid line/right) along the coding sequence of the gene. Sliding window length = 50 bp, step size = 10 bp. Introns are colored in grey (5 introns); exons are colored as follows: orange: BRICHOS domain, green: AMP domain.

at least 20 sites of amino-acid replacements among paralogs. Both the GTR+I+G model and the selected best substitution model led to the exact same tree topology among Propiece paralogs (Fig. S4).

Search for positively-selected codon sites in the propiece including the BRICHOS domain. Search for positively-selected sites was performed on a restricted dataset of 8 consensus sequences (Table 2), but also over all the recovered non-recombining alleles (79 sequences for the propiece and 36 sequences for the BRICHOS domain, see extended experimental procedures, Tables S6 and S7). Comparing the free-ratio Branch model (M_1) with the one-ratio Branch model (M_0) did not produce significant differences in goodness-of-fit for either the BRICHOS (LRT = 9.44, df = 13, NS) or the propiece (LRT = 19.80, df = 13, NS) domains, suggesting that non-synonymous changes may be equally distributed among branches of the paralogous tree. The LRT was also not significant when comparing the no-clock model with the model assuming a global molecular clock for both domains, suggesting that the substitution rates may have been constant over time during duplications. Likelihood ratio tests performed between Site models using the consensus sequence dataset were significant for the propiece region (p-value < 0.01), not the BRICHOS. This finding was also supported by analyses in which polymorphic sequences were also included for the two genic regions under scrutiny (Tables S6 and S7). The comparison between models M_0 and M_3 indicated that ω is heterogeneously distributed among codon sites along the two domains (Table S6). Other comparisons (M_{1a} vs. M_{2a}; M_7 vs. M_8) went a step further and gave credence to the hypothesis of positive selection with a non-negligible proportion of the codon

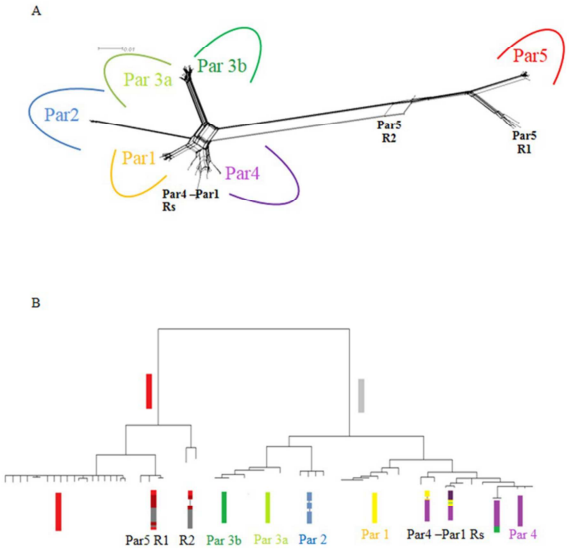


Figure 3. Splits Tree reticulated network of all alleles encoded by the preproalvinellacin multigenic family in *A. pompejana:* (**A**) the NeigborNet network together with (**B**) the schematic reconstruction of natural intergenic recombinants using the NJ tree topology obtained with the MEGA 5 software and the HKY model of substitutions. Sequences from the 5′ region were used in the two reconstruction methods and the six paralogs (Parx) as well as their intergenic recombinants (Rx) are represented by boxes in which the red, dark green, light green, blue, yellow and purple colors represent the genetic paralogous background and their proportions in recombinants, darker colors, the portions of the recombining alleles which accumulated their own set of mutations, and bars correspond to indel polymorphisms (including alleles with the 34 codons deletion in Par5).

sites under diversifying selection for the propiece region only (34%, $\omega = 4.84$; Table 2). Eight positively selected sites were identified in the BRICHOS domain using the Naive Empirical Bayes (NEB) method, some of which were previously identified in Figs 4 and S1. None of these remained significant with the BEB method (Table 2). In the propiece domain, fifteen codon sites exhibited a significant NEB probability, with two still significant with the BEB method under the M8 model (sites P60Y* and Q68E*). Results were nearly similar with all polymorphic sequences (Tables S6 and S7).

Preproalvinellacin gene induction upon abiotic stress. The level of expression of the BRICHOS containing preproalvinellacin gene was evaluated under two stressful conditions previously shown to induce the synthesis of molecular chaperones (Heat shock proteins, Hsp) in *Alvinella pompejana*[48]. Both the preproalvinellacin and the Hsp70, quantified by RT-qPCR, produced exactly the same pattern of gene expression in worms submitted to various temperatures and pressure stresses (Fig. 5). This data supports the conclusion that the *Alvinella* BRICHOS, in addition to its sequence similarities, behaves like a molecular chaperone.

Discussion

The evolution of a newly described AMP gene involved in the external immunity of annelids was investigated in two sister alvinellid species. This AMP participates in maintaining a highly conserved ecto-symbiotic microflora vital to life in the highly fluctuating vent habitat. In order to explain the lack of genetic diversity of the AMP itself in spite of a very high level of gene diversification, signs of positive selection were sought within the different functional domains of the propiece. The roles of the BRICHOS chaperone and the alvinellacin AMP have been teased apart in the context of the abrupt thermal variations encountered by the worm.

'Hot spots' of high genetic diversity result from tandem duplications and intergenic recombination. The role of duplication in the diversification of AMP genes has been well documented, leading to complex multigenic families[49–52]. This process may itself be adaptive in the co-evolutionary arms race between the host and pathogens by generating new copies of AMPs able to evolve more rapidly, and thus displaying new antimicrobial properties against newly encountered microbes, without erasing old functions[9]. Our genetic dataset demonstrates that the preproalvinellacin peptide follows this rule and is encoded by a multigenic family of at least six genes, some of which are repeated in tandem. Reciprocal monophyly of the coalescence trees combined with the non-juxtaposition of the intronic 'hot spots' of diversity between the two *Alvinella* sister species indicates that the diversification of the preproalvinellacin gene occurred independently through recurrent and recent duplication events after speciation. The two *Alvinella* species separated a long time ago (several tens of millions of years) with about 23% divergence on the mtCOI[53]. Independent and recent duplications have already been reported, e.g. in murine beta-defensins for which gene duplications took place after mice and rats diverged ca. 40 million

	N	L	S(ex)	Nsites (ex)	Pi exon	Nsites (ex+int)	S (ex+int)	pi (ex+int)	ThetaW	Hd	Tajima	Fu&Li F	Fu & Li D
par1	12	1293	10	315	0.0078	978	25	0.0048	0,0064	0.985	−1.145	−1.218	−1.365
par2	6	1124	2	214	0.0029	910	8	0.0026	0,0031	1	−1.072	−1.063	−1.145
par3a	13	1242	8	315	0.0058	927	30	0.0059	0,0078	1	−1.047	−0.998	−1.157
par3b	9	1239	8	316	0.0055	923	22	0.0055	0,0065	1	−0.804	−0.674	−0.791
par4	14	1114	12	316	0.0112	798	54	0.0121	0,0158	0.978	−1.039	−0.617	−0.842
par5	25	1270	8	349	0.0024	921	37	0.0032	0,0077	0.983	−2.207**	−3.360**	−3.518**

Table 1. Genetic diversities and associated statistical tests for the six preproalvinellacin paralogs of *Alvinella pompejana*. N: number of sequences; L: length of the paralog. S: number of polymorphic sites in both exonic (ex) and entire region (ex+int), Nsites: effective length of each domain, Pi: nucleotide diversity, ThetaW: Theta Watterson, Hd: Haplotypic diversity, Tajima: Tajima's D test, Fu&Li F: Fu and Li's F test, Fu&Li D: Fu and Li's D test. Statistical significance: ** for $P < 0.05$.

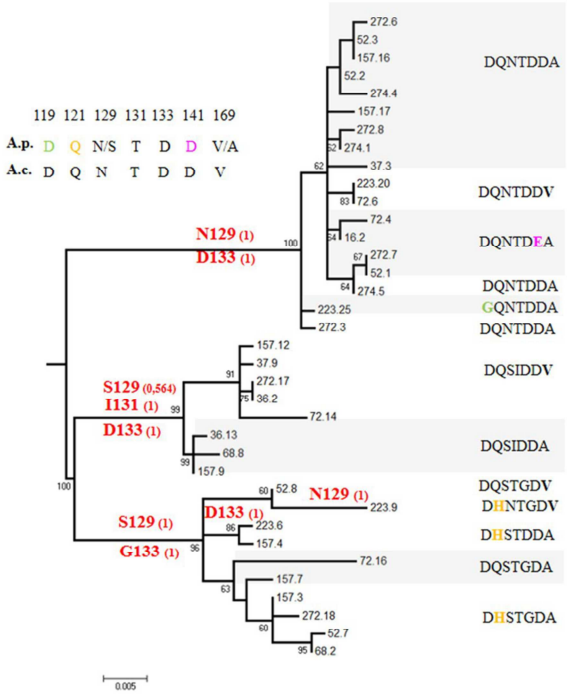


Figure 4. Ancestral reconstruction of polymorphic amino-acid replacements in the BRICHOS domain and the mature AMP in *Alvinella pompejana* (Ap) using the aaML package of PaML4.0 and *Alvinella caudata* (Ac) as an outgroup to orientate mutations. Positions of amino-acid replacements are labeled according to the start (methionine) codon. The Bayesian probability of occurrence of an amino-acid in the ancestral sequence is given in brackets. The tree topology and associated bootstrap values (obtained from 1000 replicates) was obtained by the ML method in Mega 5.0 using the full sequences of the 3′ region of the gene. Sequence labels represent the individual number and the clone number and are representative of the 6 paralogous clades (excluding natural recombinants) and subsequently used in the mapping of the BRICHOS mutations (see BRICHOS amino-acid alignment in Fig. S1).

years ago[54]. Consequently, these *Alvinella* species can be considered as ecological replicates of the same symbiotic model in which independent duplications arose probably in response to the gradual appearance of epibiosis, suggesting that the ancestor was not fully equipped to sustain durable interactions with existing microbial communities. Tandem duplications seem to have been the most appropriate mechanism for the long-term acquisition of a strict complex multi-specific epibiosis.

Series of intra- and inter-genic 'natural' recombinants were also identified, mainly in the first part of the preproalvinellacin gene without exon shuffling or deleterious frame-shifts of the coding sequence. Natural intergenic recombinants are rarely present in tandemly duplicated genes because of concerted evolution[55]. In the immune system, however, recombination seems to play a positive role in producing allelic diversity as previously shown for

BRICHOS	Branch Models				Site Models			
	Model	**M0**	**M0 (with clock)**	**M1 (free-ratio)**	**M1a**	**M2a**	**M7**	**M8**
	Log likelihood	−375.37	−381.68	−370.65	−374.63	−374.33	−374.63	−374.33
	Parameters	ω = 0,612, kappa = 14.4	ω = 0,625, kappa = 14.4	8 branches with ω > 1, kappa = 14.3	(ω_0 = 0) p_0 = 0.49, (ω_1 = 1) p_1 = 0.51, kappa = 13.9	(ω_0 = 0) p_0 = 0.67, (ω_1 = 1) p_1 = 0, (ω_2 = 1.92) p_2 = 0.33	p = 0.0050, q = 0.0052, kappa = 13.8	p_0 = 0.67 (ω_0 = 0), p = 0.005, p_1 = 0.33 (ω = 1.92) q = 1.75
	Sites with dN/dS > 1 (NEB analysis)	n.a.	n.a.	n.a	n.a.	(8 sites**) D119G; Q121H; R124R; N129S; T131I; D133G; D141E; V169A	n.a.	(8 sites**) D119G; Q121H; R124R; N129S; T131I; D133G; D141E; V169A
	Sites with dN/dS > 1 (BEB analysis)	n.a.	n.a.	n.a.	n.a.	V169A (ns)	n.a.	N129S; V169A (ns)
	LRT		12.62^{NS} (df = 7)	9.44^{NS} (df = 13)		0.6^{NS} (df = 2)		0.6^{NS} (df = 2)
PROPIECE	**Branch Model**				**Site Model**			
	Model	**M0**	**M0 (with clock)**	**M1 (free-ratio)**	**M1a**	**M2a**	**M7**	**M8**
	Log likelihood	−406.49	−412.34	−396.57	−405.02	−400.41	−405.02	−400.41
	Parameters	ω = 1.179, kappa = 9.2	ω = 1.189, kappa = 9.3	8 branches with ω > 1, kappa = 9.3	p_0 = 0.39 (ω_0 = 0), p_1 = 0.61 (ω_1 = 1)	p_0 = 0.66 (ω_0 = 0), p_1 = 0 (ω_1 = 1), p_2 = 0.34 (ω_2 = 4.84)	p = 0.009, q = 0.005, kappa = 7.1	p_0 = 0.66 (ω_0 = 0) p = 0.005, p_1 = 0.34 (ω = 4.84) q = 12.52
	Sites with dN/dS > 1 (NEB analysis)	n.a.	n.a.	n.a.	n.a.	(15 sites**) M22I: W24R; L26Q; N30S; A31V; H33D; I36T; E37K; P38Y; D57E; T60I; Q68E; D72N; H77R; L78S	n.a.	(15 sites**) M22I: W24R; L26Q; N30S; A31V; H33D; I36T; E37K; P38Y; D57E; T60I; Q68E; D72N; H77R; L78S
	Sites with dN/dS > 1 (BEB analysis)	n.a.	n.a.	n.a.	n.a.	(4 sites$^{\$}$ with p > 0.85) M22I: W24R; L26Q; N30S; A31V; H33D; I36T; P38Y$^{\$}$; D57E; T60I$^{\$}$; Q68E$^{\$}$; H77R$^{\$}$; L78S	n.a.	(4 sites$^{\$}$ with p > 0.85 and 2 sites*) M22I: W24R; L26Q; N30S$^{\$}$; A31V$^{\$}$; H33D; I36T; P38Y*; D57E; T60I$^{\$}$; Q68E*; H77R$^{\$}$; L78S
	LRT		11.70^{NS} (df = 7)	19.84^{NS} (df = 13)		9.22** (df = 2)		9.22** (df = 2)

Table 2. Log-likelihood values and parameter estimates for the BRICHOS and propiece domains of the preproalvinellacin gene using models implemented in the CodeML program of the PAML package with the alignments of consensus paralogous sequences (8) and the reference tree selected by jModelTest 2.1.7 (BRICHOS: K80+I, Propiece: Tim2ef I+G). M0 (one-ratio); M1 (free-ratio); M1a (nearly-neutral); M2a (selection); M7 (β); M8 (β+ω > 1) and the estimated log-likelihood values (l) by the CodeML program, ω = dN/dS nonsynonymous/synonymous rate ratio; p = proportion of sites for each site class. M0: one ω for the tree; M1: one ω per branch, M1a: p_0 = proportion of sites with $\omega_0 = 0$, $p_1 = 1 - p_0$, proportion of sites with $\omega_1 = 1$; M2a: p_0 ($\omega_0 = 0$), p_1 ($\omega_0 = 1$), and ω_2, $p_2 = 1 - p_0 - p_1$. M7: p and q (parameters of β distribution of ω between 0 and 1). M8: same as M7 except the addition of one site class in which ω is greater than one. Positively Selected Sites: Codon positions predicted to be under positive selection with a posterior probability of acceptance of $^{\$}$>85%, *>95% and **>0.99 (identification of sites exhibiting dN/dS ratio > 1) with either Naive Empirical Bayes (NEB) or Bayesian Empirical Bayes (BEB). Numbers refer to amino-acid positions from the initial methionine. LRT: likelihood ratio tests between substitution models (between nested branch models: M0 vs. M0 with clock, M0 with clock vs. M1; between nested site models: M1a vs. M2a, M7 vs. M8). NS: not significant, **p < 0.05.

MHC class I genes[56]. Successful recombination events have been frequent in the preproalvinellacin gene since the first duplication event, with some early duplications subsequently kept in populations and now displaying their own pattern of accumulated mutations (0.052 substitution per site). Mutation and recombination are two major evolutionary mechanisms driving genetic diversity and likely to promote an adaptive response to cope with biotic interactions, but their relative contribution varies greatly between genes and organisms[57]. In the specific case of preproalvinellacin, the main positive outcome of intergenic recombination would be the spread of positively selected mutations between duplicates leading to the observed patterns of "shared polymorphism" in the propiece and BRICHOS domains between paralogs. Although we cannot rule out the hypothesis of retention of an ancestral polymorphism due to balancing selection occurring within duplicated genes, a more likely explanation would be that the observed mutation reversals are the result of a transfer of newly gained positive mutations from one duplicated gene to another via gene conversion or unequal crossing-over as previously proposed in newly duplicated genes in tandem[58]. Balancing selection within duplicates would have led to high π_a/π_s ratios around the non-synonymous shared polymorphisms, a situation not recorded here.

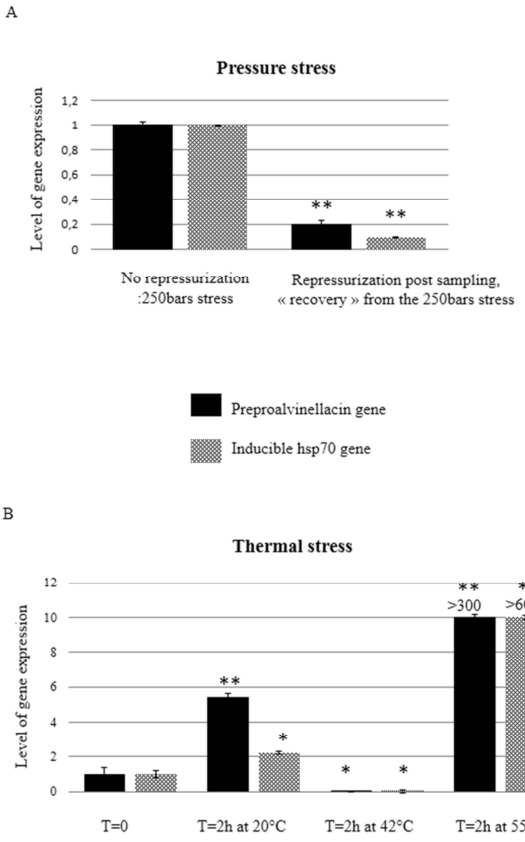


Figure 5. Upregulation of expression of the preproalvinellacin gene and the molecular chaperone hsp 70 in *Alvinella pompejana* submitted to thermal or pressure stress. (**A**) The level of transcription is significantly higher for both genes in "not re-pressurized animals" compared to those re-pressurized immediately after raising from −2500 m at the *in situ* pressure of 250 bars in the DESEARES vessel. (**B**) In animals kept under *in situ* pressure in the BALIST device, thermal stress led to an up-regulation of the two genes. P-values from Student's tests were calculated *versus* the control treatment (normalized to 1), based on the experimental measures performed in triplicates (**$p < 0.01$, *$p < 0.05$).

An antimicrobial function under strong purifying selection. The alvinellacin AMP is strictly monomorphic and identical between paralogs. No variation has even been observed between individuals issued from geographically-disjoint populations of *A. pompejana* situated on either side of the equator and thus physically separated since at least two million years[59]. AMP sequences of the two *Alvinella* species diverge only by one amino-acid replacement (N- > S) despite speciation having occurred a long time ago[53,60]. Thus, a strong directional purifying selection must have occurred on the antimicrobial effector. The strong purifying selection acting on the antimicrobial function is probably a consequence of a convergent evolution of the two species to the same microbial environment (at least one dominant ε-proteo phylobiont shared between species[20]). The two *Alvinella* species live syntopically on the walls of high-temperature chimneys along the East-Pacific Rise. This strongly suggests that host-symbiont interactions have co-evolved very slowly, possibly due to the high level of specialization of the vent microflora since the beginning of cretaceous, despite the highly chaotic variations of vent fluid discharge[61]. By contrast, changing/fluctuating biotic conditions, such as those exerted by rapidly co-evolving parasites, often result in signs of balancing selection, thus contributing to the maintenance of AMP polymorphisms as observed in arthropods[11,62]. Such patterns of purifying selection have also been observed in defensin genes of primates, which are associated with function(s) that became fixed quite early in mammalian evolution, making them less variable than expected under neutral evolution[63]. More generally, the evolution of AMPs appears to be a slow process compared to the situation observed for other immune-related genes[62,64]. It is remarkable and puzzling that AMPs are so well conserved at the species level and maintain their activity against co-evolving microbes, especially against symbionts that could have developed cheater strategies after such a long co-evolutionary time. One hypothesis is that AMPs act in complex combinations that can be synergistic, making the establishment of microbial resistance against a particular AMP difficult[62,65–69].

The BRICHOS domain prevents the incorrect folding of alvinellacin as temperature fluctuates. It is of particular interest that *A. pompejana*, which is still considered as one of the most thermotolerant and eurythermal animals on Earth, synthesizes a precursor containing an AMP with β sheet structure (alvinellacin) together with a molecular chaperone (BRICHOS). The BRICHOS domain constitutes the first example of a chaperone-like domain with a high propensity and specificity for β-prone regions[70,71]. In mammals, this chaperone binds with β hairpin motifs in order to prevent β sheet aggregation and amyloid fibril formation, which is highly toxic for organisms as exemplified in various diseases such as Alzheimers syndrome[72]. As such, this domain ensures correct β sheet folding and the subsequent activity of the protein released from the same precursor. Mutations in this domain are causative agents of multiple diseases e.g. stomach cancer, dementia, and respiratory distress in humans[73]. Even though a BRICHOS domain of over 300 proteins from 12 different protein families was identified in both protostomian and deuterostomian models in 2009[74], its association with an AMP constitutes the first observation of its direct implicit folding function in an invertebrate.

There is still no experimental method to simply assign a molecular chaperone function to a protein: the strongest presumption that a protein plays this 'repairing' role remains, even today, its induction upon stressful conditions. Our data show a 350-fold up-regulation of the gene encoding the preproalvinellacin precursor with increasing thermal shocks, thus supporting the hypothesis of its chaperone function in addition to its sequence homology with BRICHOS. Interestingly, BRICHOS and alvinellacin are both released from the epidermis cells into the acidic and thermally highly fluctuating environment of the Pompeii worm. These conditions typically favor the auto-aggregation and accumulation of molecules having a β sheet tertiary structure (like alvinellacin), generating (i) toxic amyloid fibrils, and (ii) loss of the antimicrobial properties of the molecule. Consequently, BRICHOS presumably ensures the correct folding of the secreted AMP over a wide range of thermal conditions and thus the optimal functioning of the AMP to shape and control the vital ectosymbiosis. As such, it could be a likely target for positive diversifying selection to cope with this extremely variable thermal habitat.

Evolutionary dynamics of the preproalvinellacin gene. Maximum likelihood ratio tests (CodeML) and both neutral (Tajima's D) and 'selection' (McDonald-Kreitman) tests were performed to determine whether evolution of the preproalvinellacin multigenic family is driven by positive selection and whether selection acts differentially on the subsequent functional domains of the propiece region. Values of both $\theta\pi$ and d_N/d_S along the gene indicate the occurrence of an unexpected very sharp 'hot spot' of non-synonymous mutations on the BRICHOS domain. Mapping these variants on the tree of paralogs suggests that this domain duplication was followed by a positive diversifying selection. Unlike the propiece region, the MK tests and our likelihood/Bayesian tests fail to discriminate positive selection from selective relaxation in the chaperone domain of the gene, probably as a consequence of the high rate of recombination between paralogs in this specific region. Overall, our results highlight a complex selective situation in which duplicates have first been subjected to diversifying selection (*i.e.* positive diversification by duplication: see results on the propiece domain), then partially compensated by sporadic genetic exchanges due to gene conversion/recombination, which in turn have acted in order to maximize genetic diversity over the whole set of these tandemly repeated genes (*i.e.* maintenance of the whole system by balancing selection). In contrast to the alvinellacin peptide, the high number of non-synonymous variants of BRICHOS may reflect a response to selective environmental constraints that act specifically on this region. Adaptation to high temperatures is a complex evolutionary process that can involve modifications of the intrinsic stability of proteins[75] and/or interactions with molecular chaperones that help stabilize or re-fold the focal protein. Here, it appears that adaptation to highly fluctuating temperatures has acted more specifically on the propiece region and its molecular chaperone. As the conformation of a molecule is key to its biological activity, we hypothesize that the observed allelic variants of BRICHOS contribute to the stabilization of the alvinellacin hairpin in the context of the variable abiotic (thermal) conditions of the tube habitat, thus maintaining an efficient external immunity against pathogenic bacteria and an efficient control of vital epibiota.

Conclusion

Highly fluctuating physico-chemical conditions have not promoted diversifying selection on alvinellacin *per se* in contrast to the situation generally observed in other metazoan AMPs. On the contrary, a strong purifying selection is evident, despite the duplication-driven diversification of its chaperone containing precursor. Duplication of genes has often been viewed as a molecular mechanism by which animals or plants adapt to changing environmental conditions with little cost[76]. Here, we demonstrate that exhibiting a vital and highly conserved ecto-symbiosis in the face of thermal fluctuations has led to a peculiar selective trend promoting the adaptive diversification of the molecular chaperone of the AMP, but not of the AMP itself. This finding significantly differs from previous results, as no polymorphism (following the "matching allele" model of Red Queen theory), nor duplication and ensuing divergence (following the "gene-for-gene" model) was observed for alvinellacin. As a consequence, because of the uniqueness of its chaperone, the preproalvinellacin gene family represents an interesting model to better understand the evolution of external immunity *in natura*. Our results fill some knowledge gaps concerning the function of BRICHOS and revive innovative topics that question the evolutionary success of the BRICHOS domain in a large variety of animal proteins, notably its anti-amyloid function which may have appeared early in the history of life.

References

1. Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389–395 (2002). doi:10.1038.
2. Maroti, G., Kereszt, A., Kondorosi, E. & Mergaert, P. Natural roles of antimicrobial peptides in microbes, plants and animals. *Res Microbiol* **162**, 363–374 (2011). doi:S0923/j.resmic.2011.02.005.
3. Bulet, P., Stocklin, R. & Menin, L. Anti-microbial peptides: from invertebrates to vertebrates. *Immunol Rev* **198**, 169–184 (2004).
4. Login, F. H. *et al.* Antimicrobial peptides keep insect endosymbionts under control. *Science* **334**, 362–365, doi:10.1126/science.1209728 (2011).

5. Salzman, N. H. *et al.* Enteric defensins are essential regulators of intestinal microbial ecology. *Nat Immunol* **11**, 76–83, doi:10.1038/ni.1825 (2009).
6. Tasiemski, A. *et al.* Reciprocal immune benefit based on complementary production of antibiotics by the leech Hirudo verbana and its gut symbiont Aeromonas veronii. *Sci Rep* **5**, 17498, doi:10.1038/srep17498 (2015).
7. Franzenburg, S. *et al.* Distinct antimicrobial peptide expression determines host species-specific bacterial associations. *Proc Natl Acad Sci USA* **110**, E3730–3738, doi:10.1073/pnas.1304960110 (2013).
8. Gallo, R. L. & Nakatsuji, T. Microbial symbiosis with the innate immune defense system of the skin. *J Invest Dermatol* **131**, 1974–1980, doi:10.1038/jid.2011.182 (2011).
9. Tennessen, J. A. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. *J Evol Biol* **18**, 1387–1394 (2005). doi:10.1111/j.1420.
10. Gosset, C. C., Do Nascimento, J., Auge, M. T. & Bierne, N. Evidence for adaptation from standing genetic variation on an antimicrobial peptide gene in the mussel Mytilus edulis. *Mol Ecol* **23**, 3000–3012, doi:10.1111/mec.12784 (2014).
11. Unckless, R. L., Howick, V. M. & Lazzaro, B. P. Convergent Balancing Selection on an Antimicrobial Peptide in Drosophila. *Curr Biol* **26**, 257–262, doi:10.1016/j.cub.2015.11.063 (2016).
12. Unckless, R. L. & Lazzaro, B. P. The potential for adaptive maintenance of diversity in insect antimicrobial peptides. *Philos Trans R Soc Lond B Biol Sci* **371**, doi:10.1098/rstb.2015.0291 (2016).
13. Salathé, M., Kouyos, R. D. & Bonhoeffer, S. The state of affairs in the kingdom of the Red Queen. *Trends in Ecology & Evolution* **23**, 439–445 (2008).
14. Otti, O., Tragust, S. & Feldhaar, H. Unifying external and internal immune defences. *Trends Ecol Evol* **29**, 625–634, doi:10.1016/j.tree.2014.09.002 (2014).
15. Salzet, M., Tasiemski, A. & Cooper, E. Innate immunity in lophotrochozoans: the annelids. *Curr Pharm Des* **12**, 3043–3050 (2006).
16. Tasiemski, A. & Salzet, M. Leech immunity: from brain to peripheral responses. *Adv Exp Med Biol* **708**, 80–104 (2010).
17. Conlon, J. M. The contribution of skin antimicrobial peptides to the system of innate immunity in anurans. *Cell Tissue Res* **343**, 201–212 (2011). doi:10.1007.
18. Schikorski, D. *et al.* The medicinal leech as a model for studying the antimicrobial response of the central nervous system. *J Immunol* **181**, 1083–1095 (2008).
19. Tasiemski, A. *et al.* Characterization and function of the first antibiotic isolated from a vent organism: the extremophile metazoan Alvinella pompejana. *PLoS ONE* **9**, e95737, doi:10.1371/journal.pone.0095737 (2014).
20. Cary, S. C., Cottrell, M. T., Stein, J. L., Camacho, F. & Desbruyeres, D. Molecular Identification and Localization of Filamentous Symbiotic Bacteria Associated with the Hydrothermal Vent Annelid *Alvinella pompejana*. *Appl Environ Microbiol* **63**, 1124–1130 (1997).
21. Le Bris, N. & Gaill, F. How does the annelid *Alvinella pompejana* deal with an extreme hydrothermal environment? *Rev Environ Sci Biotechnol* **6**, 197–221 (2007).
22. Bonch-Osmolovskaya, E. A. *et al.* Activity and distribution of thermophilic prokaryotes in hydrothermal fluid, sulfidic structures, and sheaths of alvinellids (East Pacific Rise, 13 degrees N). *Appl Environ Microbiol* **77**, 2803–2806, doi:10.1128/AEM.02266-10 (2011).
23. Di Meo-Savoie, C. A., Luther, G. W. & Cary, S. C. Physicochemical characterization of the microhabitat of the epibionts associated with *Alvinella pompejana*, a hydrothermal vent annelid. *Geochim Cosmochim Acta* **68**, 2055–2066 (2004).
24. Luther, G. W. 3rd *et al.* Chemical speciation drives hydrothermal vent ecology. *Nature* **410**, 813–816, doi:10.1038/35071069 (2001).
25. Le Bris, N., Zbinden, M. & Gaill, F. Processes controlling the physico-chemical microenvironments associated with Pompeï worms. *Deep-sea research* **52**, 1071–1083 (2005).
26. Cary, S. C., Shank, T. & Stein, J. Worms basks in extreme temperatures. *Nature* **391**, 545–546 (1998).
27. Chevaldonné, P., Desbruyères, D. & Le Haître, M. Time-series of temperature from three deep-sea hydrothermal vent sites. *Deep-Sea Res. Part A* **38**, 1417–1430 (1991).
28. Grzymski, J. J. *et al.* Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. *Proc Natl Acad Sci USA* **105**, 17516–17521, doi:10.1073/pnas.0802782105 (2008).
29. Doyle, J. & Doyle, J. L. Genomic plant DNA preparation from fresh tissue-CTAB method. *Phytochem. Bull* **19**, 11–15 (1987).
30. Jolly, M., Viard, F., Weinmayr, G., Gentil, F., Thiébaut, E. & Jollivet, D. Does the genetic structure of *Pectinaria koreni* (Polychaeta: Pectinariidae) conform to a source-sink metapopulation model at the scale of the Baie de Seine? *Helgoland Mar. Res.* **56**, 238–246 (2003).
31. Bierne, N. *et al.* Mark–recapture cloning: a straightforward and cost-effective cloning method for population genetics of single-copy nuclear DNA sequences in diploids. *Molecular Ecology Notes* **7**, 562–566, doi:10.1111/j.1471-8286 (2007).
32. Heath, L., van der Walt, E., Varsani, A. & Martin, D. P. Recombination patterns in aphthoviruses mirror those found in other picornaviruses. *J. Virol.* **80**, 11827–11832 (2006).
33. Librado, P. & Rozas, J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **25**, 1451–1452, doi:10.1093/bioinformatics/btp187 (2009).
34. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
35. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat Meth* **9**, 772–772, doi:10.1038/nmeth.2109 (2012).
36. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution* **30**, 2725–2729, doi:10.1093/molbev/mst197 (2013).
37. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology* **59**, 307–321, doi:10.1093/sysbio/syq010 (2010).
38. Tavaré, S. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *American Mathematical Society: Lectures on Mathematics in the Life Sciences* **17**, 57–86 (1986).
39. Shoemaker, J. S. & Fitch, W. M. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. *Molecular Biology and Evolution* **6**, 270–289 (1989).
40. Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution* **39**, 306–314, doi:10.1007/bf00160154 (1994).
41. Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254–267, doi:10.1093/molbev/msj030 (2006).
42. Yang, Z. PaML4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
43. Goldman, N. & Yang, Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**, 725–736 (1994).
44. Thornton, K. & Long, M. Excess of Amino Acid Substitutions Relative to Polymorphism Between X-Linked Duplications in Drosophila melanogaster. *Molecular Biology and Evolution* **22**, 273–284, doi:10.1093/molbev/msi015 (2005).
45. Arguello, J. R., Chen, Y., Yang, S., Wang, W. & Long, M. Origination of an X-Linked Testes Chimeric Gene by Illegitimate Recombination in Drosophila. *PLOS Genetics* **2**, e77, doi:10.1371/journal.pgen.0020077 (2006).
46. Meisel, R. P., Hilldorfer, B. B., Koch, J. L., Lockton, S. & Schaeffer, S. W. Adaptive Evolution of Genes Duplicated from the Drosophila pseudoobscura neo-X Chromosome. *Molecular Biology and Evolution* **27**, 1963–1978, doi:10.1093/molbev/msq085 (2010).

47. Hahn, M. W. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity* **100**, 605–617, doi:10.1093/jhered/esp047 (2009).
48. Ravaux, J. *et al.* Thermal limit for metazoan life in question: *in vivo* heat tolerance of the pompeii worm. *PLoS ONE* **8**, e64074, doi:10.1371/journal.pone.0064074 (2013).
49. Schutte, B. C., Mitros, J. P. & Bartlett, J. A. Discovery of five conserved β-defensin gene clusters using a computational search strategy. *Proceedings of the National Academy of Sciences* **99**, 2129–2133 (2002).
50. Semple, C. A. M., Rolfe, M. & Dorin, J. R. Duplication and selection in the evolution of primate b-defensin genes. *Genome Biol.* **4** (2003).
51. Tennessen, J. A. & Blouin, M. S. Selection for antimicrobial peptide diversity in frogs leads to gene duplication and low allelic variation. *J Mol Evol* **65**, 605–615 (2007). doi:10.1007.
52. Lynn, D. J. *et al.* Bioinformatic discovery and initial characterisation of nine novel antimicrobial peptide genes in the chicken. *Immunogenetics* **56**, 170–177 (2004). doi:10.1007.
53. Vrijenhoek, R. C. On the instability and evolutionary age of deep-sea chemosynthetic communities. *Deep Sea Res.* Part II, 189–200 (2013).
54. Maxwell, A. I., Morrison, G. M. & Dorin, J. R. Rapid sequence divergence in mammalian beta-defensins by adaptive evolution. *Mol Immunol* **40**, 413–421 (2003).
55. Elder, J. F. Jr. & Turner, B. J. Concerted evolution of repetitive DNA sequences in eukaryotes. *The Quarterly Review of Biology* **70**, 297–320 (1995).
56. Zhao, M. *et al.* Evolution by selection, recombination, and gene duplication in MHC class I genes of two Rhacophoridae species. *BMC Evol Biol* **13**, 113, doi:10.1186/1471-2148-13-113 (2013).
57. Awadalla, P. The evolutionary genomics of pathogen recombination. *Nat Rev Genet* **4**, 50–60 (2003). doi:10.1038/nrg964.
58. Fawcett, J. A. & Innan, H. Neutral and non-neutral evolution of duplicated genes with gene conversion. *Genes (Basel)* **2**, 191–209, doi:10.3390/genes2010191 (2011).
59. Plouviez, S., Le Guen, D., Lecompte, O., Lallier, F. H. & Jollivet, D. 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 Evol Biol* **10**, 220 (2010). doi:10.1186.
60. Little, C. T. S. & Vrijenhoek, R. C. Are hydrothermal vent animals living fossils? *Trends Ecol. Evol.* **18**, 582–588 (2003).
61. Haymon, R. M., Koski, R. A. & Sinclair, C. Fossils of hydrothermal vent worms from cretaceous sulfide ores of the samail ophiolite, oman. *Science* **223**, 1407–1409, doi:10.1126/science.223.4643.1407 (1984).
62. Rolff, J. & Schmid-Hempel, P. Perspectives on the evolutionary ecology of arthropod antimicrobial peptides. *Philos Trans R Soc Lond B Biol Sci* **371**, doi:10.1098/rstb.2015.0297 (2016).
63. Crovella, S. *et al.* Primate beta-defensins–structure, function and evolution. *Curr Protein Pept Sci* **6**, 7–21 (2005).
64. Lazzaro, B. P. Natural selection on the Drosophila antimicrobial immune system. *Curr Opin Microbiol* **11**, 284–289, doi:10.1016/j.mib.2008.05.001 (2008).
65. Cassone, M. & Otvos, L. Jr. Synergy among antibacterial peptides and between peptides and small-molecule antibiotics. *Expert Rev Anti Infect Ther* **8**, 703–716, doi:10.1586/eri.10.38 (2010).
66. Lauth, X. *et al.* Bass hepcidin synthesis, solution structure, antimicrobial activities and synergism, and *in vivo* hepatic response to bacterial infections. *J Biol Chem* **280**, 9272–9282, doi:10.1074/jbc.M411154200 (2005).
67. Nagaoka, I., Hirota, S., Yomogida, S., Ohwada, A. & Hirata, M. Synergistic actions of antibacterial neutrophil defensins and cathelicidins. *Inflamm Res* **49**, 73–79, doi:10.1007/s000110050561 (2000).
68. Rosenfeld, Y., Barra, D., Simmaco, M., Shai, Y. & Mangoni, M. L. A synergism between temporins toward Gram-negative bacteria overcomes resistance imposed by the lipopolysaccharide protective layer. *J Biol Chem* **281**, 28565–28574, doi:10.1074/jbc.M606031200 (2006).
69. Yan, H. & Hancock, R. E. Synergistic interactions between mammalian antimicrobial defense peptides. *Antimicrob Agents Chemother* **45**, 1558–1560, doi:10.1128/AAC.45.5 (2001).
70. Knight, S. D., Presto, J., Linse, S. & Johansson, J. The BRICHOS domain, amyloid fibril formation, and their relationship. *Biochemistry* **52**, 7523–7531, doi:10.1021/bi400908x (2013).
71. Sanchez-Pulido, L., Devos, D. & Valencia, A. BRICHOS: a conserved domain in proteins associated with dementia, respiratory distress and cancer. *Trends Biochem Sci* **27**, 329–332 (2002).
72. Willander, H., Hermansson, E., Johansson, J. & Presto, J. BRICHOS domain associated with lung fibrosis, dementia and cancer–a chaperone that prevents amyloid fibril formation? *FEBS J* **278**, 3893–3904, doi:10.1111/j.1742 (2011).
73. Landreh, M., Rising, A., Presto, J., Jornvall, H. & Johansson, J. Specific chaperones and regulatory domains in control of amyloid formation. *J Biol Chem* **290**, 26430–26436, doi:10.1074/jbc.R115.653097 (2015).
74. Hedlund, J., Johansson, J. & Persson, B. BRICHOS - a superfamily of multidomain proteins with diverse functions. *BMC Res Notes* **2**, 180 (2009). doi:10.1186.
75. Jollivet, D. *et al.* Proteome adaptation to high temperatures in the ectothermic hydrothermal vent Pompeii worm. *PLoS One* **7**, e31150, doi:10.1371/journal.pone.0031150 (2012).
76. James, T. C., Usher, J., Campbell, S. & Bond, U. Lager yeasts possess dynamic genomes that undergo rearrangements and gene amplification in response to stress. *Current Genetics* **53**, 139–152 (2008).

Acknowledgements

We thank the captain and crew of the RV Atalante, the DSV Nautile group (IFREMER), along with N. Le Bris (UMR8222) and F. Lallier (UMR7144), chief scientists of the MESCAL cruises. We thank A.S. Leport (UMR7144) for her technical assistance and B. Shillito (BOREA) and S. Hourdez (UMR7144) for the use of BALIST and DESEARES, respectively. We also acknowledge I. Probert (MBRC, Roscoff) for the English proofreading. This work was funded by the Université de Lille (BQR), the CNRS, the GDR ECCHIS and the Région Nord Pas de Calais-FRB (VERMER project).

Author Contributions

C.P. generated and analyzed the sequence dataset. D.J. designed and contributed to the genetic analysis and performed the sampling with A.T. A.T. conducted and performed the experiments under pressure and the RT-qPCR. F.M. and C.P. participated in writing the paper. A.T. and D.J. supervised the work and wrote the paper.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-01626-2

Competing Interests: The authors declare that they have no competing interests.

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

SUPPLEMENTARY FILES :

Antagonistic evolution of an antibiotic and its molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat

Claire Papot, François Massol, Didier Jollivet, Aurélie Tasiemski

---

SUPPLEMENTARY INFORMATION

EXTENDED EXPERIMENTAL PROCEDURES

Gene amplification. The preproalvinellacin gene displays a 5 introns/6 exons structure, with a first large intron of 449 bp after the signal peptide sequence [1] (Fig. 2). Because of its length, nested primers with a 100 bp overlap in the middle of the gene were designed to split the gene at the beginning of the BRICHOS domain (see table S1): the 5' part comprises both the peptide signal and the linker region (3 exons + 3 introns) whereas the 3' part contains the BRICHOS domain and the antimicrobial peptide (3 exons + 2 introns).

Specimen sampling

The genus *Alvinella* is made of only two species: the Pompeii worm *A. pompejana* and its closely related syntopic species *A. caudata*. Both species live on the wall of vent chimneys distributed along the East Pacific Rise (EPR). Animals were collected using the arm of the manned submersible Nautile and brought back to the surface inside an insulated basket. Until sampled, animals were measured, sexed and frozen immediately at -80°C for further laboratory analysis or alternatively freshly dissected to perform a DNA extraction directly onboard from the anterior part of the animal (gills and prostomium, which are devoid of epibionts). The two sampling sites are separated by a distance of about 3000km and a major faulting system (Quebrada/Discovery/Gofar fracture zone) around the Equator known to

represent a biogeographic barrier for the vent fauna [2]. *Alvinella* specimens from both sides of the EPR therefore display a *Cox-1* specific mitochondrial signature with 2% divergence.

PCR amplification, cloning and sequencing

Four *A. caudata* specimens and 96 *A. pompejana* were sampled from populations on both sides of the Equatorial barrier (50 individuals from Fromveur and 50 individuals from Bio9). Each individual was amplified by PCR separately with a different tag combination. PCR were conducted in a 25µL volume including: 1X buffer, 2mM MgCL2, 0.05mM of each dNTP, 0.4µM of each primer, 1U of Taq polymerase (Uptitherm^{TM}, Interchim). Thermal cycling parameters used an initial denaturation step at 96°C for 4 min, followed by 40 cycles at 96°C for 30 s, 60°C for 45 s and 72°C for 2 min, before a 10 min final extension at 72°C. PCR products were then pooled together before cloning. In *A. pompejana*, the technique consisted in pooling the same amount of DNA from PCR amplification of 16 previously tagged individuals for each cloning assay and, to perform six distinct cloning assays for both the 5' and 3' regions of the gene. In other words, 12 pools were made representing 96 individuals for each part of the gene (i.e. 192 amplifications representing 384 allelic fragments for an autosomal locus). Pools of tagged PCR products were purified using QIAquick^{TM} columns and ligated into a Bluescript vector using the TOPO_TA cloning kit (Invitrogen^{TM}) and subsequently transformed into top10 competent *E. coli* strain following the manufacturer's instructions. For each cloning assay, after amplification of the insert with Puc-specific plasmid (outside the polyclonal region) BS1 primers, 96 positive clones (i.e. containing an insert of the right size) were sequenced on both strands with the sequencing primers M13F or M13R, leading to a total number of 1152 sequences and a recapture effort of 3.0 under the single locus assumption. The mark-recapture cloning method led to more than 900 proof-read sequences in both directions (321 in the 5' region and 566 in the 3' region) for the focus species *A. pompejana* and only 20 and 58 sequences in the 5' and 3' regions for the outgroup

species *A. caudata*. A consensus sequence of the two forward and reverse reads was produced for each clone. Sequences recaptured more than once were the only sequences kept for the first allelic assignment to duplicated loci. The mark-recapture cloning technique generates about 30% of artifactual recombinants with our complex dataset when looking at the most recaptured individuals (c.a. > 20 clones). These chimeric alleles were either due to intra-locus or inter-loci recombination during the PCR. Some recombinants were, however, considered to be natural when recaptured in more than two individuals (i.e. alleles with the same recombination breakpoints in distinct individuals).

Cleaning sequence datasets from artifactual mutational events

Global alignments of consensus sequences were obtained with the Geneious software using ClustalW with the free ends gap option, a gap penalty of 1.0 and a cost matrix option of 51% similarity. Chimeras of alleles for heterozygous individuals or chimeras of alleles between closely-related loci in the specific case of duplicated genes were of frequent occurrences in our sequence datasets. Tracing back recombinants was, however, only possible for the most recaptured individuals displaying at least 20 clones mainly because of the high number of duplicated genes in our set of sequences. Hence, intra-individual *in vitro* recombination points between alleles and/or duplicated genes were searched using RDP4.0. First, for each set of clonal sequences attributed to one individual, the alignment-based "Automated MAxChi" procedure [3] was performed with all settings left as default. The within-individual alignment with 'true' allelic sequences and identified chimeras was thereafter checked manually to visualize any additional intra-individual recombinants. Second, putative recombinants in a given individual were compared with the whole sequence dataset to see whether they might be shared with other individuals. Recombinants found in more than one individual were kept and assumed to represent 'natural' intra- or inter-loci recombination events. Third, a second set of recombinant search was performed with the MAxChi procedure over the multi-

individual alignment of the remaining sequences to identify additional recombining points across distinct individuals. These new recombinants were also removed from the dataset if not observe in at least two distinct individuals. This allowed us to confirm the natural existence of previously described alleles. Finally, alleles from individuals weakly recaptured or only recaptured once were added to the final dataset if they were able to match at least one sequence of the curated alignment. On this 'cleaned' dataset, artifactual/somatic mutations were also removed taking advantage of the multiple recaptures (i.e. >20 sequences) by suppressing singletons between intra-individual sequences that referred to a well-assigned allele. This allowed us to calculate a rate at which artifactual mutations occur in the dataset and to apply this rate on singletons found in the other less recaptured individuals.

Paralog identification and individual genotyping

Combining the 5' and 3' regions of the gene (separate PCR amplification) and thus, the exact correspondence of 5' and 3' alleles, was not possible in this study due to the high rate of recombination and the lack of diagnostic sites in the 100 bp-overlapping region of the gene, leading to a disjoint assignment of paralogs in the two genic regions. Exact allelic concordance was only met for 5' paralog 5 and the 3' paralog E as they both display the highest level of divergence with the other paralogs, respectively. A detailed analysis of paralogs was performed at the intraspecific level from the whole 'cleaned' sequence dataset recovered from *A. pompejana* on the 5' region of the gene. This region was chosen because it contains long intronic regions that help to discriminate more easily between paralogs (i.e. specific signatures of linked sites). Forward and reverse paralog-specific primers were positioned on specific mutation signatures typifying each putative paralog with a final amplicon size of less than 400-nucleotides long. For each paralogous gene, direct sequencing allowed us to search for heterozygous individuals (double peaks in the chromatogram) at diagnostic sites using an alignment performed with the *de novo* Assemble module of the

Geneious software. Gene orthology was confirmed for a given set of primers when both homozygous and heterozygous individuals co-occur at previously chosen diagnostic sites.

The evolutionary history of paralogs was inferred with the Maximum Likelihood method of the software MEGA6 using the GTR model of substitutions and the allelic alignment of either the 5' or 3' regions (coding and non-coding region) of the gene. Initial tree(s) for the heuristic search were obtained by applying the BioNJ method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (4 categories (Gamma shape parameter = 0.13) with no invariable sites). The tree was drawn to scale with branch lengths measured as the number of substitutions per site. A search for the best model of substitutions was also performed using jModelTest 2.1.7. The tree topology obtained with the GTR+I+G model was compared with possible alternative trees. Results using the whole set of allelic sequences of the 5' region of the gene indicated that the best substitution model is the GTR+G according to the AIC or the TPM3uf+G model according to the BIC, but the three models (GTR+I+G, GTR+G and TPM3uf+G) fall within the 95% confidence intervals of the AIC/BIC analyses (i.e. models which have a substantial support for the dataset by summing the ranked weight (ωi) of each model (i) that uses the difference between each model-specific value of AIC/BIC and the minimum one). There was no significant difference between the GTR+G and GTR+G+I models (LRT=2.22, df=1) and the TPM3uf model produced a significantly decreased likelihood than the GTR+G+I model (LRT=2.45, df=4). Comparing the topologies obtained with PhyML under the two selected best models and the GTR+G+I model did not give much difference in topology (see alternative topologies below). Slight differences in the coalescence of alleles within each paralog were observed but were not taken into account, as they have no influence on the arrangement of clades.

The phylogenetic network constructed via the NeighborNet method implemented in the program SplitsTree4 [4] indicated that the preproalvinellacin is encoded by a multigenic family of six genes (par1, par2, par3a, par3b, par4 and par5), some of the alleles being recently derived recombinants (Par5 R2) while others (Par5 R1, Par4-1 R) represent older recombinants that have already accumulated their proper set of mutations. Twelve unambiguous sequences were kept for par1, 6 for par2, 13 for par3a, 9 for par3b, 14 for par4 and 25 for par5. The length of alleles dramatically varied between paralogs, mainly because of indels in the intronic regions. No indel was depicted in the exonic regions with the exception of Par2 which lacks a piece of 34 codons located at the end of the first exon and the beginning of the second one without changing the reading frame. Par4 displayed the lowest length due to a major deletion in the first intronic region. Par5 was the most divergent lineage mainly because of the first intron, which exhibited a tandem repeat region and could not be aligned with the other sequences (foreign insertion due to an unequal crossing over with another gene).

Strength of selection along the preproalvinellacin gene

Intensity of selection acting on each domain of the gene (i.e. signal peptide, propiece, BRICHOS and AMP) according to each paralog was measured using the ratio of non-synonymous substitution rate (d_N), which are usually subject to selective pressure, and the synonymous substitution rate (d_S), which is assumed to be (nearly) neutral [5,6]. Values greater than one were assumed to show positive diversifying selection on the divergence between two paralogous domains, and thus positive diversification of duplicates.

Search for positive selection in the propiece and BRICHOS regions of preproalvinellacin

Paralogous consensus sequences of the propiece (79 sequences) and the BRICHOS plus the alvinellacin AMP (36 sequences) were aligned together using ClustalW of the alignment

module of Geneious for *A. pompejana* and *A. caudata*. All positions with less than 95% site coverage were eliminated, leading to a total of 681 site positions in the final dataset. A search for the best model of substitutions was performed with jModelTest 2.1.7 and the tree topologies obtained from the PhyML reconstruction with these models were compared to the topology obtained with the GTR+I+G, as implemented in MEGA6 (Tables S2 and S3). Using the BRICHOS alignment, a hLRT backward selection procedure showed that none of the nested substitution models had significantly poorer goodness-of-fit than the GTR+I+G model (log likelihood=-439.625, BIC=1299.380) with the exception of the JC+I+G model (log likelihood= -452.175). Because the best model was the K80+I according to the BIC (log likelihood=-441.781, BIC=1261.15), we carefully examined the topology of the BIC-based best tree when compared to the GTR+I+G used for CodeML and aaML analyses (Fig. S3). Though more simplistic, the K80+I topology between the paralogous clades was not different from the one given by the GTR+I+G model and thus does not affect either the ancestral reconstruction or the search for positive selection on codons. For the propiece CodeML analysis (79 sequences), the AIC-based best model was also the K80+I model (log likelihood=-607.94, AICc= 3657.6) but the TPM2+I+G model (log likelihood=-588.29, BIC=1997.1) according to the BIC. Both models had better goodness-of-fit criteria than the GTR+I+G model (log likelihood=-579.8, AICc=5402.5, BIC=2016.5). However, hLRTs in backward selection indicated no significant differences in goodness-of-fit between nested models until the Tim2ef+I+G model. Tree reconstruction with this specific model did not modify the topology of the reference tree used for the Propiece analyses. Models H80+I and Tim2ef+I+G were then used for the tree reconstruction of a smaller set of duplicate-specific consensus sequences for either the BRICHOS or the Propiece domain (8 sequences each) and, subsequently used in the CodeML analyses for the article in order to remove polymorphic sequences from the analysis (Figs S3 and S4). Results from the small sets of consensus

sequences are now provided in Table 2 and results from all sets of sequences (including polymorphic ones) are provided in Tables S6 and S7. Several nested models of codon selection (i.e. M_3, M_{2a} and M_8) were subsequently tested against their 'nearly neutral' counterparts (i.e. M_0, M_{1a} and M_7, respectively) using a likelihood ratio test (LRT). The codon sequence dataset was first fitted on the 'nearly neutral' model M_{1a}, which divides codon sites into two categories, those under purifying selection ($\omega_0<1$) and the others under relaxed selection ($\omega_1=1$) using our reference tree. This model was then compared to the alternative nested 'selection' model M_{2a} where a third category of codon sites under positive selection ($\omega_2>1$) is added, thus accommodating positively selected sites. More sophisticated alternative nested models - the 'nearly neutral beta' M_7 and the 'selective beta' M_8 - were also compared. These models assume an omega distribution that follows a β (p,q) distribution with the shape parameter estimated in the interval [0, 1]. The difference between the two nested models lies in the fact that M_8 includes one additional substitution rate ω_1 with a probability p_1 that accounts for positively selected sites. In these two models, the rate of synonymous substitutions (d_S) is fixed among sites, while the rate of non-synonymous substitutions (d_N) remained variable along the gene. The significance of selection models using a likelihood ratio test with a degree of freedom equal to the difference between the number of parameters estimated for the 2 models when comparing M_{2A} vs. M_{1A} and, M_8 vs. M_7 (adapted from [7]).

MacDonald-Kreitman test between pairs of paralogs

Another approach to detect signs of positive selection, the MacDonald-Kreitman test, was also performed onto the propiece and BRICHOS regions of the *preproalvinellacin* taking advantage of the fixed divergence between paralogs. Under strict neutrality, both rates of synonymous and non-synonymous substitutions are expected to be equal in either the species divergence or the within-species polymorphisms, but d_N/d_S would become much greater than p_N/p_S under positive diversifying selection. None of the tests was significant but MK test can

be easily biased if the constancy of the neutral accumulation of mutations is not met over time for duplicates [8]. In order to test whether selective relaxation has occurred prior to the duplication events, we performed a Branch model analysis with CodeML by comparing the 'free ratio' model M_1 to the 'one ratio' model M_0 and checked whether the d_N/d_S ratios were higher in the internal branches leading to the duplicates when compared to their associated terminal branches. For the BRICHOS domain, M_1 (log likelihood=-370.654, np=29) was not significantly better (LRT=9.44, df=13, NS) than M_0 (log likelihood=-375.374, np=16), with a nearly neutral evolution before and after duplication (overall omega=0.64). For the Propiece domain, M_1 (log likelihood=-396.573, np=29) was also not significantly better (LRT=19.84, df=13, NS) than M_0 (log likelihood=-406.491, np=16), but its evolution was even more relaxed with positive selection (overall omega=1.18). In this case, terminal branches (6 out of eight with omega>1 for the Propiece) produced much higher d_N/d_S ratios than their internal counterparts when using the M_1 branch-model.

SUPPLEMENTARY FIGURES AND TABLES

Figure S1. BRICHOS and alvinellacin amino-acid alignment. Sequence labels represent the individual and clone number and are representative of the 6 paralogous clades (excluding natural recombinants, see Fig. 2) subsequently used in the mapping of the BRICHOS mutations. Ac: *A. caudata*; Ap *A. pompejana*

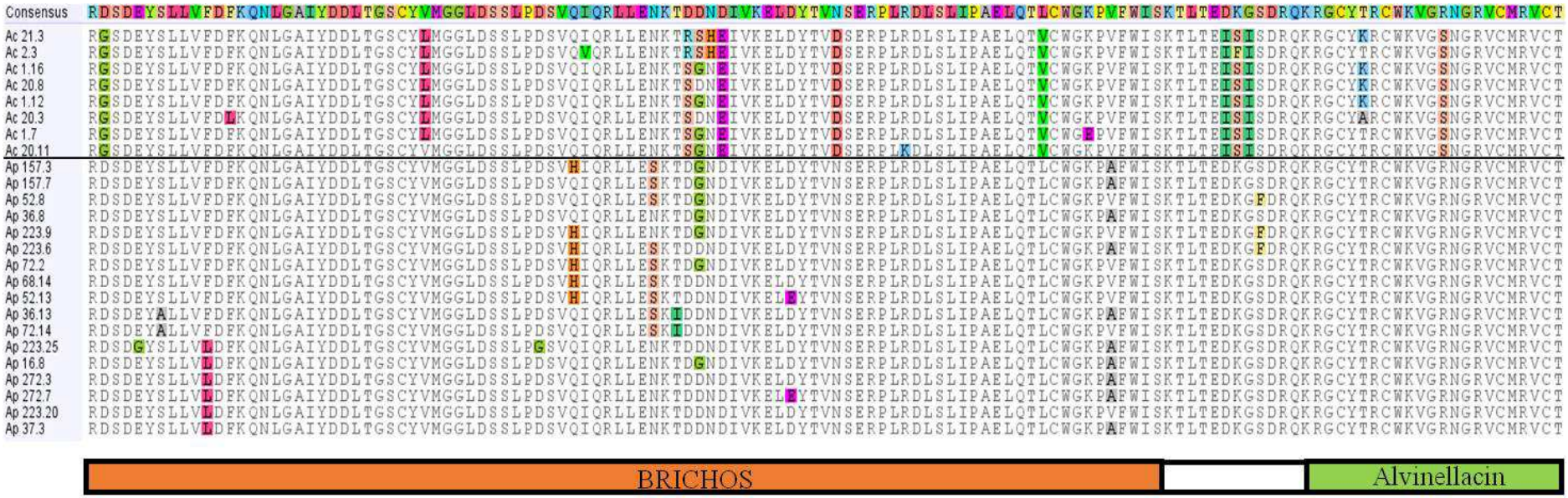


Figure S2. Tree topology comparisons between GTR+I+G and the AIC- and BIC-based best models for the most variable 5' region of the preproalvinellacin used in the identification of paralogous genes. Comparison between topologies of the alvinellacin paralogous MEGA tree obtained in Figure 2 using (**A**) the GTR+I+G model implemented in MEGA 6.0 (log likelihood= 5367.044, AIC=14030.71, BIC=14981.24), (**B**) the selected best model (GTR+G) obtained with jModelTest v2.1.7 based on the AIC criterion (log likelihood=-5365.93, AIC=14025.04) and, (**C**) the selected best model (TPM3uf+G) obtained with jModelTest v2.1.7 based on the BIC criterion (log likelihood=-5368.27, BIC=14963.74). Note that these three models fall within the 95% confidence intervals of the jModelTest analysis, and that the three models are not significantly different according to hierarchical likelihood ratio tests (LRT). Slight differences can be observed in the gene genealogies of each paralogous clade but do not affect the clade rearrangement.

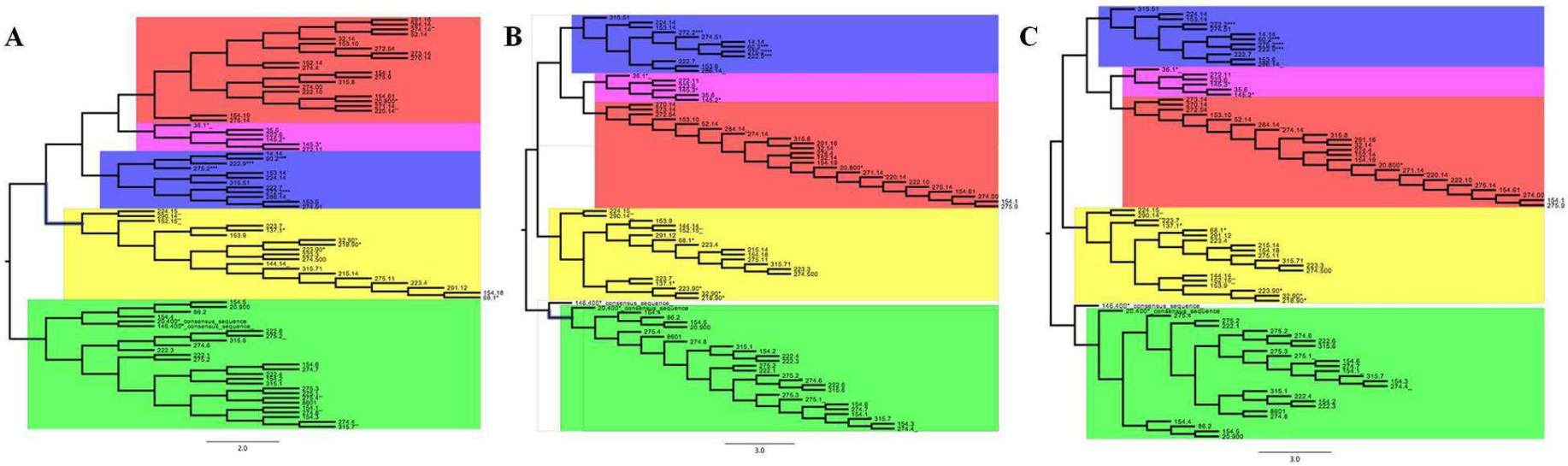


Fig. S3. PhyML tree of the BRICHOS domain using the K80+I model selected by jmodeltest according to the BIC. Topology of BRICHOS trees obtained using the K80+I model selected by jModelTest 2.1.7 according to the BIC criterion (log likelihood=-441.781, BIC=1261.15). This tree topology was not different from the one obtained using the GTR+I+G model implemented in MEGA 6.0 (Fig. 4 in main text) and led to the exact same conclusions when used as the reference tree in the CodeML and aaML analyses.

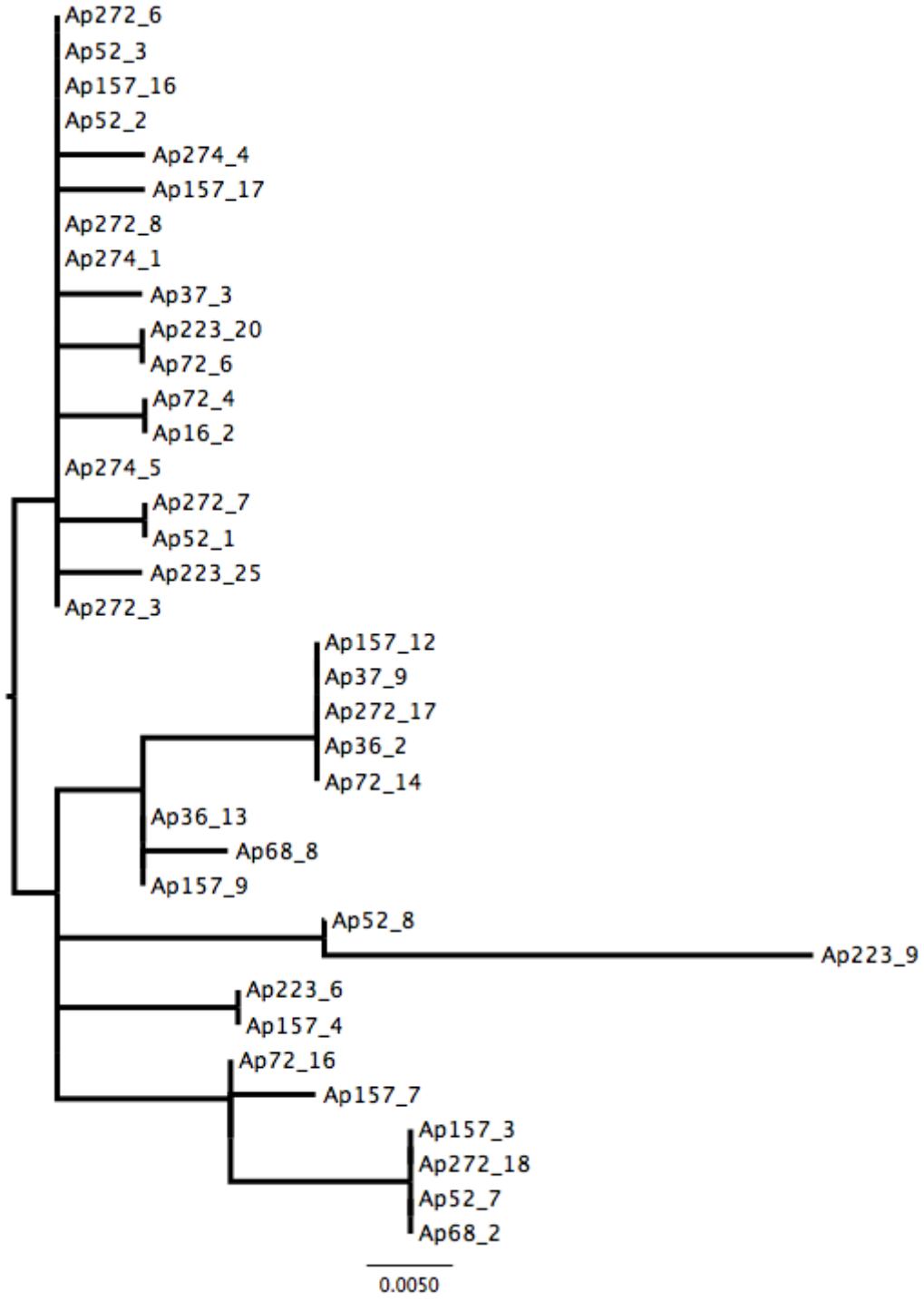


Fig. S4. Propiece ML tree topology using the GTR+I+G model implemented in MEGA 6.0 and the BIC-based best model (TPM2ef+I+G) obtained from the jModelTest analysis. Both models led to the same topology and all the 88 substitution models tested fall within the 95% confidence interval of the AIC/BIC analyses (see text). (**A**) Ultrametric and (**B**) chronogram representations of the same tree topology.

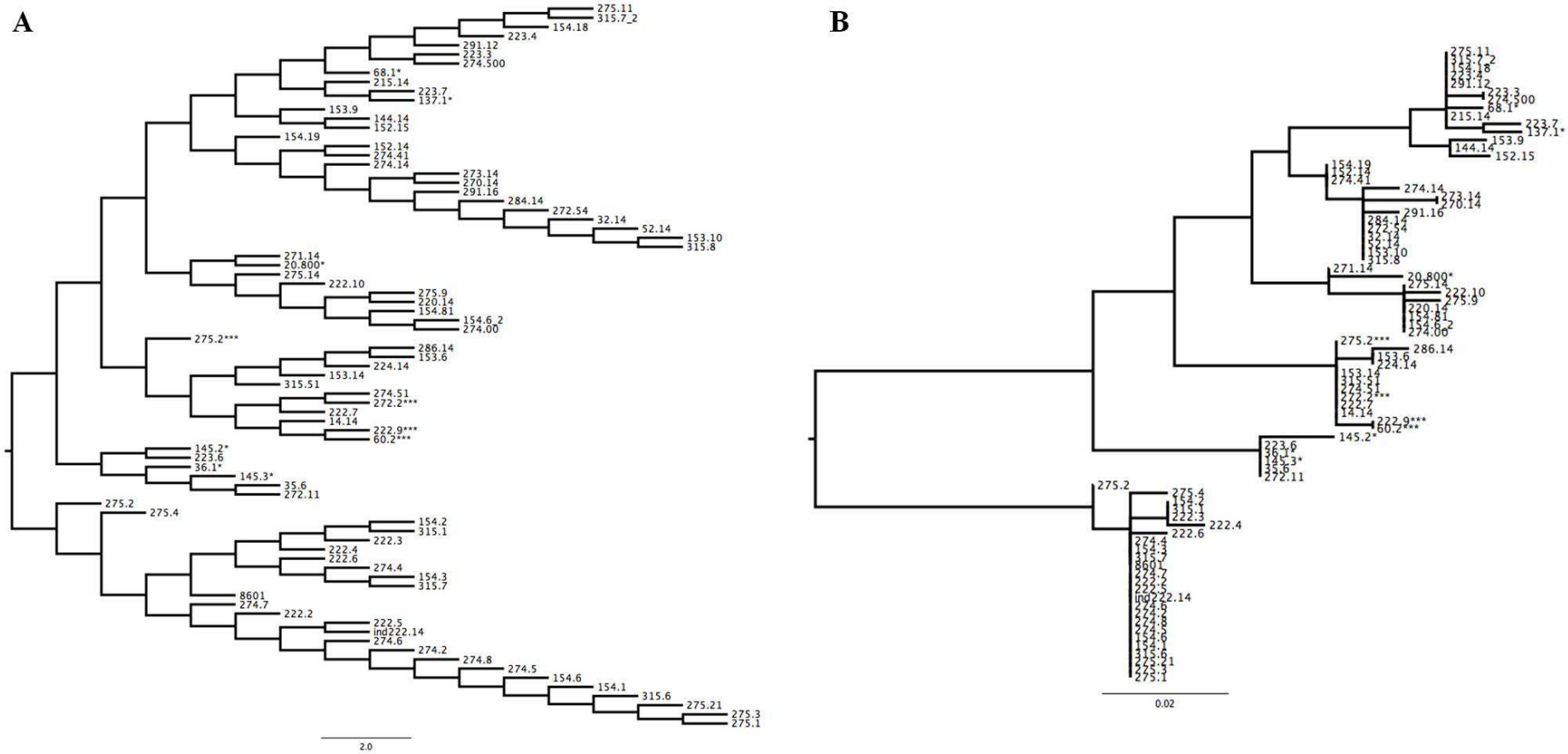


Table S1. Primers sequences (Forward and Reverse) used for the amplification of the preproalvinellacin gene from both *Alvinella pompejana* and *Alvinella caudata* and used for the genotyping of each paralogs (Px) for the *Alvinella pompejana* species.

Alvinella pompejana		
	3' Forward	ATCGTGTTACGTCATGGGTGGCCTTG
	3' Reverse	CTCAGTGAAATGAAGCAGGTGAGTTATG
	5' Forward	ATGACGTATTCTGTAGTTGTGACGCTGGTC
	5' Reverse	ATCCGGTAAGATCGTCGTAAATGGCTCC
Genotyping		
	P1_Forward	ACATCTACAGATTGGTGCTATCGAC
	P2_Forward	CTACAGATTGGTGCAGCCGAC
	P3_Forward	CATCTACAGATTGGTGCTGTGGAT
	P4_Forward	AACAGATTGGTGCTGTCGCC
	P5_Forward	TTTACATAGATTGGTGTTTCTTCTCTGAG
	P1_Reverse	GTTGAGGTGGCCAGCTGC
	P2-Reverse	GTTGGGGTGGCCAACTGC
	P3_Reverse	ATGTTGGGGTGATCAGCTGC
	P4_Reverse	GATGTTGAGGTGGCCAGCTAT
	P5_Reverse	GTTTCATGAAATGTGGCAGATG
Alvinella caudata		
	5' Forward	GTTACGTATTCTGTAGTCACGACGCTG
	5' Reverse	GGTAAGATCGTCGTAAATGGCTCC
	3' Forward	GTCGTGTTACCTGATGGGTGGC
	3' Reverse	AATATGCCAAAACAGGCGAATTACG

Table S2. AIC-based goodness-of-fit indicators of substitution models for the most variable 5' region of the preproalvinellacin used in the identification of paralogous genes, obtained from jModelTest. Models are ordered by increasing AIC.

Model	- log likelihood	number of parameters	AIC	ΔAIC	Akaike weight	Cumulative weight
GTR+G	6833.52	179	14025.04	0.00	0.75	0.75
TVM+G	6836.38	178	14028.75	3.71	0.12	0.87
GTR+I+G	6835.36	180	14030.71	5.67	0.04	0.92
TIM3+G	6838.39	177	14030.79	5.75	0.04	0.96
TVM+I+G	6836.73	179	14031.46	6.42	0.03	0.99
TPM3uf+G	6841.17	176	14034.33	9.29	0.01	1.00
TIM3+I+G	6840.53	178	14037.07	12.03	0.00	1.00
TPM3uf+I+G	6841.72	177	14037.43	12.39	0.00	1.00
TIM2+G	6842.97	177	14039.95	14.91	0.00	1.00
TPM2uf+G	6844.39	176	14040.79	15.75	0.00	1.00
TrN+G	6847.11	176	14046.22	21.18	0.00	1.00
TIM2+I+G	6845.54	178	14047.08	22.04	0.00	1.00
TPM2uf+I+G	6847.15	177	14048.29	23.25	0.00	1.00
HKY+G	6850.36	175	14050.72	25.68	0.00	1.00
TIM1+G	6848.95	177	14051.90	26.86	0.00	1.00
TPM1uf+G	6850.07	176	14052.15	27.11	0.00	1.00
TrN+I+G	6849.80	177	14053.60	28.56	0.00	1.00
GTR+I	6848.00	179	14054.00	28.96	0.00	1.00
HKY+I+G	6851.19	176	14054.38	29.34	0.00	1.00
TIM1+I+G	6849.49	178	14054.97	29.93	0.00	1.00
TVM+I	6849.67	178	14055.35	30.31	0.00	1.00
TPM1uf+I+G	6850.88	177	14055.76	30.72	0.00	1.00
TIM3+I	6853.24	177	14060.49	35.45	0.00	1.00
TPM3uf+I	6854.83	176	14061.66	36.63	0.00	1.00
TIM2+I	6858.26	177	14070.53	45.49	0.00	1.00
TPM2uf+I	6859.96	176	14071.93	46.89	0.00	1.00
TrN+I	6862.51	176	14077.01	51.97	0.00	1.00
HKY+I	6864.07	175	14078.14	53.10	0.00	1.00
TIM1+I	6862.28	177	14078.57	53.53	0.00	1.00
TPM1uf+I	6863.76	176	14079.51	54.47	0.00	1.00
TPM3+I+G	6870.38	174	14088.76	63.72	0.00	1.00
TIM3ef+I+G	6869.91	175	14089.82	64.78	0.00	1.00
TVMef+I+G	6869.05	176	14090.10	65.06	0.00	1.00
SYM+I+G	6868.59	177	14091.18	66.14	0.00	1.00
TPM3+G	6873.36	173	14092.73	67.69	0.00	1.00
K80+I+G	6873.63	173	14093.25	68.21	0.00	1.00
TPM1+I+G	6872.68	174	14093.36	68.32	0.00	1.00
TIM3ef+G	6872.94	174	14093.88	68.84	0.00	1.00
TIM1ef+I+G	6872.21	175	14094.42	69.38	0.00	1.00
TVMef+G	6872.27	175	14094.54	69.50	0.00	1.00
TrNef+I+G	6873.27	174	14094.55	69.51	0.00	1.00
TPM2+I+G	6873.52	174	14095.05	70.01	0.00	1.00

SYM+G	6871.85	176	14095.69	70.65	0.00	1.00
TIM2ef+I+G	6873.06	175	14096.12	71.08	0.00	1.00
K80+G	6876.61	172	14097.22	72.18	0.00	1.00
TPM1+G	6875.85	173	14097.70	72.66	0.00	1.00
TrNef+G	6876.18	173	14098.37	73.33	0.00	1.00
TIM1ef+G	6875.42	174	14098.84	73.80	0.00	1.00
TPM2+G	6876.54	173	14099.08	74.04	0.00	1.00
TIM2ef+G	6876.12	174	14100.24	75.20	0.00	1.00
TPM3+I	6889.00	173	14123.99	98.95	0.00	1.00
TIM3ef+I	6888.42	174	14124.83	99.79	0.00	1.00
TVMef+I	6887.70	175	14125.40	100.36	0.00	1.00
SYM+I	6887.12	176	14126.24	101.20	0.00	1.00
K80+I	6892.58	172	14129.16	104.12	0.00	1.00
TPM1+I	6891.71	173	14129.42	104.38	0.00	1.00
TIM1ef+I	6891.14	174	14130.29	105.25	0.00	1.00
TrNef+I	6892.23	173	14130.45	105.41	0.00	1.00
TPM2+I	6892.44	173	14130.89	105.85	0.00	1.00
TIM2ef+I	6891.88	174	14131.77	106.73	0.00	1.00
F81+G	6931.81	174	14211.62	186.58	0.00	1.00
F81+I+G	6932.11	175	14214.22	189.18	0.00	1.00
F81+I	6945.08	174	14238.17	213.13	0.00	1.00
TVM	6943.77	177	14241.53	216.49	0.00	1.00
GTR	6943.17	178	14242.33	217.29	0.00	1.00
TPM3uf	6948.84	175	14247.69	222.65	0.00	1.00
TIM3	6948.29	176	14248.57	223.53	0.00	1.00
JC+G	6954.45	171	14250.90	225.86	0.00	1.00
JC+I+G	6955.38	172	14254.76	229.73	0.00	1.00
TPM2uf	6956.76	175	14263.52	238.48	0.00	1.00
TIM2	6956.10	176	14264.19	239.15	0.00	1.00
HKY	6960.49	174	14268.98	243.94	0.00	1.00
TrN	6959.87	175	14269.75	244.71	0.00	1.00
TPM1uf	6960.14	175	14270.28	245.24	0.00	1.00
TIM1	6959.54	176	14271.08	246.04	0.00	1.00
JC+I	6970.40	171	14282.79	257.75	0.00	1.00
TPM3	6987.80	172	14319.61	294.57	0.00	1.00
TVMef	6985.82	174	14319.64	294.60	0.00	1.00
TIM3ef	6987.64	173	14321.27	296.23	0.00	1.00
SYM	6985.65	175	14321.31	296.27	0.00	1.00
TPM1	6991.78	172	14327.57	302.53	0.00	1.00
K80	6993.01	171	14328.01	302.97	0.00	1.00
TIM1ef	6991.59	173	14329.18	304.14	0.00	1.00
TPM2	6992.77	172	14329.54	304.50	0.00	1.00
TrNef	6992.87	172	14329.74	304.70	0.00	1.00
TIM2ef	6992.58	173	14331.15	306.11	0.00	1.00
F81	7034.17	173	14414.33	389.29	0.00	1.00
JC	7061.49	170	14462.99	437.95	0.00	1.00

Table S3. BIC-based goodness-of-fit indicators of substitution models for the most variable 5' region of the preproalvinellacin used in the identification of paralogous genes, obtained from jModelTest. Models are ordered by increasing BIC.

Model	- log likelihood	number of parameters	AIC	ΔAIC	Akaike weight	Cumulative weight
TPM3uf+G	6841.17	176	14963.74	0.00	0.62	0.62
TIM3+G	6838.39	177	14965.47	1.73	0.26	0.88
TVM+G	6836.38	178	14968.72	4.98	0.05	0.93
TPM2uf+G	6844.39	176	14970.19	6.45	0.02	0.96
GTR+G	6833.52	179	14970.28	6.55	0.02	0.98
TPM3uf+I+G	6841.72	177	14972.12	8.38	0.01	0.99
TIM2+G	6842.97	177	14974.63	10.89	0.00	0.99
HKY+G	6850.36	175	14974.84	11.11	0.00	1.00
TrN+G	6847.11	176	14975.62	11.89	0.00	1.00
TVM+I+G	6836.73	179	14976.70	12.97	0.00	1.00
TIM3+I+G	6840.53	178	14977.03	13.30	0.00	1.00
GTR+I+G	6835.36	180	14981.24	17.50	0.00	1.00
TPM1uf+G	6850.07	176	14981.55	17.81	0.00	1.00
TPM2uf+I+G	6847.15	177	14982.98	19.24	0.00	1.00
HKY+I+G	6851.19	176	14983.79	20.05	0.00	1.00
TIM1+G	6848.95	177	14986.58	22.84	0.00	1.00
TIM2+I+G	6845.54	178	14987.04	23.31	0.00	1.00
TrN+I+G	6849.80	177	14988.28	24.55	0.00	1.00
TPM1uf+I+G	6850.88	177	14990.44	26.70	0.00	1.00
TPM3uf+I	6854.83	176	14991.07	27.33	0.00	1.00
TIM1+I+G	6849.49	178	14994.94	31.20	0.00	1.00
TIM3+I	6853.24	177	14995.17	31.43	0.00	1.00
TVM+I	6849.67	178	14995.31	31.57	0.00	1.00
GTR+I	6848.00	179	14999.25	35.51	0.00	1.00
TPM2uf+I	6859.96	176	15001.33	37.59	0.00	1.00
HKY+I	6864.07	175	15002.26	38.52	0.00	1.00
TIM2+I	6858.26	177	15005.21	41.47	0.00	1.00
K80+G	6876.61	172	15005.50	41.76	0.00	1.00
TPM3+G	6873.36	173	15006.29	42.55	0.00	1.00
TrN+I	6862.51	176	15006.41	42.68	0.00	1.00
K80+I+G	6873.63	173	15006.81	43.08	0.00	1.00
TPM3+I+G	6870.38	174	15007.60	43.86	0.00	1.00
TPM1uf+I	6863.76	176	15008.91	45.18	0.00	1.00
TPM1+G	6875.85	173	15011.26	47.52	0.00	1.00
TrNef+G	6876.18	173	15011.93	48.19	0.00	1.00
TPM1+I+G	6872.68	174	15012.20	48.47	0.00	1.00
TPM2+G	6876.54	173	15012.64	48.91	0.00	1.00
TIM3ef+G	6872.94	174	15012.72	48.98	0.00	1.00
TIM1+I	6862.28	177	15013.25	49.52	0.00	1.00
TrNef+I+G	6873.27	174	15013.39	49.65	0.00	1.00
TPM2+I+G	6873.52	174	15013.89	50.15	0.00	1.00
TIM3ef+I+G	6869.91	175	15013.95	50.21	0.00	1.00

TIM1ef+G	6875.42	174	15017.68	53.95	0.00	1.00
TIM1ef+I+G	6872.21	175	15018.54	54.80	0.00	1.00
TVMef+G	6872.27	175	15018.66	54.92	0.00	1.00
TIM2ef+G	6876.12	174	15019.08	55.34	0.00	1.00
TVMef+I+G	6869.05	176	15019.50	55.77	0.00	1.00
TIM2ef+I+G	6873.06	175	15020.24	56.51	0.00	1.00
SYM+G	6871.85	176	15025.10	61.36	0.00	1.00
SYM+I+G	6868.59	177	15025.86	62.12	0.00	1.00
K80+I	6892.58	172	15037.44	73.70	0.00	1.00
TPM3+I	6889.00	173	15037.55	73.82	0.00	1.00
TPM1+I	6891.71	173	15042.98	79.24	0.00	1.00
TIM3ef+I	6888.42	174	15043.67	79.94	0.00	1.00
TrNef+I	6892.23	173	15044.01	80.28	0.00	1.00
TPM2+I	6892.44	173	15044.45	80.71	0.00	1.00
TIM1ef+I	6891.14	174	15049.13	85.39	0.00	1.00
TVMef+I	6887.70	175	15049.52	85.79	0.00	1.00
TIM2ef+I	6891.88	174	15050.61	86.87	0.00	1.00
SYM+I	6887.12	176	15055.64	91.90	0.00	1.00
F81+G	6931.81	174	15130.46	166.73	0.00	1.00
F81+I+G	6932.11	175	15138.34	174.60	0.00	1.00
JC+G	6954.45	171	15153.90	190.16	0.00	1.00
F81+I	6945.08	174	15157.01	193.27	0.00	1.00
JC+I+G	6955.38	172	15163.04	199.31	0.00	1.00
TPM3uf	6948.84	175	15171.81	208.07	0.00	1.00
TVM	6943.77	177	15176.21	212.48	0.00	1.00
TIM3	6948.29	176	15177.98	214.24	0.00	1.00
GTR	6943.17	178	15182.30	218.56	0.00	1.00
JC+I	6970.40	171	15185.79	222.05	0.00	1.00
TPM2uf	6956.76	175	15187.64	223.90	0.00	1.00
HKY	6960.49	174	15187.82	224.09	0.00	1.00
TIM2	6956.10	176	15193.60	229.86	0.00	1.00
TrN	6959.87	175	15193.87	230.13	0.00	1.00
TPM1uf	6960.14	175	15194.40	230.66	0.00	1.00
TIM1	6959.54	176	15200.48	236.75	0.00	1.00
TPM3	6987.80	172	15227.89	264.15	0.00	1.00
K80	6993.01	171	15231.01	267.27	0.00	1.00
TIM3ef	6987.64	173	15234.83	271.10	0.00	1.00
TPM1	6991.78	172	15235.85	272.11	0.00	1.00
TPM2	6992.77	172	15237.82	274.08	0.00	1.00
TrNef	6992.87	172	15238.02	274.29	0.00	1.00
TVMef	6985.82	174	15238.48	274.74	0.00	1.00
TIM1ef	6991.59	173	15242.74	279.01	0.00	1.00
TIM2ef	6992.58	173	15244.71	280.98	0.00	1.00
SYM	6985.65	175	15245.43	281.69	0.00	1.00
F81	7034.17	173	15327.89	364.16	0.00	1.00
JC	7061.49	170	15360.71	396.97	0.00	1.00

Table S4. D_{xy} between paralogs

	par5	par1	par2	par3a	par3b	par4
par5						
par1	0.306					
par2	0.374	0.0921				
par3a	0.307	0.0752	0.122			
par3b	0.304	0.0769	0.111	0.0169		
par4	0.281	0.046	0.0987	0.0759	0.07789	

Table S5. Pairwise comparisons of K_a/K_s ratios between paralogs (par) for each domain of the gene. SP: signal peptide; PR: propiece; BRICHOS: BRICHOS domain, AMP: Antimicrobial Peptide mature domain. M refers to monomorphic sequences: no genetic diversity can be depicted for the pairwise comparison.

Region	Domain	Paralog			Paralog			
			par5	par1	par2	par3a	par3b	par4
		par5						
		par1	1.5842					
	SP	par2	0.3786	0.8005				
		par3a	0.3597	0.7612	0			
		par3b	0.3597	0.7612	0	0		
5’		par4	0.7772	0.7612	0	0	0	
			par5	par1	par2	par3a	par3b	par4
		par5						
		par1	1.5667					
	PR	par2	1.9158	0.858				
		par3a	1.3084	0.2798	0.3757			
		par3b	1.0855	0.3998	0.4868	1.4101		
		par4	0.9976	0.327	0.492	0.252	0.5152	
			parE	parA	parB	parC	parD	
		parE						
	BRICHOS	parA	0.5741					
		parB	1.8245	0.214				
		parC	0.5629	0.4216	0.8065			
3’		parD	0.8603	0.8649	1.8177	0.2782		
			parE	parA	parB	parC	parD	
		parE						
	AMP	parA	M					
		parB	M	M				
		parC	M	M	M			
		parD	M	M	M	M		

Table S6. Log-likelihood values and parameter estimates for the BRICHOS domain region and the propiece of the preproalvinellacin gene. Maximum-likelihood models implemented in the codeML program of the PAML package for models that allow positive selection (M2, M3,M8) and those that do not (M0, M1, M7). M0, one-ratio; M1, neutral; M2, selection; M3, discrete; M7, β; M8, β+ω and the estimated log-likelihood values (l) by the codeml program, ω = dN/dS nonsynonymous/synonymous rate ratio; p = proportion of sites for each site class. M0: one estimated ω for all sites; M1a: estimate p_0 = proportion of sites with $\omega_0 = 0$,$p_1 = 1 - p_0$, proportion of sites with $\omega_1 = 1$; M2a: estimate p_0 ($\omega_0 = 0$), p_1 ($\omega_0 = 1$), and ω_2, $p_2 = 1 - p_0 - p_1$. M3: estimate p_0, p_1, ω_0, ω_1, and ω_2; $p_2 = 1 - p_0 - p_1$. M7: estimates p and q (parameters of β distribution of ω between 0 and 1). M8: same as M7 except additional site class where an estimated ω is allowed. Positively Selected Sites: Codon positions predicted to be under positive selection with a posterior probability of ** >0.99 and *>95% (identification of sites exhibiting dn/dS ratio >1). Sites refer to amino acids positions from the first M. *: P>95%; **: P>99 in the Naive Empirical Bayes (NEB) analyses of PAML.

Model		M0	M3 (Discrete)	M1 (Neutral)	M2 (Selection)	M7 (β)	M8 (β+ω)
BRICHOS	**Log likelihood**	-443.76	-437.09	-439.14	-437.09	-439.17	-437.09
	Parameters estimates	ω= 0,612	ω0=0, p0=0.79, ω1=2.78772 p1=0.18, ω2=2.78775 p2=0.03	ω0= 0, p0=0.67, (ω1=1) p1=0.33	ω0=0, p0=0.79, ω1=1 p1=0, ω2=2.78 p2=0.21	p=0.005, q=0.0117	p0=0.78, p=0.005, p1=0.21, q= 2.74, ω=2.78
	Sites with dN/dS>1 (NEB analysis)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all **)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all **)	n.a.	D119G; Q121H; N129S; T131I; D133G; D141E; V169A (all **)
	Sites with dN/dS>1 (BEB analysis)	n.a.	n.a.	n.a.	Q121H; N129S; D133G; D141E; V169A (not significant)	n.a.	Q121H; N129S; D133G; D141E; V169A (not significant)
PROPIECE	**Log likelihood**	-593.27	-585.85	-591.82	-585.88	-592.88	-585.88
	Parameters estimates	ω=1.2438 5	ω0=0.49, p0=0.76, ω1=3.33 p1=0.14, ω2=7.32 p2=0.1	ω0= 0, p0=0.29, (ω1=1) p1=0.71	ω0=0.45, p0=0.61, ω1=1 p1=0.21, ω2=5.96 p2=0.18	p=0.012, q=0.005	p0=0.81, p=5.23, p1=0.19, q= 3.99, ω=5.85
	Sites with dN/dS>1 (NEB analysis)	n.a.	N22I; W24R; L26Q; N30S; A31V; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S	n.a.	W24R; L26Q; N30S; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S	n.a.	N22I; W24R; L26Q; N30S; A31V; H33D*; P38YS*; D57E; T60IAS**; Q68E*; H76RD; L78S
	Sites with dN/dS>1 (BEB analysis)	n.a.	n.a.	n.a.	W24R; L26Q; N30S; H33D; P38YS;T60IAS*; Q68E*; H76RD	n.a.	W24R; L26Q; N30S; H33D; P38YS;T60IAS*; Q68E*; H76RD; L78S

Table S7. Statistical likelihood ratio tests comparing substitution models on BRICHOS and Propiece sequences. The deviances (LRT) calculated from paired CodeML models are compared with the critical values of chi-square asymptotic distribution with appropriate degrees of freedom.

BRICHOS	**Models**		
	M0 versus M3	**M1 versus M2**	**M7 versus M8**
deviance	13.32	4.1	4.16
df	4	2	2
p-value	**0.0357**	0.1287	0.1249
PROPIECE	**Models**		
	M0 versus M3	**M1 versus M2**	**M7 versus M8**
deviance	14.84	11.88	14
df	4	2	2
p-value	**0.0005**	**0.0026**	**0.0009**

References

1 Tasiemski, A. *et al.* Characterization and function of the first antibiotic isolated from a vent organism: the extremophile metazoan *Alvinella pompejana*. *PLoS ONE* **9**, e95737, doi:10.1371/journal.pone.0095737 (2014).
2 Plouviez, S., Le Guen, D., Lecompte, O., Lallier, F. H. & Jollivet, D. 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 Evol Biol* **10**, 220, doi:10.1186/1471-2148-10-220 (2010).
3 Smith, J. M. Analyzing the mosaic structure of genes. *J Mol Evol* **34**, 126-129 (1992).
4 Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**, 254-267, doi:10.1093/molbev/msj030 (2006).
5 Nielsen, R. Molecular signatures of natural selection. *Annu Rev Genet* **39**, 197-218, doi:10.1146/annurev.genet.39.073003.112420 (2005).
6 Yang, Z. & Bielawski, J. P. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**, 496-503 (2000).
7 Parmakelis, A. *et al.* Anopheles immune genes and amino acid sites evolving under the effect of positive selection. *PLoS One* **5**, e8885, doi:10.1371/journal.pone.0008885 (2010).
8 Hahn, M. W. Distinguishing Among Evolutionary Models for the Maintenance of Gene Duplicates. *Journal of Heredity* **100**, 605-617, doi:10.1093/jhered/esp047 (2009).

Annexe 3. Valeurs de F_{st} calculées pour les clades de la capitellacine au sein des deux régions 5' et 3' par pairs de populations et de façon globale.

Région 5'

Clade B

Fst	Boulogne	Roscoff	PLF
Boulogne			
Roscoff	**0,49**		
PLF	**0,3**	**0,27**	

Fst global=0,32 pvalue=0,0001

0,008		
0,005	**0,0001**	

Clade A

Fst	Boulogne	Roscoff	PLF
Boulogne			
Roscoff	0,034		
PLF	0	0,02	

Fst globa =0,02 pvalue=0,35

0,61		
0,99	0,46	

Région 3'

Clade B

Fst	Boulogne	Roscoff	PLF
Boulogne			
Roscoff	0,16		
PLF	0,13	0,033	

Fst global=0,09 p value 0,023

0,11		
0,084	0,46	

Clade A

Fst	Boulogne	Roscoff	PLF
Boulogne			
Roscoff	0,11		
PLF	**0,23**	**0,18**	

Fst global=0,18 p value 0,0012

0,15		
0,033	**0,03**	

Annexe 4. Protocole de productions des variants du BRICHOS

PROTOCOLE

Les transcrits complets (obtenus dans l'Annexe 3 : totalité du codant du précurseur protéique de l'alvinellacine) avaient précédemment été insérés dans le vecteur pcr2.1 (selon le protocole classique du fournisseur : INVITROGEN TA CLONING KIT en bactérie TOP10). Ceci a été effectué en utilisant les amorces 5' F et 3'R du chapitre 2 qui permettent l'amplification de la séquence codante complète à partir des cDNA avec le même protocole que l'amplification d'ADN (2min d'élongation). A l'issu du clonage, chaque clone (contenant le transcrit inséré) avait été séquencé puis avait été glycérolé et conservé à -80°C.

A partir de ces transcrits immortalisés, la région codant le domaine BRICHOS a été amplifiée à l'aide des primers du tableau X pour les trois variants selon le protocole : Tampon PCR (1X), MgCl2 (2,0 mM), dNTP (0,4 µM), amorces (0,2µM chaque), 0,04U de Taq Uptitherm (Interchim) selon les conditions d'amplification par PCR suivantes : 94°C/3min ; 60°C/2min ; 72°C/3min suivi de 35 cycles de 94°C/30s ; 55°C/30s et 72°C/1min avec une élongation finale de 7min.

Amorces BRICHOS:	
5'-2-tous	AAAAAACCATGGTACGCGATAGTGATG
3' SNC10	TTTTTTCACTGGATCCAGAAGACAGGC
3'SNCs	TTTTTTCACTGGATCCAGAAGGCAGG
Amorces sur le vecteur:	
Pet32c_F	GCCAGCACATGGACAGCC
Pet32c_R	CCCATGGCTTAGCAGCCG

Tableau 1. Séquences des amorces utilisées pour la sur-expression des variants du BRICHOS

A ces primers, les sites de restrictions des enzymes BAMHI et NcoI avec ajout de 5A ont étés ajoutés pour pouvoir :

1) ré-insérer les séquences amplifiées (contenant uniquement les régions BRICHOS donc + tag) dans un premier temps en vecteur pcr2.1 (TA cloning, même protocole en TOP10). Après transformation, un criblage de clones blanc dans le but de vérifier qu'aucune mutation non synoyme s'est produite lors de l'amplification PCR a été effectué suivie d'une extraction plasmidique Miniprep selon le protocole du fournisseur QIAprep Miniprep (QIAGEN)

2) Ces extractions plasmidiques ont ensuite été utilisées dans le but de les digérer avec les enzymes de restrictions et les insérer dans le plasmide d'expression choisi après digestion : le pet32c. Le vecteur a été dephosphorilé ave la shrimp alkalyne phosphatase : 0,05U per pmol, 30min à 37°C.

Le protocole de digestion est le suivant : 5U de NcoI avec 5µg d'ADN, 1X de NEBuffer (protocole New England BioLabs) dans 10µL final. Après 15minutes de digestion à 37°C, l'enzyme BAMHI (5U) est ajoutée pour 15minutes de plus. L'inactivation des enzymes se fait pendant 20minutes à 80°C La digestion se fait de façon séquentielle car les deux sites de restriction sont séparés de moins de 10nucléotides.

Le plasmide ainsi que les trois variants ont été ligués à l'aide de la T4 DNA ligase de thermoFisher (20ng de vecteur,ADN ratio 1:1, 2µL de 10X T4 DNA ligase buffer, 1U T4 DNA ligase ; qsp 20µL). Le produit de ligation était incubé la nuit à 22°C.

1) Production des protéines

Après ligation, les produits étaient transformés dans des Escherichia coli Origami(DE3)pLysS (Novagen) selon un protocole standard du fournisseur de transformation (50µL de bactéries transformées par 20ng de vecteur d'expression par choc thermique de 40s à 42°C) et sont ensuite étalées sur un milieu LB contenant 100µg/mL d'ampiciline à 37°C sur la nuit. Le lendemain, des colonies sont criblées avec des amorces dessinées sur le plasmide (Tableau 1) et séquençées sur le sequençeur ABI3100 du laboratoire pour vérifier qu'aucune mutation non synonyme artéfactuelles n'a été ajoutée. Une fois vérifié par séquençage, une préculture est lancée dans 5mL de LB+ampicilline sur la nuit à 37°C. Quelques mL ont été récupérés ensuite et une extraction plasmidique a été effectuée selon le protocole du fournisseur QIAprep Miniprep (QIAGEN). La préculture est ensuite mise dans 1L de LB+amp et l'expression des protéines est induite en ajoutant 0.25mM d'ITPG (isopropylthiogalactoside) après que la D.O soit arrivée à 0.5 à 600nm).

2) Lyse des bactéries et séparation des fractions solubles et insolubles

Après 8h à 25°C et 250rpm, les cellules sont collectées par centrifugation (1000g, 20min, 4°C) et resuspendues dans du tampon contenant 20mM de sodium phosphate ph7.0 contenant 5mM d'imidazole. A ceci est ajouté du lysozyme (1mg/mL) incubé 1h sur glace

puis les culots bactériens sont passés à la Press de French. Cet appareil permet de déchirer les membranes des cellules (déjà fragilisé par l'action du lysozyme) en les forçant à passer dans un espace de petit diamètre à forte pression. Elles sont ensuite centrifugées (30min à 14000rpm à 4°C) pour séparer fraction soluble et insoluble. Le surnageant est récupéré et 20µL est mélangé à du Tampon Laemmli pour vérification par SDS-PAGE (12%acrylamide running gel, 4%acrylamine stacking gel). Une étape supplémentaire de vérification a été effectuée en réalisant un Western Blot avec un anticorps anti-BRICHOS.

3) Purification sur colonne de nickel –His Trap

La chromatographie de pseudo affinité sur résine de nickel HisTrap HP de 1mL (GE HealthCare Life Science) permet, après lavage de la résine (débit 1mL/min) avec un tampon de fixation (20mM sodium phosphate, 0,5M NacL, pH 7,4) contenant 5mM d'imidazole, une élution de la protéine taguée a une forte concentration en imidazole (même tampon avec 250mM d'imidazole). En effet, l'élution (sur 5 volumes de colonne à chaque fois) a été réalisée en faisant un gradient de 50mM (permettant l'élimination de la plupart des protéines non recombinante) puis à 250mM et à 500mM. Toutes ces élutions ont été collectées par fractions de 1mL et ont ensuite été analysées sur gel SDS-PAGE (12% acrylamide running gel, 4% acrylaminde stacking gel). Un lavage final est réalisé à 1M d'imidazole.

4) Concentration des protéines

Un filtre à centrifuger a été utilisé (MILLIPORE 10K) dans le but de concentrer les protéines après élution (15minutes à 4000g).

5) Clivage du vecteur d'expression

Après avoir quantifié la quantité de protéine par méthode Bradford, la protéase enterokinase (NEB) a été utilisée pour couper la partie en aval du domaine BRICHOS du vecteur (0,00014µg d'enterokinase pour 25µg de protéines, 16h à 25°C).

Annexe 5. Liste des articles et valorisations.

Articles dans l'ordre de publications :

Cascella, K., Jollivet, D., Papot, C., Léger, N., Corre, E., Ravaux, J., Clark, M.S., and Toullec, J.-Y. (2015). Diversification, evolution and sub-functionalization of 70kDa heat-shock proteins in two sister species of Antarctic krill: differences in thermal habitats, responses and implications under climate change. PLoS One *10*, e0121642.

Papot, C., Cascella, K., Toullec, J.-Y., and 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. Ecol. Evol. *6*, 1555–1575.

Papot, C., Massol, F., Jollivet, D., and Tasiemski, A. (2017). Antagonistic evolution of an antibiotic and its molecular chaperone: how to maintain a vital ectosymbiosis in a highly fluctuating habitat. Sci. Rep. *7*. 1454-1460.

Cuvillier-Hot, V., Gaudron, S.M., Massol, F., Boidin-Wichlacz, C., Pennel, T., Lesven, L., Net, S., Papot, C., Ravaux, J., and Vekemans, X. (2018). Immune failure reveals vulnerability of populations exposed to pollution in the bioindicator species Hediste diversicolor. Sci. Total Environ.

Valorisations scientifiques :

- Posters :

1 Poster au congrès international BSE/SFE, Lille, Décembre 2014.

- Communications orales :

1 Communication Orale au congrès Petit Pois Déridé, Orsay, Aout 2014.

2 Communications Orale à IMMUNINV, Dijon 2014 et Lille 2016.

1 Communication Orale au congrès international AMP 2016, Montpellier, Mai 2016.

2 Communications Orales au GDR MufoPAM, Orléans et Dourdan, 2014 et 2016.

1 Communication Orale au congrès CONNECT2, Plouzané, Avril 2016.

Résumé

Les peptides antimicrobiens (PAMs) font partie intégrante du système immunitaire inné de la plupart des organismes en constituant une première ligne de défense contre un large éventail d'agents pathogènes et peuvent également être impliqués dans le contrôle et/ou le confinement de la microflore symbiotique. Le but de cette thèse était d'étudier l'évolution moléculaire de deux gènes codant pour deux précurseurs protéiques de PAMs (preprocapitellacine et preproalvinellacine) caractérisés chez deux annélides extrémophiles : le ver côtier *Capitella spp (Cc)* et le ver hydrothermal *Alvinella pompejana (Ap).* Ces précurseurs à partir desquels sont maturés les PAM présentent une structure typique des protéines à BRICHOS: un peptide signal, une propiece, un domaine chaperon BRICHOS et un peptide en épingle à cheveux beta (ici le PAM). Les résultats ont montrés que le même type de mécanismes pourrait coexister entre les deux taxons annélides étudiés pour promouvoir et maintenir la diversité génétique des deux effecteurs immunitaires dans les différents domaines des précurseurs protéiques (duplication, recombinaison, sélection positive, introgression). Une différence majeure peut être mise en évidence dans la région du PAM qui est monomorphe (sélection purifiante) pour *Ap* et polymorphe pour *Cc*. Ceci serait dû à l'absolue nécessité de cultiver une communauté épibiotique hautement spécialisée et obligatoire pour le ver hydrothermal malgré des conditions abiotiques très fluctuantes alors que les espèces côtières de *Capitella spp* évoluent dans un environnement pathogène dans lequel la diversification de l'arsenal immunitaire constitue un avantage pour renforcer leur potentiel défensif.

Abstract

Antimicrobial peptides (AMP) are integral components of the innate immune system of most organisms in which they provide an early and a first line of defense against a wide range of microbial and microeukaryotic agents. They are also known to shape and control the symbiotic microflora. The aim of this thesis was to study the molecular evolution of two antimicrobial peptides encoding the genes: preproalvinellacin and preprocapitellacin that have been characterized from two annelids: the coastal species *Capitella spp* (*Cc*) and the hydrothermal species *Alvinella pompejana (Ap)*. These precursors from which are matured the AMPs alvinellacin and capitellacin, display an original structure of a BRICHOS chaperon: a signal peptide, a propiece, a BRICHOS domain and a beta hairpin peptide (here the AMP). Results show that the same kind of mechanisms might co-occur between the two distinct annelid taxa to promote and maintain genetic diversity for both immune effectors in the precursor domains (duplication, recombination, positive selection, introgression). One major difference can be highlighted in the AMP region that is strictly monomorphic for the *Ap* species (purifiying selection) and is highly polymorphic in the *Cc* species. This can be due to the absolute need of farming a highly specialized epibiotic community for the hydrothermal worm despite highly fluctuating abiotic conditions whereas the coastal species of *Capitella spp* evolved in a more pathogenic environment in which the immune arsenal diversification should be an advantage in enhancing their defensive potential.