L'apport de la génomique comparative dans la compréhension des interactions évolutives au sein des holobiontes

Jonathan Filée

Habilitation à diriger des recherches

université PARIS-SACLAY

L’apport de la génomique comparative dans la compréhension des interactions évolutives au sein des holobiontes

Habilitation à diriger des recherches de l'Université Paris-Saclay

présentée et soutenue à Gif-sur-Yvette, le 13/03/2023, par

JONATHAN FILÉE

Composition du jury

Eric BAPTESTE Directeur de Recherche CNRS-Sorbonne Université	Rapporteur
Richard CORDAUX Directeur de Recherche CNRS-Université de Poitiers	Rapporteur
Jean-Michel DREZEN Directeur de recherche CNRS-Université de Tours	Rapporteur
Mohammed JEBBAR Professeur Université de Bretagne Occidentale	Examinateur
Olivier LESPINET Professeur Université Paris-Saclay	Examinateur

Two centuries after Darwin's birth, 150 years after the publication of his 'Origin of Species', and 50 years after the consolidation of the Modern Synthesis, comparative analysis of hundreds of genomes from many diverse taxa offers unprecedented opportunities for testing the conjectures of (neo)Darwinism and deciphering the mechanisms of evolution.

Eugene V. Koonin (2009)

Biodiversité de mes travaux de recherche

Remerciements

En premier lieu, je voudrais remercier les membres du jury qui m'ont fait l'honneur d'accepter de juger ces travaux.

Ma gratitude est grande envers tout ceux qui, de prêt ou de loin, ont participé a tout les travaux de recherche qui ont aboutis aux publications qui sont décrites dans ce manuscrit. Je pense en premier lieu a tout les étudiants et étudiantes, aux personnels techniques ainsi qu'aux nombreux collaborateurs qui ont participé a ces travaux.

Je ne saurais continuer ces remerciements sans mentionner mes collègues de Gif (et d'ailleurs !) dont la bienveillance ont beaucoup compté tout au long de ma carrière au laboratoire EGCE. Je pense en particulier à la période qui a suivis ma longue absence du laboratoire après mes graves ennuis de santé : ils ont joué un rôle très important dans mon rétablissement.

Enfin je mesure la chance que j'ai d'avoir toujours eu autour de moi une famille aimante dont le support et les attentions ont (et vont) toujours jouer un rôle fondamental dans ma carrière scientifique.

Préambule

Ce manuscrit d'Habilitation a Diriger des Recherches (HDR) décris d'une façon synthétique mes travaux de recherche en génomique comparative et évolutive au seins des holobiontes. Le concept d'holobionte étant ici pris au sens large (ensembles des organismes cellulaires vivant en association ainsi que leurs virus et les autres éléments génétiques mobiles égoïstes). Il s'articule autour d'une brève introduction générale qui présente les grandes lignes des travaux de recherches qui seront ensuite détaillés dans la partie Résultats & Discution. Cette partie n'est pas complètement exhaustive sur mes travaux effectués, notamment ceux qui ont été réalisé en collaboration avec d'autres équipes sur des thématiques accessoires. J'ai donc fait le choix de présenter les résultats les plus marquants. Le dernier chapitre de cette partie sera consacré a mes travaux en génomique entomologique et présentera une partie de mes projets pour les prochaines années. A la fin de chaque chapitre consacré aux travaux effectués, une liste des publications auxquelles je suis associé sera proposée. Enfin, le lecteur trouvera en annexe un CV complet comprenant la liste complète de mes publications, encadrements, financements et participations a des tâches collectives.

Introduction générale

1. La génomique comparative: histoire et technologies

La génomique comparative pourrait se définir par la Lapalissade suivante: il s'agit de la science qui étudie les génomes et les compare entre eux. D'un point de vue historique, la génomique est la fille aînée de la génétique dont elle a hérité de nombreux paradigmes. A quoi les progrès technologiques en matière de séquençage de l'ADN ont ajouté un corpus considérable de données et de concepts qui ont fait entrer de plein pied la biologie dans la « science des données » (« data science »). La génomique a grandement participé, et bénéficié, a l'émergence de la bio-informatique , cette science pluri-disciplinaire qui a permis notamment d'automatiser et de traiter de vaste jeux de données de séquences. En ce sens, la génomique comparative a toujours avancé sur deux pieds : les progrès du séquençage d'une part, les progrès en terme d'analyse bio-informatique d'autre part.

Si le début de ma thèse en 2001 a coïncidé avec la publication du premier génome humain, l'obtention des premiers génomes complets de phages remontait aux années 70' avec la mise au point des techniques de séquençage de type Sanger[1]. Les années qui ont suivis ont vu le développement de technique de séquençage de deuxième et troisième générations qui, tout en comprimant les coûts, ont permis d'obtenir des lectures («reads») de plus en plus abondantes, et de plus en plus longues (mais parfois au prix d'erreurs de séquences)[2]. Nous sommes passé en moins de deux décennies d'une période où l'information génétique et génomique était rare et précieuse, a une période d'abondance où le surplus a rapidement engendré un goulot d'étranglement en matière de capacité d'analyse. Si l'on prends la base de données de séquences de référence RefSeq du NCBI, a l'heure où ces lignes sont écrites, nous pouvons y trouver plus de 250 000 génomes complets de procaryotes et plusieurs milliers de génomes eucaryotes. Il ne s'agit pourtant là que des génomes de références, c'est a dire présentant un niveau de complétude et d'erreur de séquences considéré comme optimaux! L'addition de tout les projets génomes de moins bonnes qualités, ainsi que des projets de métagénomes c'est a dire de communauté de génomes cohabitant dans un même environnement, feraient grimper ces chiffres de plusieurs niveau de magnitude!
Cet avènement du séquençage de masse, a bas coût, a historiquement toujours posé des problèmes d'analyse en terme de ressources de calcul, et secondairement en capacité d'archivage. C'est pour cette raison que les progrès dans les traitements informatiques de ces masses de données ont très tôt joué un rôle fondamental en génomique comparative. Il est a ce titre fascinant de voir comment la baisse spectaculaire du coût du séquençage dans les années 2000' a même été encore plus rapide que la loi empirique de Moore sur l'augmentation de la puissance de calculs des microprocesseurs (Figure 1) . Symétriquement, l'abondance des séquences des génomes a ouvert un vaste champ d'étude sur le développement de méthodes d'analyses et a participé a faire émerger la bioinformatique comme un discipline centrale et a part entière en biologie. Se n'est donc pas un hasard si le logiciel BLAST, utilisé pour fouiller des banques de séquences, est plus de 30 ans après sa publication, un des papiers les plus cité de l'histoire des sciences[3,4].

Pour ces raisons, on peut penser que la génomique comparative a été au centre de la révolution techno-scientifique des progrès conjoints du séquençage et de la bioinformatique. Un exemple emblématique de cette bipédie scientifique est donné par le récent prix Nobel de médecine 2022 au paléo-génomicien Svante Paabo dont les travaux pionniers sur l'ADN ancien ont combiné a la fois des progrès en matière de séquençage de l'ADN d'os fossiles a très faible concentrations et très dégradés, mais aussi

des innovations bioinformatiques dans la matière de corriger, d'assembler et d'analyser ces « reads » génomiques. Logiquement, cette science a débouché sur de très nombreuses découvertes dont un survol rapide en lien avec mes travaux de recherche sera proposé dans le chapitre suivant.

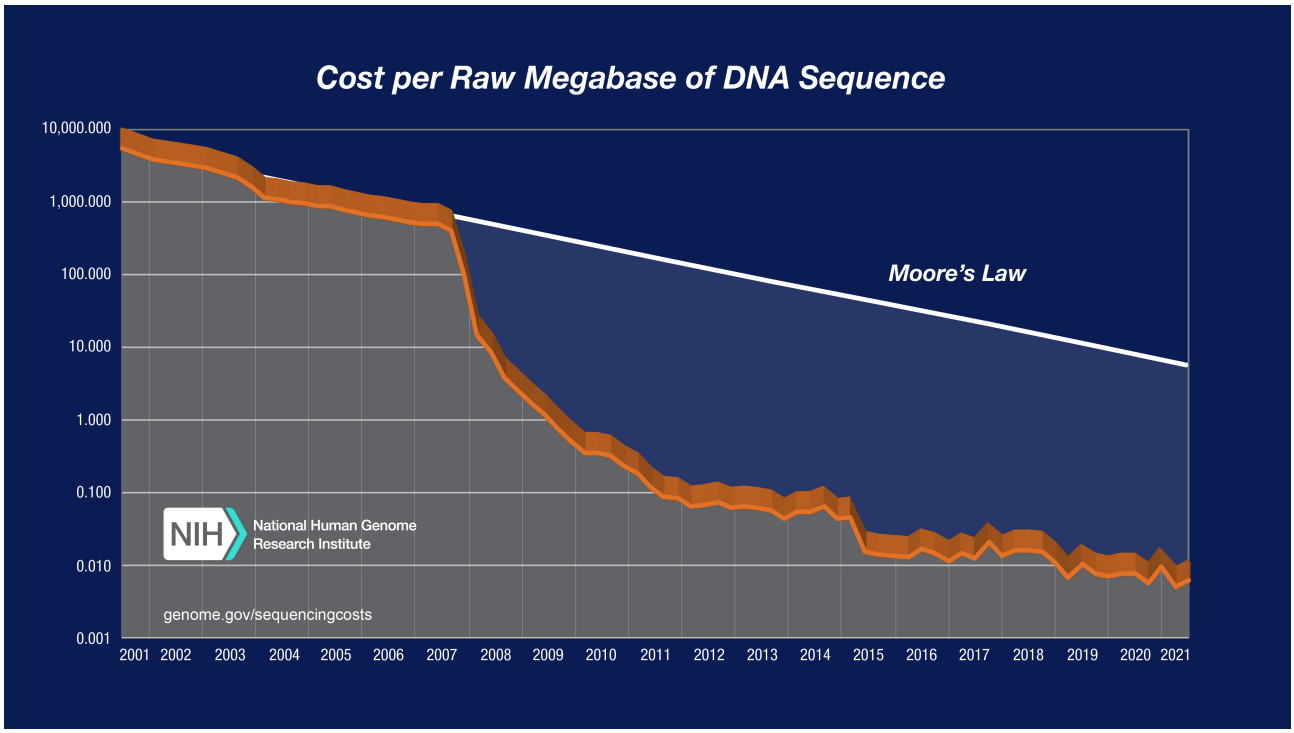

Figure 1: Comparaison entre le prix du séquencage de l'ADN et le doublement de la capacité de calcul des microprocesseur tout les 2 ans prédites par la loi de Moore (source : https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data)

2. La génomique comparative: applications en biologie.

Avant de traiter dans le chapitre suivant des apports de la génomique comparative en biologie évolutive proprement dit, il m'a paru important de donner un aperçu des applications dans d'autres domaines de la biologie car le caractère transversal de l'étude de l'Évolution du vivant implique de fréquent allers/retours dans des champs thématiques variés de la biologie.

Génomique comparative en médecine. L'obtention du premier génome humain a ouvert la voie a la médecine génomique qui a considérablement fait progresser le diagnostic et la prévention de nombreuses maladies, en particulier en cancérologie et maladies d'origine génétique (Figure 2)[5]. L'obtention a bas coût de génomes humains de bonne qualités a rapidement permis d'opérer des comparaisons fines entre les génomes, en particulier pour détecter des mutations ponctuelles mais aussi des variation structurales des génomes (duplications de gènes par exemples). Combiné a des méthodes statistiques de type GWAS (Genome Wide Association Studies) qui étudie les corrélations entre un phénotype donné et la distribution des variants génomiques, il a été possible d'identifier de nombreux gènes *potentiellement* impliqués dans de nombreuses maladies d'origine génétique ou d'identifier des facteurs génétiques de sensibilité a telle ou telle maladie. Un cas d'école a été la découverte des mutations sur les gènes BRCA1 et BRCA2 associées a un risque très élevé de cancer du sein. Il est intéressant de constater que beaucoup de méthodes et logiciel bio-informatique

initialement développé en médecine génomique ont été appliqué avec succès dans l'étude de traits chez bien d'autres espèces (notamment pour les GWAS mais aussi dans la détection des variations structurales des génomes). Si ces études se sont souvent heurtées aux caractères fondamentalement polygéniques de nombreux traits avec l'implication de dizaines voire de centaines de gènes dans un phénotype donné, elles ont considérablement amélioré l'efficacité des approches de type « gène-centrées» ou «gène candidats» , en particulier depuis l'avènement des technologies d'édition du génome de type CRISPR-Cas.

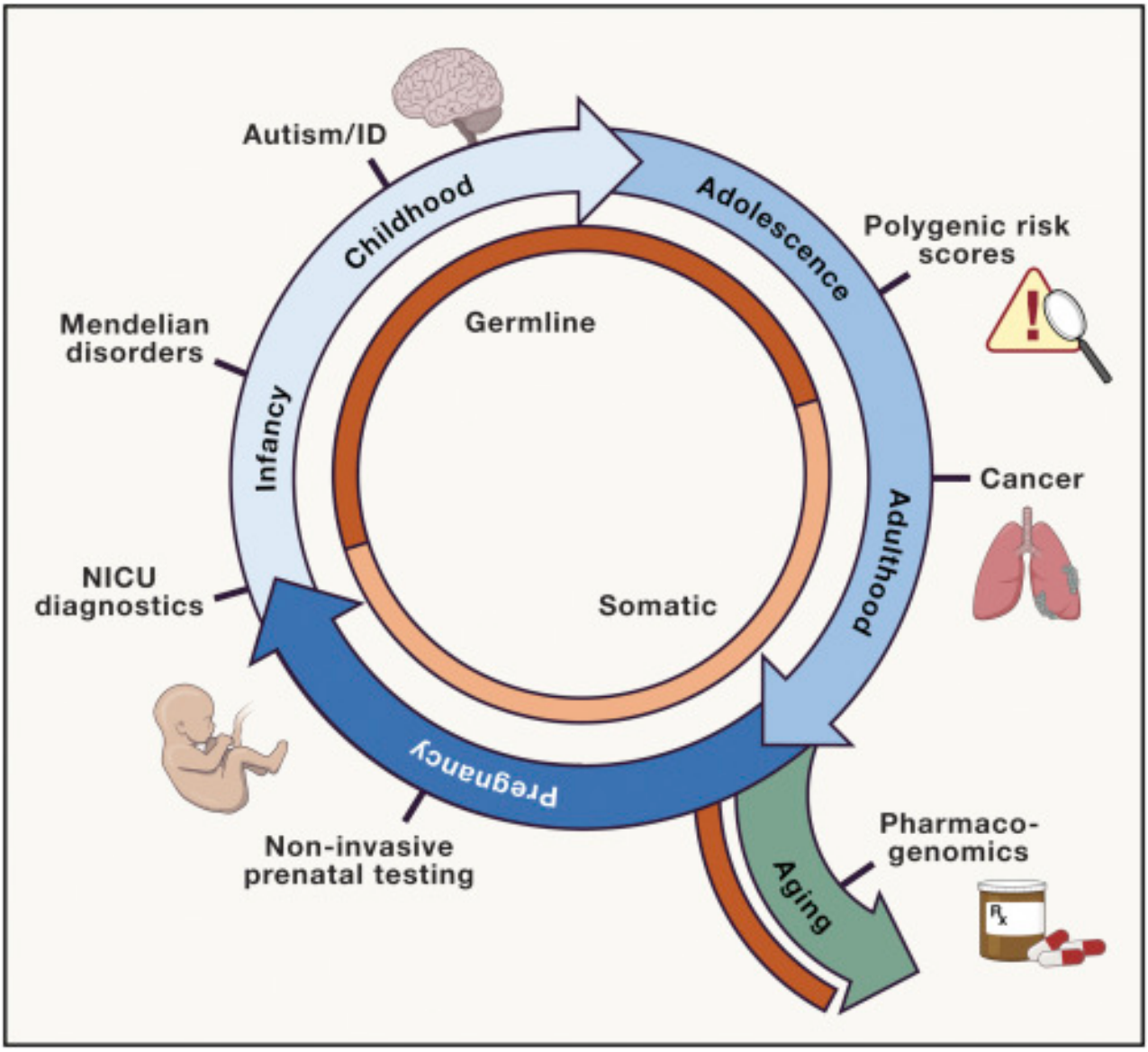

Figure 2 : Applications de la médecine génomique au cours du cycle de vie de l'être humain

Métagénomique et génomique microbienne: Notre connaissance de la diversité du monde microbien s'est pendant très longtemps heurté a la difficulté de pouvoir reproduire et cultiver de nombreux microbes. Pour autant le développement d'amorces de PCR universel ciblant l'ADNr 16S chez les procaryotes et 18S chez les eucaryotes a révélé une impressionnante diversité longtemps passée

inaperçues. La découverte des Archaea, «troisièmes branches du vivant est sans doute la découverte la plus emblématique de cette époque[6]. Néanmoins la disponibilité d'un seul gène «barcode» a considérablement limité l'analyse des communautés microbiennes et le séquençage direct de l'ADN génomique de l'environnement puis les analyses de génomique comparative de ces séquences a débouché sur de très importante découvertes sur la diversité du contenu génomique des communautés microbiennes. Ces analyses ont révélé de nombreux types d'interactions, des voies métaboliques nouvelles, des rôles écologiques insoupçonnés, des rôles dans les grands cycles geochimiques de la matière etc.[7] En conjonction avec les techniques traditionnelles de mise en culture, la génomique comparative des microbes et de leurs communauté ont fait émerger les concepts de «microbiome» (ensemble des microbes et de leurs génomes vivant dans un environnement donné) et d'«hologenome» (ensemble des génomes microbiens et de leurs hôtes multi-cellulaires constituant l'«holobionte» au sens de la microbiologiste Lynn Margulis c'est a dire donnant prise ensemble a la sélection de groupe[8]). Au cœur de ces découvertes, l'abondance des relations symbiotiques entre microbes, et entre microbes et leurs hôtes, ont ouvert un vaste champ d'étude en écologie microbienne mais aussi en santé publique sur le rôle joué par le microbiote dans l'apparition de maladie. En synthèse, la génomique comparative des communautés microbiennes a révolutionné notre perception du monde microbien, mais au-delà de sa diversité et de son rôle fonctionnel dans les écosystèmes, nous verront dans le chapitre suivant que la génomique microbienne a aussi considérablement modifié notre connaissance de l'évolution du vivant.

Génomique comparative des virus: Notre connaissance des virus a longtemps été empêchée d'une part par la méconnaissance de leurs hôtes, microbiens notamment, et d'autres part l'absence des gènes universels de type barcode facilement obtenus par PCR. L'avènement de la métagénomique a permis de révéler d'une manière indirecte combien le monde viral était divers. La fraction virale des métagénomes a cessé d'être un sous-produit de la métagénomique grâce a des design expérimentaux spécifiques, notamment par ultra-filtration, qui ont permis de concentrer l'ADN viral et de le séquencer préférentiellement. Cette métagénomique virale a révélé une quantité et une diversité complètement insoupçonnée de virus et de phages dont les catalogues de gènes dépassaient en diversité tout ce que l'on pouvait appréhender. Un exemple emblématique de cette thématique repose sur les découvertes de virus géants appartenant aux groupes des NCLDV (Nucleo Cytoplasmic Larges DNA Virus). Exemple typique de sérendipité, la découverte du Mimivirus et de son génome viral géant d'1,2Mb alors que ses découvreurs croyaient travailler sur une bactérie intracellulaire (le «Bradford coccus») a complètement relancé le débat sur l'origine des virus[9]. En moins de 20 ans, la métagénomique virale des NCLDV a considérablement étendu notre connaissance de ce groupe de virus qui sont révélés être partout, en grande quantité et capable d'infecter un grand nombre d'hôtes différents, des protistes aux animaux en passant par des algues. A l'heure actuelle, plus de 2000 génomes différents de NCLDV sont connus dont plus de 90% sont issus d'analyses métagénomiques[10] (Figure 3)!

Génomique comparative en agronomie: L'amélioration des plantes cultivées et des animaux de rente est un processus datant de l'Antiquité qui a aboutis a une extraordinaire diversité de races, lignées et croisements. Les avancées de la génomique ont permis de comprendre l'histoire et les mécanismes de ces domestications qui récapitule les mécanismes généraux de l'évolution du vivant mais accéléré par la main de l'homme (ils seront donc discutés dans le paragraphe suivant)[11]. En plus de permettre de mieux connaître la diversité génétique et génomique intra-spécifique des espèces cultivées, les analyses comparatives génotypes/phénotypes de type GWAS ou les recherches de QTL («quantitative trait loci») ont permis de comprendre les déterminants génétiques et génomiques associés a de nombreux traits particuliers d'intérêt agronomique comme des gènes de résistances a des ravageurs ou des gènes d'adaptation a des environnements particuliers par exemple. Ces données ont permis

d’améliorer et d’accélérer les processus de sélection variétale avec le développement de stratégie de sélection assistée par marqueur. En outre le développement des approches d’édition du génome par la technologie CRISPR-Cas9 a ouvert un boulevard dans la capacité des sélectionneurs a introduire dans une lignée un haplotype conférant des caractéristiques particulières et ce, en s’affranchissant des risques associés aux techniques de modifications génétiques classique[12].

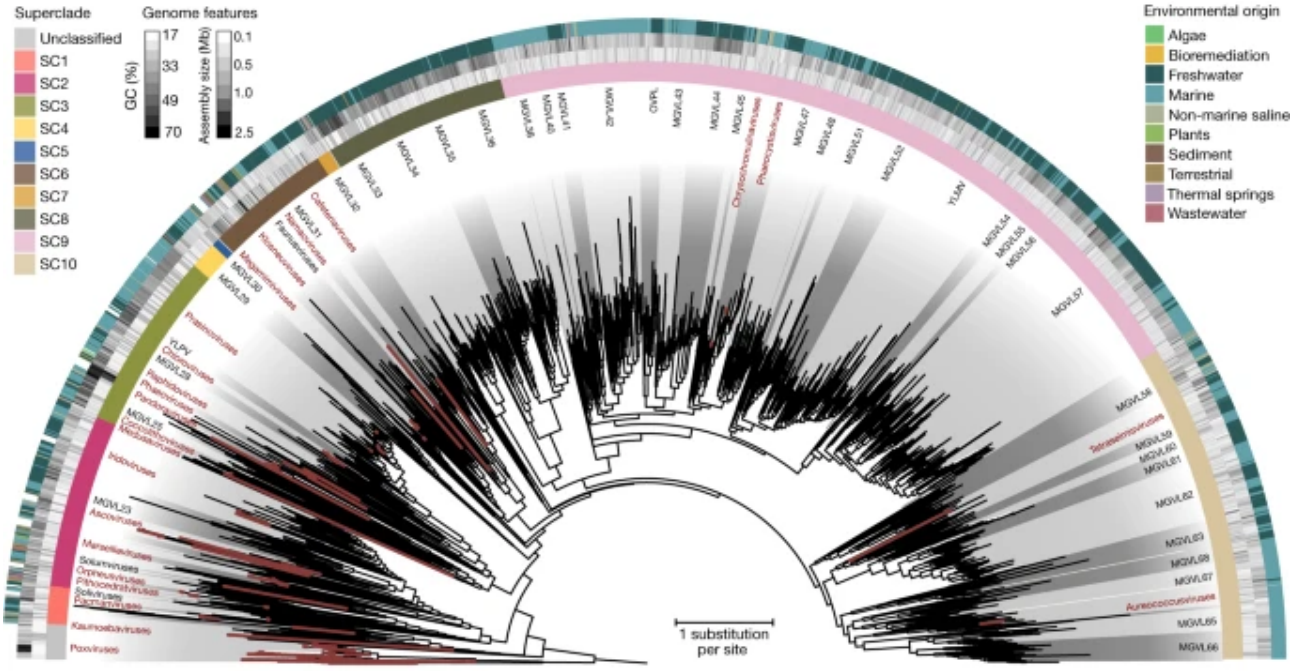

Figure 3: Phylogénie des NCLDV basé sur 5 gènes conservés. Les branches en rouge foncé correspondent aux génomes complets issus de culture, en noir ceux issu de métagénome. Les cercles extérieures correspondent de l’intérieur vers l’extérieur aux groupes phylogénétiques d’appartenances, aux caractéristiques génomiques et a l’environnement source.

3. Génomique comparative et biologie évolutive.

*LUCA et l’Eucaryogénese***:** La dénomination LUCA pour Last Universal Common Ancestor doit beaucoup aux travaux de Patrick Forterre qui a su populariser l’idée de l’unicité du vivant déjà postulée chez Darwin dans l’Origine des Espèces («*All the organic beings which have ever lived on this Earth may be descended from some one primordial form*»)[13]. La monophylie du vivant avait été confirmée par les travaux pionniers en génétique sur l’existence de gènes universels et homologues comme les gènes codant pour les ARN ribosomaux au seins des trois branches du vivant (bactérie, archée, eucaryotes). Les apports de la génomique comparative ont été ici considérable dans le «portrait robot» que l’on peut se faire de LUCA, notamment sur le fait que son répertoire génomique était déjà substantiel (plusieurs centaines de gènes) et comprenait notamment une machinerie de traduction et de transcription dont la complexité était déjà comparable a celle du vivant contemporain[14]. Néanmoins, plusieurs composants essentiels du vivant ne sont pas universellement conservés: l’exemple le plus emblématique est l’appareil de réplication de l’ADN et certains composant clé du métabolisme de

l'ADN qui ne sont pas homologues entre d'une part les bactéries et d'autres part les Archées et les Eucaryotes. Retenons que la ressemblance des gènes de la réplication entre bactéries et eucaryotes peut être interprété comme une synapomorphie permettant d'estimer qu'Archées et Eucaryotes partagent un ancêtre commun divergent de l'ancêtre des bactéries : c'est l'hypothèse de l'origine symbiotique des eucaryotes[15]. Dans ce cadre, l'eucaryogénèse résulterait d'une symbiose métabolique entre une archée et une ou plusieurs bactéries, dont au moins une alpha-protéobactérie. Si par nature cette hypothèse demeure spéculative, la découverte des archées appartenant a la lignée Asgard a considérablement conforté cette vision. En effet dans la vaste diversité des archées, les Asgard ont un positionnement basal aux eucaryotes dans les phylogénies de gènes universellement conservés[16]. Cet arbre a deux domaines (2D) contraste avec les phylogénies classiques a trois domaines (3D) (Figure 4) mais semble supporté d'une manière robuste dans les analyses phylogénomiques[17], même si le débat n'est probablement pas encore complètement clos[18].

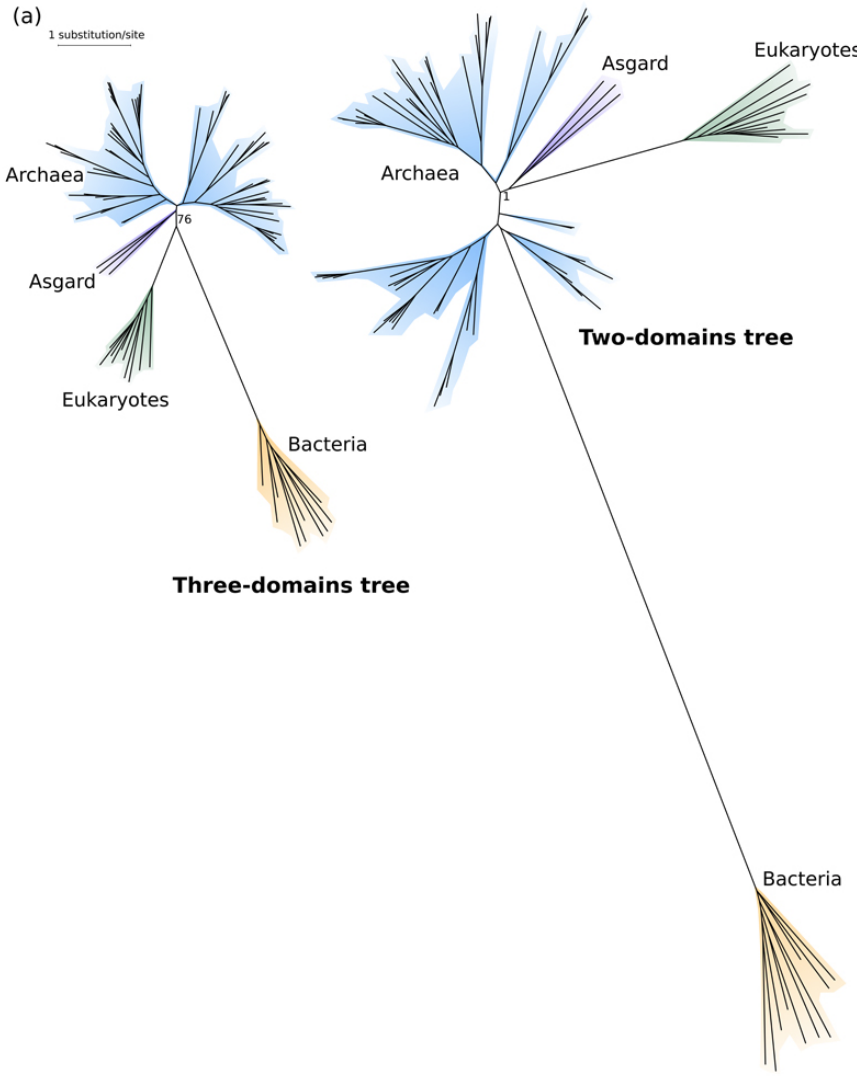

Figure 4: Phylogénie du vivant supportant un arbre 3D et un arbre 2D en utilisant des données simulées

Dans l'hypothèse 2D qui semble avoir aujourd'hui un support croissant, l'eucaryogénèse serait concomitante a l'endosymbiose d'une alpha-protéobactérie ancestrale qui deviendra une mitochondrie. Cette observation pointe du doigt l'importance fondamentale des processus symbiotiques dans l'Évolution du vivant qui seront discuté dans le paragraphe suivant.

Evolution des symbioses microbiennes et des endosymbioses: Rare sont les preuves solides dans la reconstruction de l'histoire ancienne du vivant, pourtant l'origine bactérienne des mitochondries et des chloroplastes ne fait désormais plus l'objet d'aucun débat. En effet les phylogénies des gènes d'organelles ont systématiquement pointé les alpha-protéobactéries et les cyanobactéries comme ancêtres des mitochondries[19] et des chloroplastes[20] respectivement. L'origine endosymbiotique des organelles est la partie émergée d'un iceberg de symbioses entre microbes et entre microbes et organismes multi-cellulaires qui prend la forme d'un continuum évolutif entre le parasitisme *sensu-stricto* et différents gradients de mutualismes où chaque partenaire dans le consortium trouve son avantage[21]. La génomique comparative a permis de déterminer que les symbioses microbiennes concernaient virtuellement toutes les branches du vivant et étaient la source de nombreux bénéfices: protection contre des pathogènes ou des parasites, augmentation de la fécondité, approvisionnement en nutriments (azote, vitamines...), bioluminescence, photosynthèse etc.… Il est aussi remarquable d'observer la diversité taxonomique des microbes impliqués dans ces symbioses où presque tout les grands groupes de bactéries et de nombreux champignons sont impliqués[21]. Si l'on prends l'exemple des insectes hématophages obligatoires, la quasi-totalité d'entre eux dépendent d'une manière critique de divers symbiotes pour supplémenter en vitamine B le sang de leurs proies qui en est carencé[22] (Figure 5).

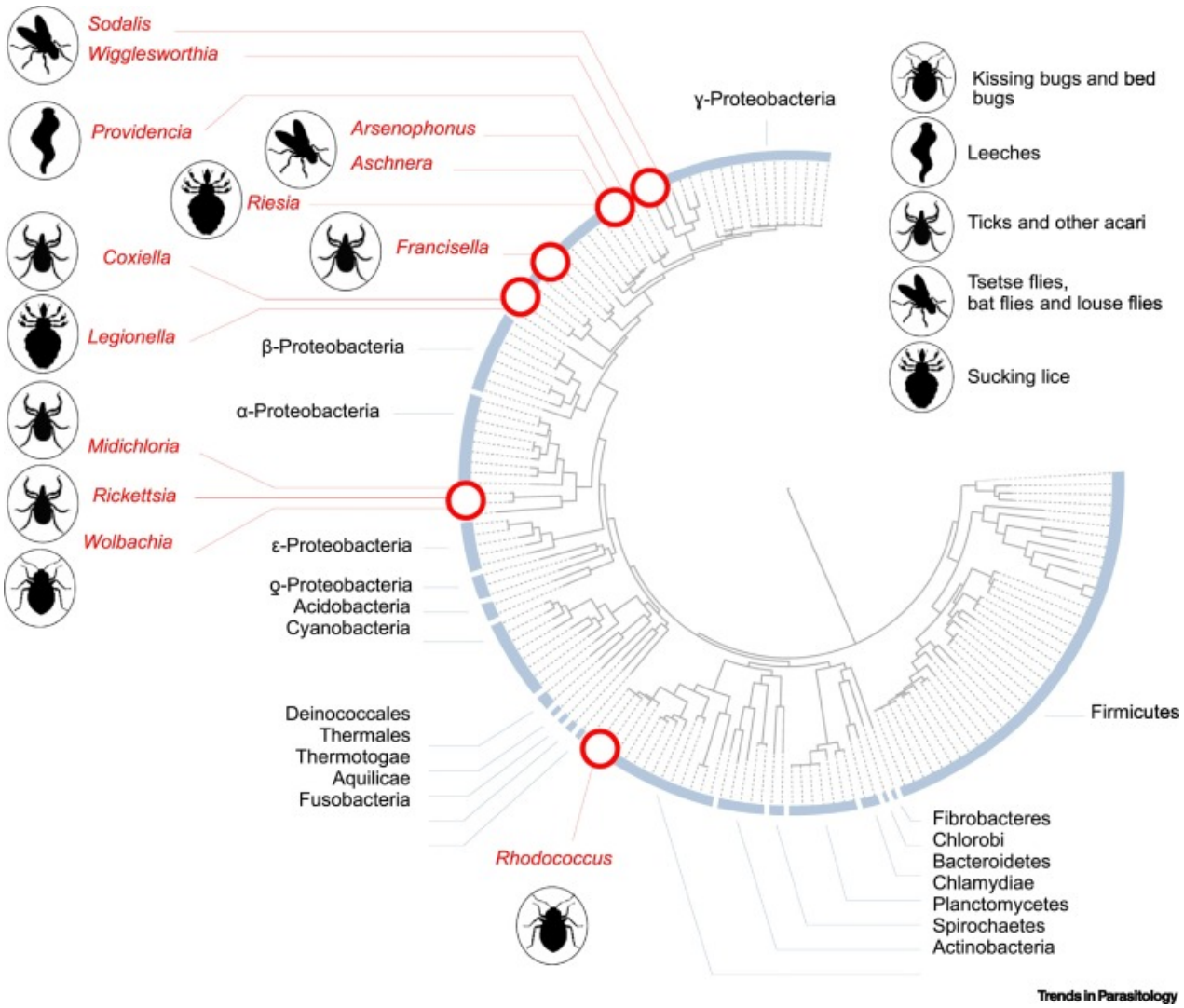

Figure 5: Relations évolutives entre les organismes hématophages obligatoires et leurs symbiotes nutritionnels.

Ce processus est appelé «symbiose nutritionnelle»[23] car l'élimination du symbiote en laboratoire peut être compensé par l'ajout, dans le repas sanguins des insectes, de vitamines B et plus précisément de biotine. Les génomes des symbiotes codent en effet tous d'un opéron de biosynthèse de cette vitamine qui est généralement le composant-clé de la symbiose nutritionnelle. D'autres types de symbiose nutritionnelle existent, on pourra citer par exemple chez les plantes les symbiotes du genre *Rhizobium* responsables de la fixation de l'azote atmosphérique et de approvisionnement de leurs hôtes en azote assimilable par la plante.

Le rôle évolutifs des éléments génétiques mobiles: La génomique comparative a révélé une extraordinaire diversité d'éléments génétiques mobiles (EGMS) présents dans la quasi totalité des génomes. On pourra y inclure les transposons, les plasmides, les prophages intégrés, les intégrons, divers systèmes de défense bactérien comme les systèmes toxine-antitoxine ou restriction-modification etc... On pourra définir les EGMs comme des gènes égoïstes qui se réplique au détriment de leurs génomes hôtes. La prolifération des EGMs est en grande partie responsable de ce que Michael Lynch appelle le «syndrome de la complexité» c'est a dire une forme d'obésité des génomes qui n'arrivent plus a éliminer ou a contre-carrer l'expansion de ces gènes égoïstes[24]. Cette hypothèse est supportée par de nombreuses preuves empiriques apportées par la génomique comparative, documentant la prolifération dans presque tout les génomes des transposons. Si l'on prends par exemple le cas des plus gros génomes animaux, ceux des salamandres (10-120 Gb), les transposons occupent jusqu'à plus de 50% du génome[25]. Si les EGMSs sont par nature des réplicateurs égoïstes, il peut être sélectivement avantageux pour ceux-ci de mitiger leurs effets potentiellement délétères par des avantages évolutifs conférés a leurs hôtes. Il existe de nombreux cas de bénéfices de type mutualistes documentés par les EGMs : en reprenant l'exemple donné précédemment par les bactéries légumineuses du genre *Rhizobium*, les gènes de fixation de l'azote atmosphériques *nod* et *nif* sont très souvent portés par un ou plusieurs plasmides[26]. En conférant un avantage sélectif a son hôte, le plasmide augmente par la même occasion sa propre diffusion. De nombreux autres cas de mutualisme sont portés par des plasmides et des pro-phages on citera des gènes de toxines, des gènes de résistances a d'autres EGMs co-infectants, des gènes de résistances a des antibiotiques etc. Chez les eucaryotes, les EGMs ont fréquemment été la source de mutation avantageuse comme par exemple chez la phalène du bouleau où l'adaptation a des environnements pollués par l'activité industrielle avec un phénotype noir cendré («*carbonaria*»)(Figure 5) est lié a l'insertion d'un élément transposable qui a modifié la cascade de régulation des gènes impliqués dans la couleur du papillon[27].

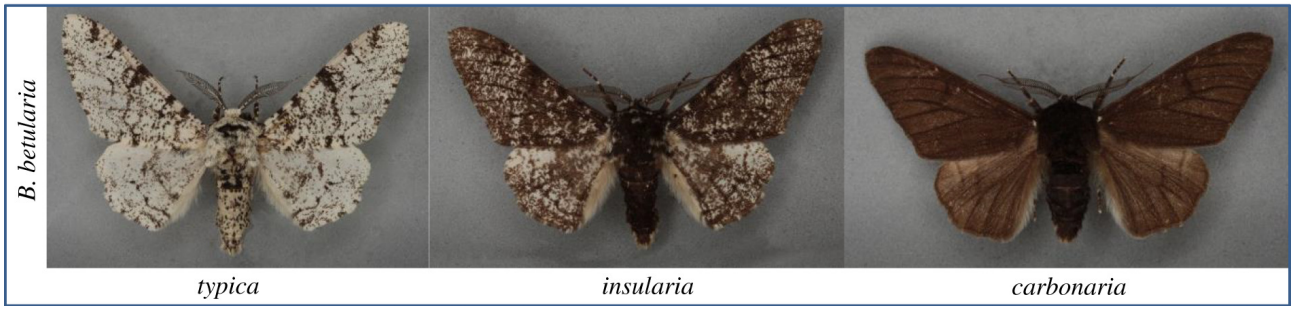

Figure 5: Mélanisme industriel chez la phalène du bouleau causé par l'insertion d'un transposon.

Finalement c'est parfois même les séquences elles mêmes des gènes de transposons qui ont été utilisées par leurs hôtes pour accomplir des fonctions cellulaires, l'exemple le plus connu de ce type d'exaptation moléculaire est celui du gène *syncytin*, codant pour une protéine-clé de la genèse du placenta chez les mammifères, qui est dérivée d'un gène d'enveloppe *env* de rétro-transposons[28].

Au sens de Lynch[24], cette prolifération pourrait être due a des phénomènes démographiques comme

des goulots d'étranglements qui par effet de dérive génétique empêcherait les organismes d'éliminer ces EGMs. C'est ce qui expliquerait pourquoi les procaryotes, qui ont généralement des tailles de populations élevées, ont pu conserver des génomes de petites tailles, comparativement peu colonisé par les EGMs par rapport aux eucaryotes. Néanmoins, chez les plantes et de nombreux métazoaires, on trouve une mécanisme de défense contre les transposons qui repose dans les lignées germinales par la complémentarité de séquence entre un vaste répertoire de piARN (pour « PIWI-interacting ARN») et des courtes séquences issu de transposons déclenchant une dégradation spécifique et mémorisée de cette cible. Ce système immunitaire anti-transposon repose sur l'intégration dans le génome de petites séquences issues des transposons regroupées dans des clusters qui produisent des piARNs. Ces piARNs s'apparient, par ressemblance, aux transcrits issus des transposons, initiant leurs dégradations spécifiques par le système PIWI et générant une boucle de rétroaction de type ping-pong qui potentialise et augmente la réponse[29] (Figure 6). Si il existe des différences dans le mode d'action biochimique des systèmes immunitaires anti-EGMs entre les plantes et les animaux, la machinerie des piARN est capable d'éteindre relativement rapidement l'activité des transposons, la seule issue possible pour ceux-ci étant leurs capacité a se transmettre horizontalement vers des hôtes encore «naïf» illustrant l'importance dans le vivant de l'hérédité génétique horizontale.

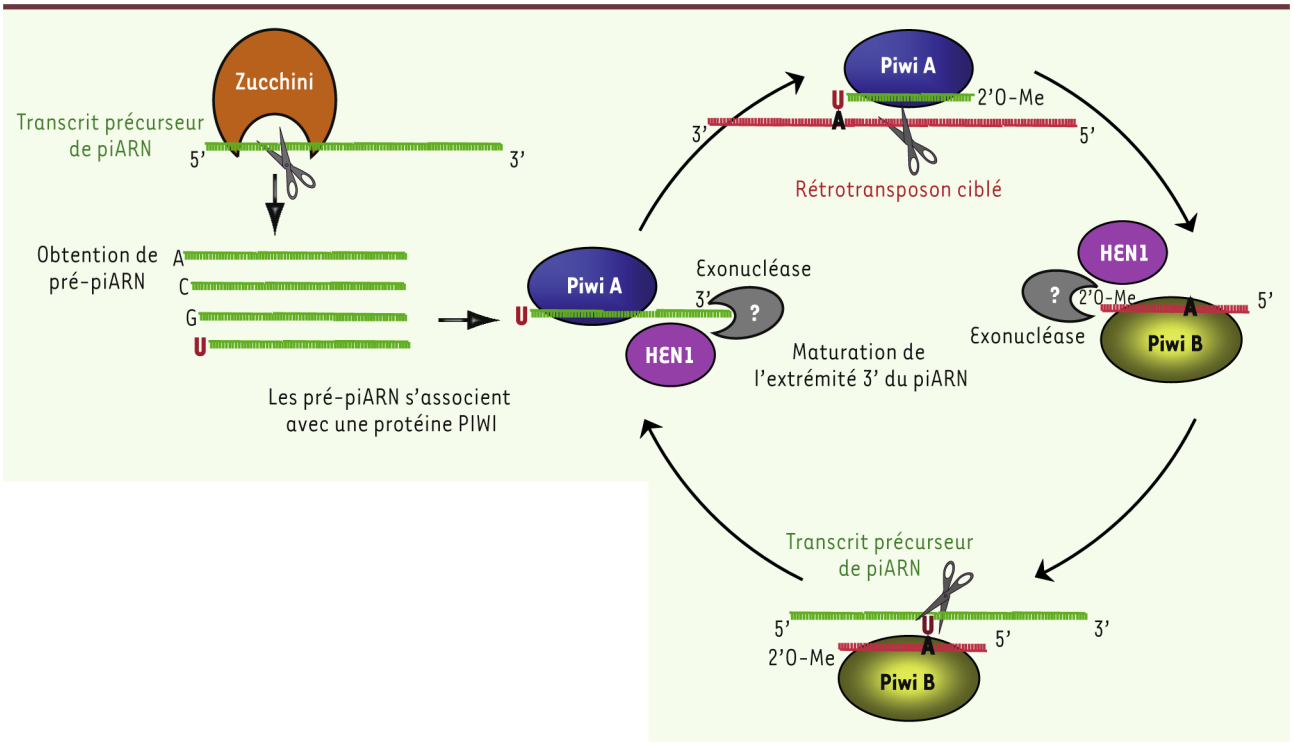

Figure 6: Fonctionnement schématique de la voie des piARN[30]

L'importance des flux horizontaux de gènes au cours de l'Évolution : L'importance des échanges de horizontaux gènes est un des acquis majeurs de la génomique comparative, et ce, même si l'estimation exacte a l'échelle d'un génome donné de la part des gènes issues de transferts horizontaux est très dépendante des approches utilisées[31]. On peut donner comme exemple la situation de la quasi-totalité des lignées d'archées dont l'origine corrèle a des transfert horizontaux inter-domaine en provenance des bactéries impliquant surtout des gènes métaboliques[32]. Cette importance quantitative a amené chez les procaryotes a la notion de «pan-génome», c'est a dire a une dichotomie génomique entre une petite fraction de gène «core» (coeur), conservé a une échelle taxonomique donnée, et un vaste répertoire de gènes accessoires pas ou peu conservés. Ainsi au seins de plus de 1300 génomes d' *E. coli* , il a été possible d'identifier plus de 25 000 familles de gènes accessoires (le pan-génome) pour un génome

cœur de moins de 3000 familles de gènes (et moins de 1000 familles strictement conservées)[33]. A plus grande échelle, les gènes conservés au seins des trois domaines du vivant sont proches de 100, principalement des gènes de la traductions et des ARN polymérases, qui ne sont pas elles-même indemnes de transferts horizontaux[34]. Si il y a débats sur l'importance des transferts horizontaux chez les eucaryotes, en particulier les multicellulaires présentant une différentiation tissulaire de type *soma*/ *germen*[35] , l'abondance des transferts chez les procaryotes a complètement remis en question le concept d'arbre du vivant, c'est a dire de phylogénies d'espèces, pour le remplacer par des structures en réseaux, mieux a même de représenter ces flux horizontaux[36]. Quand aux rares gènes conservés, leurs phylogénies représenteraient ainsi l'arbre du 1% (« the tree of one percent »)[37]et probablement d'ailleurs bien moins qu'1% des gènes d'un génome donné ! On notera que quelques rares transferts horizontaux de gènes ont été validé chez les métazoaires, conférant souvent des avantages adaptatifs très fort, ce qui indique que le processus est universel: c'est par exemple le cas du gène BtPMaT1 codant pour une glucoside malonyltransferase chez l'aleurode *Bemisia tabacii*, en provenance de bactéries, qui permet ainsi a cet insecte de détoxifier les phénols produits par les plantes parasitées[38] (Figure 7).

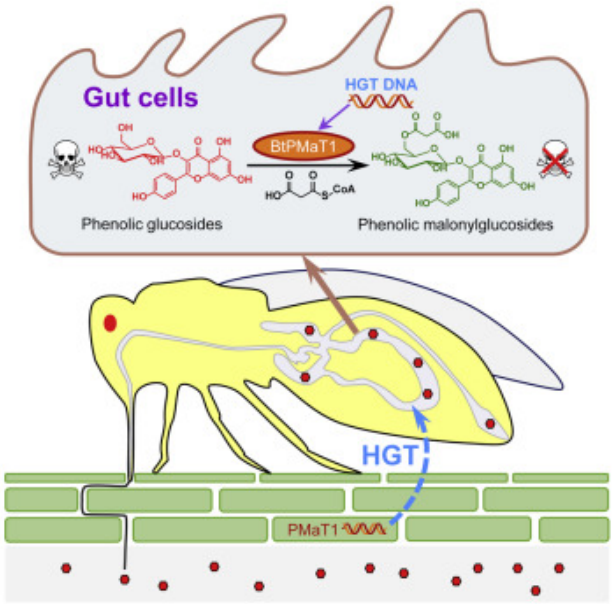

Figure 7: Adaptation a la plante-hôte chez *B. tabacci* grâce a un transfert horizontal de gène bactérien.

Si la plupart des cas de transferts horizontaux chez les eucaryotes multi-cellulaires impliquent des transferts en provenance d'endosymbiotes (incluant ceux ayant aboutis aux organelles), le phénomène semble beaucoup plus massif chez les protistes où l'existence d'un pan-génome fait même débat[39].

Au terme de cette introduction sommaire générale, il apparaît clairement l'importance de ce que l'on pourrait appeler «la génomique des interactions » en biologie évolutive c'est a dire l'étude des mécanismes et des patrons génétiques et génomiques qui explique les relations entre les organismes constituant l'holobionte (incluant les parasite ou les symbiotes) mais aussi entre ces organismes et leurs «mobilomes» (virus et autres EGMs) au sein de ce que l'on peut appeler «l'hologénome» . C'est en grande partie ceci qui motive mes recherches et qui a fait l'objet d'études sur des holobiontes très divers et qui seront détaillés dans les chapitres suivants.

Résultats & Discutions

1. Evolution de l'appareil de réplication et du métabolisme de l'ADN

Au cours de ma thèse dans le laboratoire de Patrick Forterre a l'Institut de Génétique et de Microbiologie (IGM) d'Orsay, nous avions clairement démontré la dualité des gènes codant pour les enzymes de la réplication et du métabolisme de l'ADN: chaque fonction était en effet prise en charge par au moins deux gènes non-homologues. Une partie de mes travaux consistent donc a documenter cette observation et a comprendre les scénario évolutifs sous-jacents. Nos travaux menés notamment sur les ADN polymerases et les ADN topoisomerases ainsi que les enzymes du métabolismes de l'ADN avaient initialement permis de préciser la distribution et l'évolution de ces différentes familles, j'ai choisis ici de détailler deux questions qui illustrent mes travaux sur le sujet.

a. Le rôle des virus dans l'évolution de l'appareil de réplication des organelles.

Nous avions proposé l'hypothèse que cette étonnante redondance fonctionnelle des enzymes de la réplication résultait du fait que les virus avaient transféré latéralement une partie de leurs gènes impliqués dans la réplication de l'ADN vers leurs hôtes et avaient joué ainsi un rôle de «donneur» de nouveau gènes qui avaient remplacé les gènes cellulaires. Si cette hypothèse est encore aujourd'hui très débattues, nos travaux sur l'évolution de l'appareil de réplication des mitochondries et des chloroplastes ont apporté des preuves solides sur le rôle important joué dans ce cas de figure bien précis par les virus.

En effet, nous avions vu dans l'introduction générale de ce manuscrit que les mitochondries et les chloroplastes dérivaient chacun d'un événement unique d'endosymbiose d'une bactérie par un organisme eucaryote. Si une grande partie de ces gènes bactériens ont été perdus ou ont été transféré aux noyaux eucaryotes, un chromosome vestigial perdure au sein de ces organelles sous forme d'un génome circulaire de quelques dizaines de kilobases, rarement plus de quelques centaines de kilobases. Il y a donc nécessité de transcrire et de répliquer ces génomes lors des événements mitotiques notamment. Pourtant, se n'est pas l'ADN polymerase de la famille C (replicase bactérienne) qui est impliquée dans la réplication de l'ADN mitochondrial mais bien une ADN polymerase de la famille A, non-homologue de la famille C. Nos phylogénies ont montré que ces ADN polymérases du groupe A étaient phylogénétiquement apparentées a des ADN polymérases des bactériophages du groupe T3/T7 (Figure 8). Symétriquement, l'hélicase replicative/primase (DnaB) des mitochondries et des chloroplastes est non pas apparentée aux gènes des bactéries mais bien a des hélicases de phages T3/T7. Finalement, l'ARN polymerase DNA dépendante des mitochondries qui sert a la synthèse des amorces n'est elle-même pas homologue a l'ARN polymérase bactérienne codée par plusieurs sous-unités mais bien a une ARN polymerase monomerique qui n'existe que chez les phages T3/T7. Il existe toutefois une exception chez un seul groupe de protistes, les Jakobides, qui ont conservé l'ARN polymérase multimérique de type bactérienne[40], ce qui renforce l'idée que l'appareil de réplication de la quasi-totalité des mitochondries n'est pas issu de la bactérie initiale. Les chloroplastes de plantes utilisent quand a eux une ARN polymérase multimérique de type bactérienne. L'ensemble de ces données démontre d'une manière solide que tout ou partie des gènes codant pour l'appareil de réplication des chloroplastes et des mitochondries, ancestralement d'origine bactérienne, a été remplacé par des gènes de phages T3/T7. Il restait a proposer une mécanisme possible pour ce remplacement. Nous avons proposé l'idée que l'alpha-protéobacterie ancestrale a l'origine des mitochondries possédait dans son génome un ou plusieurs prophages intégrés de type T3/T7, a l'image de protéobactéries contemporaines (Figure 9).

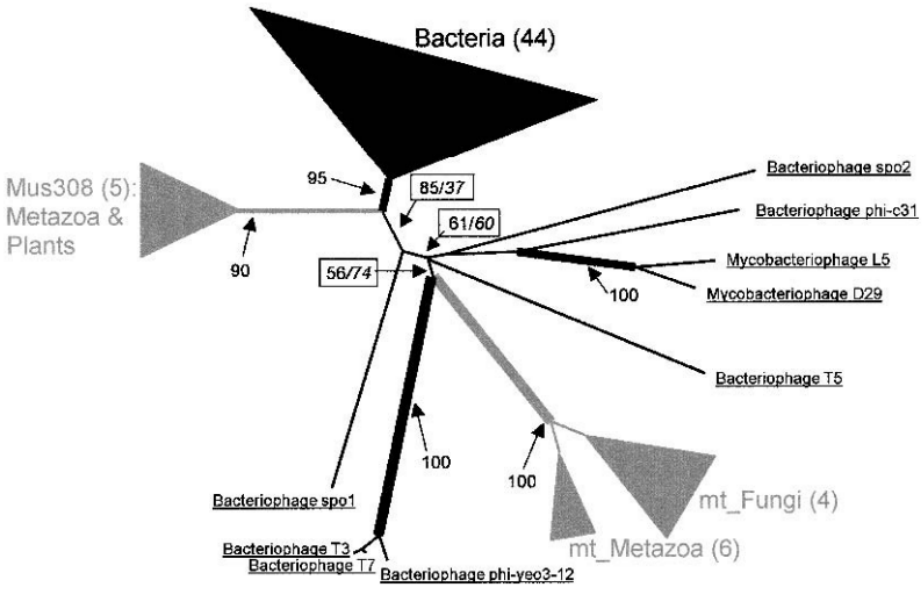

Figure 8 : Phylogénie de maximum de vraisemblance de l'ADN polymerase A

Se sont ces gènes de prophages qui ont été transférés dans le noyaux eucaryotes lors de l'endosymbiose et qui ont pu prendre en charge la réplication des mini-chromosomes mitochondriaux, a la seule exception du groupe des protistes Jakobides qui ont conservé les enzymes bactériennes originelles peut-être en raison de leurs génomes mitochondriaux de taille inhabituellement importante (jusqu'à plus d'une centaine de kilobases). Chez les chloroplastes, un mélange des enzymes bactériennes (ARN polymérase) et d'enzymes originaires des phages T3/T7 (ADN polymérase et DNA Primase/Hélicase) seraient responsables de la réplication de l'ADN chloroplastique (mais aucun de ces gènes ne serait issu de la cyanobactérie ancestrale !). Ces données constituent a ce jour la preuve la plus solide que des gènes viraux ont pu remplacer des gènes cellulaires lors de l'évolution de l'appareil de réplication de l'ADN. Certes, il s'agit d'un cas de figure un peu atypique car l'endosymbiose a engendré une réduction drastique de la taille des génomes des organelles rendant possible l'établissement d'un système minimaliste de réplication de type viral, néanmoins il indique que ce type d'événement est possible et potentiellement qu'il a pu impliquer d'autres systèmes de réplication.

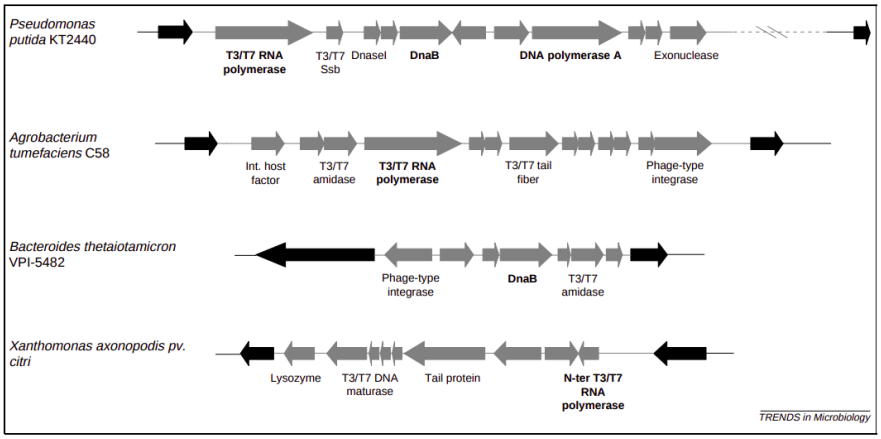

Figure 9 :Organisation des prophages de type T3/T7 (en gris) trouvé dans les génomes de bactérie (en noir).

b. Evolution des Thymidylate Synthases (et du métabolisme du folate).

La thymidylate synthase est une enzyme-clé du métabolisme de l'ADN qui convertis le 2′-deoxyuridine-5′-monophosphate (dUMP) en 2′-deoxythymidine-5′-monophosphate (dTMP), un précurseur essentiel de l'ADN. Pourtant, de nombreux génomes procaryotes ne codent pas pour l'enzyme canonique ThyA alors que cette enzyme était indispensable pour fabriquer du dTMP. Avec Hannu Myllykallio (CNRS/Ecole Polytechnique), en étudiant la distribution des gènes dans les génomes codant, et ne codant pas pour un gène *thyA* nous avions réussis a identifier un gène qui avait une répartition génomique mutuellement exclusive avec celui-ci. Ce gène nommé *thyX*, dont la fonction était inconnue, était homologue avec le gène *thy1* chez le protiste *Dictyostelium* dont le phénotype des mutants négatifs *thy1*- étaient restaurés par l'ajout de thymine dans le milieu de culture[41]. Ces prédictions bioinformatiques nous ont conduits a penser qu'il s'agissait bien là de la thymidydale synthase alternative a l'enzyme canonique ThyA. Ces suppositions ont ensuite testée fonctionnellement par le groupe d'Hannu Myllykallio qui a démontré d'un point de vue biochimique et par test de complémentation génétique que ce gène *thyX* codaient bien pour une deuxième thymidylate synthase, non homologue a ThyA, et qui opérait avec une biochimie complètement distincte. En effet alors que ThyA opère avec un co-enzyme de type méthylène-tetrahydro-Folate (C2H4-folate), ThyX a recours a un co-facteur de type flavine-adénine dinucléotide (FAD). Les phylogénies de ThyA et ThyX montraient de nombreux événements de remplacements de l'un par l'autre (et vice et versa) rendant impossible l'établissement d'un scénario évolutif crédible, notamment sur l'ancestralité de ces enzymes au sein des trois domaines du vivants.

Plus de 15 ans plus tard, la découverte du groupe des archées Asgard, qui pourrait être a l'origine des eucaryotes dans le scénario de l'eucaryogénèse par symbiose archée/bactérie, nous as incité a regarder de plus prêt l'évolution de ces enzymes chez ce groupe d'archée. Nous avons tout d'abord démontré que le gène *thyX* avait une répartition ponctuelle et erratique au sein des génomes d'Asgard, a l'image de presque tout les autres gènes impliqués dans le métabolisme du thymidylate/folate (Figure 10). Par opposition, *thyA* était beaucoup plus largement distribué au sein des Asgards (Figure 10).

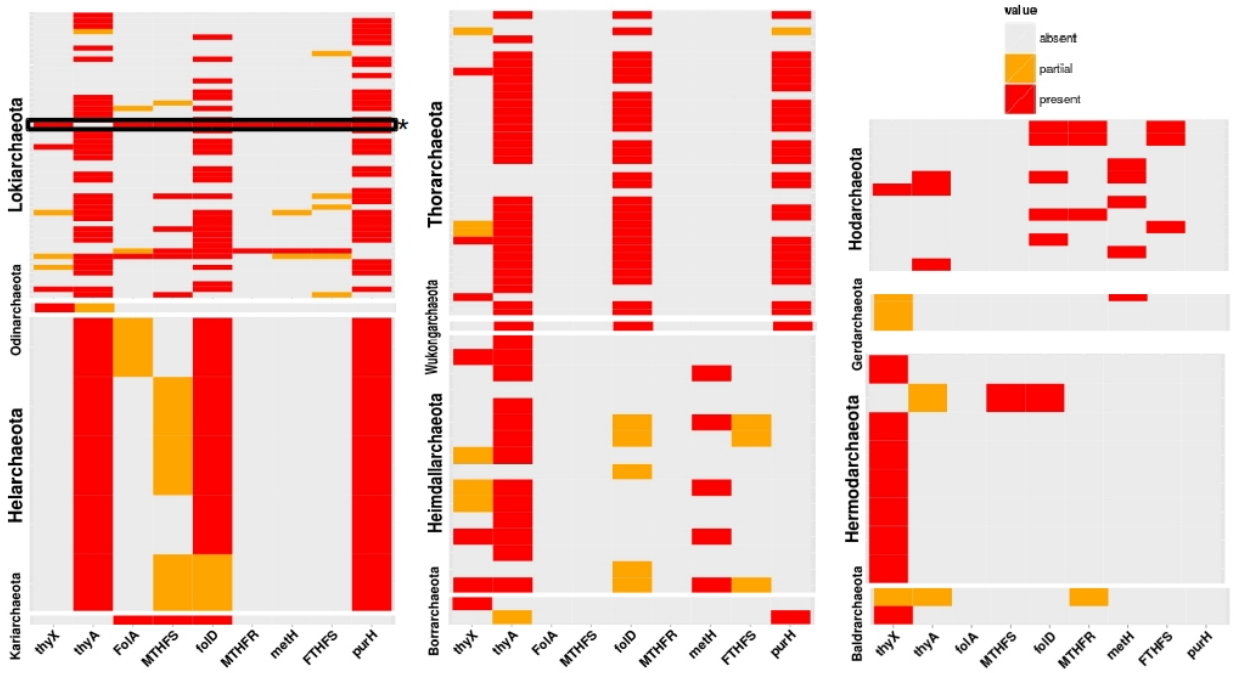

Figure 10: Distribution phylétique des gènes du métabolisme du Thymidylate/Folate au sein des (méta)génomes des différents familles d'Archées Asgard, la seule Asgard cultivée en laboratoire a été encadrée.

Les phylogénies montrent que les Thymidylate synthase de type Folate-dépendante (ThyA) d'Asgard forme un groupe monophylétique avec les autres séquences d'archées, alors que pour la version FAD-dépendante (ThyX), les Asgards sont polyphylétiques et nichées au sein des séquences bactériennes (Figure 11). On notera avec intérêt que les séquences eucaryotes du gène thyA sont phylogénétiquement très distantes des séquences d'Asgard et positionnées au seins du sous-arbre des bactéries.

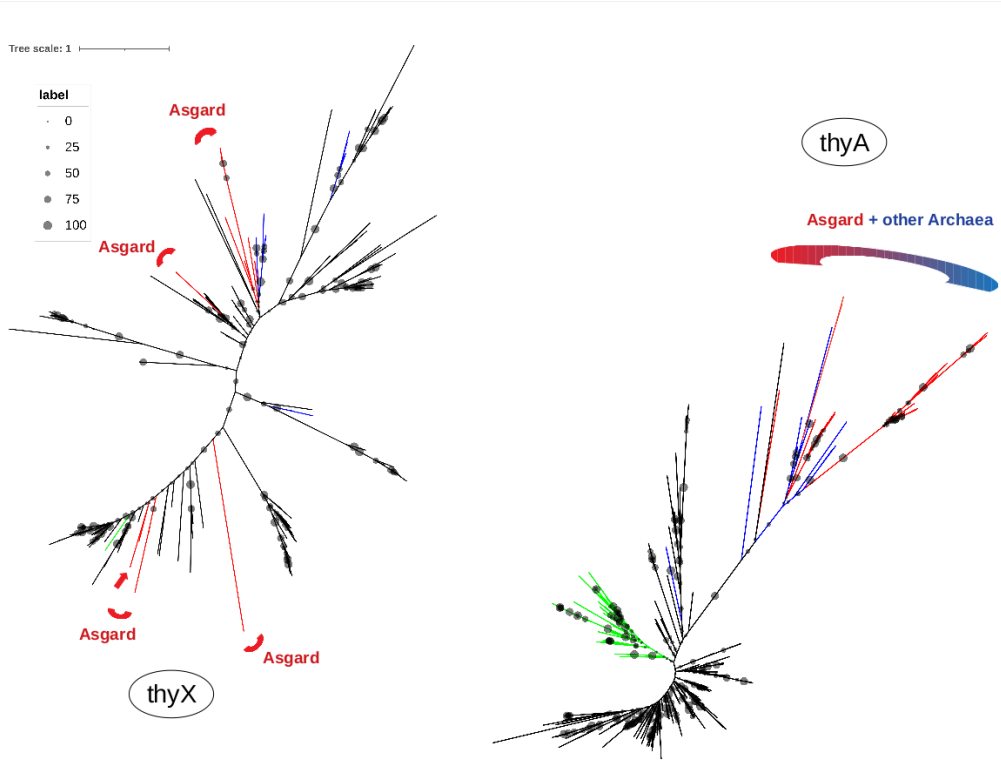

Figure 11: Phylogénie de Maximum de Vraisemblance des gènes *thyX* et *thyA*. Les Asgard sont en bleu, les autres archées sont en rouge et les eucaryotes en vert.

Les analyses de réconciliations phylogénétiques a partir d'une phylogénie d'espèce construite a partir du gène codant pour l'ADNr 16S prédisent plusieurs transferts horizontaux du gène *thyX* entre Asgard et bactérie, aucun pour *thyA*. Finalement l'analyse de la synthénie des contigs codant pour un gène *thy* montre une relativement bonne conservation pour les gènes *thyA* chez les Asgard, alors qu'on ne retrouve quasiment aucune synténie au sein des contigs comprenant gène *thyX* (Figure 12).

L'ensemble de ces données converge pour penser que ThyA est la version ancestrale de la Thymidylate synthase chez les Asgard alors que le gène *thyX* est très probablement issu de transferts horizontaux de gènes en provenance de diverses bactéries. L'existence de multiple transferts horizontaux inter-domaines est par ailleurs soutenue par les distributions en patch et les phylogénies de tout les autres gènes impliqués dans la voie métabolique du folate. Finalement, la validation fonctionnelle des prédictions *in silico* sur la fonction des gènes par des tests biochimiques et des expériences de complémentations fonctionnelles de mutant *E. coli thy-* ont été effectuées par l'équipe d'Hannu Myllykallyo. Ces données permettent de montrer que l'enzyme ThyX d'Asgard, d'origine bactérienne, est effectivement capable de synthétiser *in vitro* et *in vivo* du dTMP. Pris ensemble, ces résultats apportent des enseignements importants sur le scénario évolutif des Thymidylate synthase :

ThyA serait la version ancestrale chez les Archées et donc chez les Asgard. Mais paradoxalement l'enzyme ThyA, qui est ancestrale elle aussi chez les Eucaryotes, seraient issues au sein de ce domaine d'au moins une source bactérienne (Figure 11) et ne dériveraient donc pas des Asgards comme on pouvait le penser dans le cadre de la théorie symbiotique de l'eucaryogénèse. La présence du gène *thyX* chez un nombre limité d'Asgard, et encore plus réduits d'eucaryotes, seraient quand a eux le résultat de transferts horizontaux récents et multiples en provenance des bactéries.

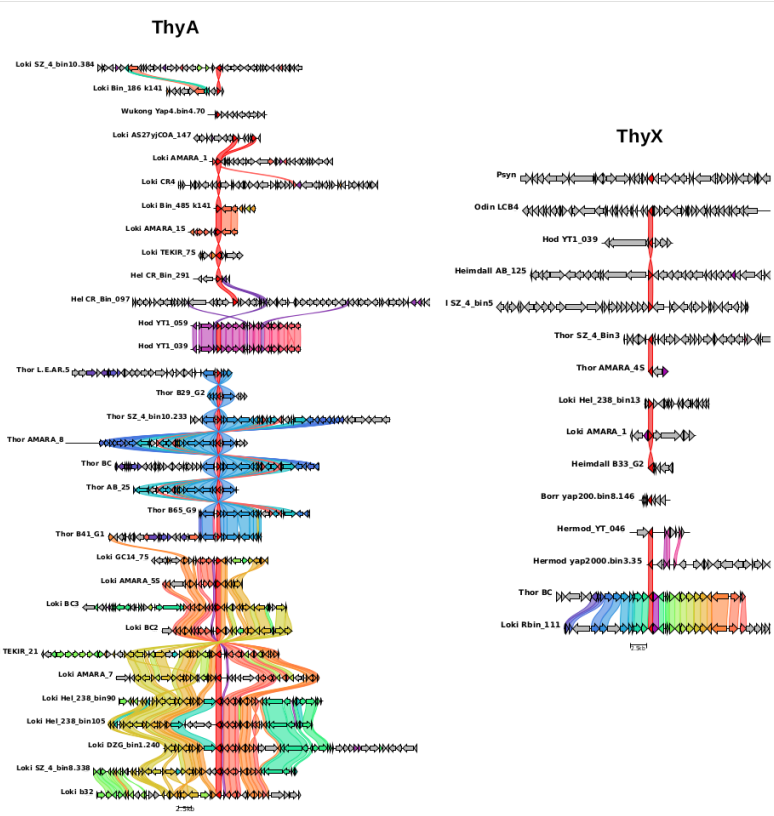

Figure 12: Analyse de la synténie des contigs d'Asgard contenant un gène codant pour une Thymidylate synthase de type ThyA (a gauche) et de type ThyX (a droite). Les contigs sont alignés au centre sur les gènes *thyA et thyX* (ruban rouge)

Publications associées a ce chapitre:

Filee J*, Becker HJ*, Mellottee L, Zhihui L, Lambry JC, Liebl U, Myllykallio H (2022) Bacterial origin of thymydilate and folate metabolism in Asgard Archea. *BioRxiv*

Myllykallio H, Leduc D, **Filée J**, Liebl U (2003) Life without dihydrofolate reductase FolA. *Trends Microbiol.* 11(5):220-3

Gadelle D*, **Filée J***, Buhler C, Forterre P (2003) Phylogenomics of type II DNA topoisomerases. *BioEssay* 25: 232-242.

Filée J, Forterre P (2005) Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol.* 13(11):510-3

Filée J, Forterre P, Laurent J (2003) *The role played by viruses on the Evolution of their cellular host: a view on informational proteins phylogenies.* *Res. Microbiol.* 154:237-43

Myllykallio H, Lipowsky G, Leduc D, **Filée J**, Forterre P, Liebl U (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* 5578:105-7.

Filée J, Forterre P, Sen-Lin T, Laurent J (2002) *Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins*. *J Mol Evol* 54:763-73.

2. Evolution des Virus a grand génome ADN

Ce chapitre sur la génomique évolutive des virus a débuté lors de mon postdoc dans le laboratoire d'Henri Krisch (LMGM/Toulouse) au milieu des années 2000' sur un des groupes de virus les plus abondants dans la nature : les phages du groupe T4. Ces virus de bactéries étaient étudiés de longue date par les biologistes moléculaires sur le phage éponyme du groupe mais nos travaux, a l'époque pionnier, sur les virus de l'environnement avait permis de révéler l'étonnante diversité de ce groupe dans les environnements marins. Nous avions alors recours a des expériences de PCR avec des amorces dirigées contre le gène de structure majeure de ces phages. Si cette approche est aujourd'hui obsolète, dépassée par la métagénomique moderne, elle avait permis de poser les premiers jalons sur l'étude de l'énorme diversité des phages T4 dans la nature. Symétriquement, nous avions été les premiers a séquencer une gamme assez complète de génomes de plusieurs phages apparentés a T4 et de montrer combien ces génomes s'articulaient au cours de l'évolution autour d'un cœur conservé d'une vingtaine de gènes de structure et de la réplication de l'ADN ainsi que d'un vaste répertoire de gènes accessoires, peu ou pas conservés. On ne parlait pas encore de «pan-génome» et ces résultats ont été par la suite largement confirmé par la métagénomique moderne et la disponibilité de vaste collection de génome environnementaux de phage T4. En parallèle a ces travaux sur les T4, je me suis intéressé a un autre groupe de virus singulier qui venait d'être découvert et dont je détaillerais une partie de ces travaux ici : les Virus Géants.

Les Virus Géants (VGs) appartiennent a une famille extrêmement diverse de virus eucaryotes : la famille des Nucleo Cytoplasmic Large DNA Viruses (NCLDVs). Ce groupe contient des virus ayant des génomes de taille très diverse de 100kb à 2M, infectant une grande variété d'hôtes, des vertébrés (Poxvirus), aux algues (Phycodnavirus), insectes (Iridovirus) ou des protistes (Mimivirus). Un très vif débat s'est ouvert pour comprendre l'origine de ces virus, la manière dont leurs génomes avaient évolués et pourquoi leurs succès évolutif est si important. L'objectif de mes travaux est notamment de mieux comprendre l'origine et l'évolution du répertoire génomique des virus géants, quantifier l'importance des transferts latéraux de gènes entre ces virus et les organismes cellulaires et déterminer le rôle des éléments génétiques mobiles présent en abondance dans ces génomes.

a. Le rôle déterminants des transferts latéraux de gènes en provenance des procaryotes

En analysant les premiers génomes disponibles de ces virus géants, nous avons tout d'abord détecté la présence de nombreux éléments génétiques mobiles typiques des procaryotes : des ISs de la famille IS607 et des endonucléases mobiles de la famille HNH. Nous avons montré que ces éléments mobiles co-localisaient avec des îlots de gènes dont l'origine phylogénétique est clairement bactérienne (Figure 13). Chez le Mimivirus, les gènes potentiellement acquis en provenance de bactéries représentent plus de 20% des gènes non orphelins (plus de 100 gènes). De plus, la comparaison de génomes proches de Phycodnavirus montre que ce processus d'acquisition latéral de gènes est un phénomène continu, dynamique et qui est responsable de l'essentiel de l'augmentation de la taille du génome chez ces virus. Par opposition, chez les autres familles de NCLDVs, parasitant des insectes et des vertébrés, la densité de gènes d'origine bactérienne est très faible voire nulle et chez les virus de métazoaires, nos analyses phylogénomiques montrent que la totalité des gènes acquis latéralement proviennent de leurs hôtes (Figure 14).

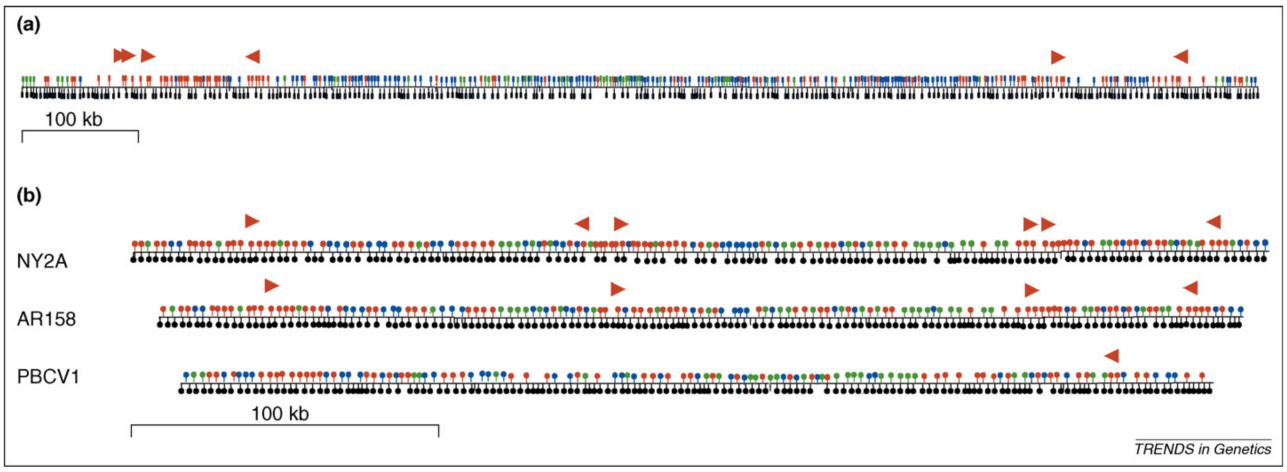

Figure 13: Carte génomique des génomes de VGs, Mimivirus en a) et les Phycodnavirus en b) . Les points rouges indiquent les gènes d'origine bactériennes, les points en verts indiquent les gènes « cœurs » des NCLDV , ceux en bleus les gènes ayant une affinité eucaryotes et en noir ceux qui sont orphelines. Les flèches rouges matérialisent la position des IS607.

Ces résultats montrent donc que les transferts horizontaux sont un élément fondamental de l'évolution des virus géants, qui expliquent en partie la taille actuelle des génomes de ces éléments. De plus, ils permettent de réfuter l'idée que les virus géants résultent de la simplification/réduction génomique d'organismes cellulaires ancestraux. Ces résultats ont contribué d'une manière décisive a installer l''idée, depuis largement confirmée par avec des jeux de données bien plus importants, que le gigantisme des VGs est un caractère dérivé récent, en lien avec les flux latéraux très importants de gènes qui affectent ces génomes.

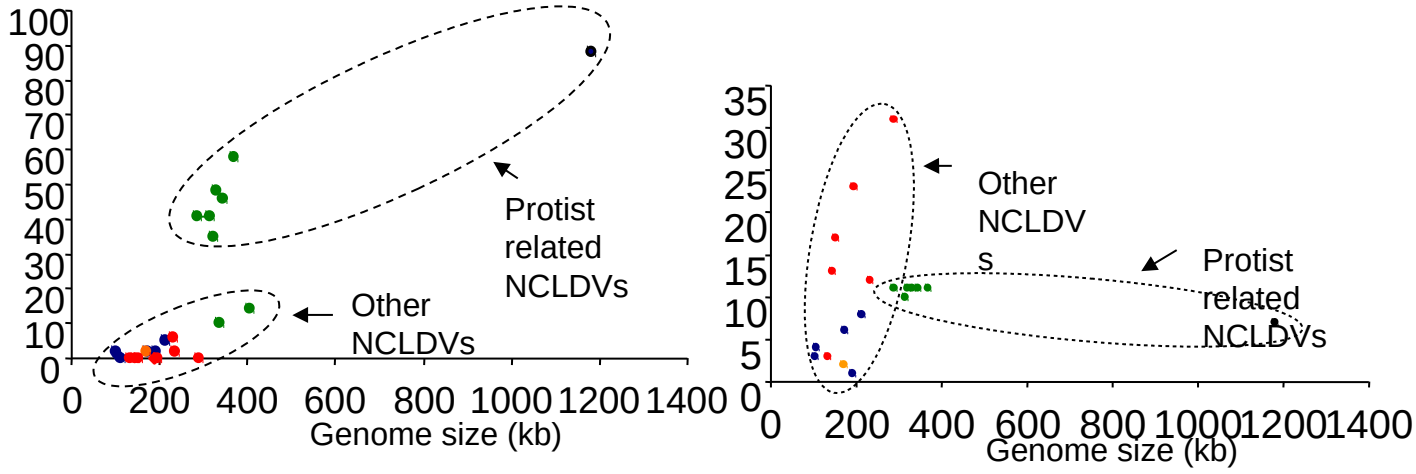

Figure 14: Nombre de gènes d'origine bactérienne (a gauche) et acquis en provenance des eucaryotes (a droite)

b. Un modèle d'évolution des génomes en accordéon chez les VGs

Une seconde approche de l'évolution des VGs repose sur la disponibilité croissante de séries de génomes phylogénétiquement très proches, pour lequel on peut aligner la quasi-totalité des séquences complètes. En alignant ces génomes, on peut déterminer le type de variations structurelles : perte ou gain d'un gène, duplication, mouvement d'un élément génétique mobile, apparition d'un gène orphelin etc... L'approche étant automatisable, on peut comparer plusieurs séries de génomes apparentés et ainsi mieux comprendre l'évolution de ces génomes à une petite échelle temporelle.

Nos résultats montrent qu'à une petite échelle évolutive, il n'y a pas de tendance globale à gagner ou à perdre des gènes au sein des 5 familles. Sur 463 événements détectés dont on peut sans ambiguïté déterminer l'origine, les pertes de gènes (212) sont à peu près compensées par des gains (251). De plus, il y a une corrélation positive entre le temps de divergence entre ces virus et le nombre d'événements génomiques : plus les virus sont phylogénétiquement éloignés (sur la base de 4 marqueurs taxonomiques), plus ils ont accumulés de variations. Ces données suggèrent fortement que les variations génomiques au sein des VGs suivent un modèle neutre avec une accumulation régulière et graduelle de mutations génomiques.

J'ai ensuite cherché à déterminer *qualitativement,* pour chacun des 5 groupes de virus, le type de variations génomiques qui ont eu cours. Dans le groupe des Megavirus, des Mimivirus et des *Chlorella* Phycodnavirus (Figure 15), les génomes sont principalement affectés par des pertes et des duplications de familles de gènes très répétés.

De plus, nos résultats indiquent aussi que les mouvements d'éléments génétiques mobiles comme les introns auto-épissables ou les transposons de la famille IS607 sont l'autre principale force évolutive agissant sur les génomes des VGs. En effet, chez les *Chlorella* Phycodnavirus, l'insertion et l'excision de ces éléments totalisent 23% des événements. Enfin, il faut noter que pour ces 3 groupes de virus, les transferts horizontaux de gènes ne représentent qu'une fraction marginale des variations génomiques observées. A la différence des 3 groupes de virus précédents, les *Ostréoccocus* et *Micromonas* Phycodnavirus présentent une situation complètement différente. Chez ces 2 groupes, caractérisés par des génomes de tailles plus faibles, se sont bien les transferts horizontaux qui représentent la force évolutive prépondérante, ces transferts étant partiellement compensés par des pertes de gènes en copie unique. La phylogénie des 71 gènes impliqués dans les transferts horizontaux indiquent que la majorité des gènes proviennent de sources procaryote bien que quelques cas d'acquisitions de gènes de l'hôte soient aussi mises en évidence.

Globalement, ces résultats montrent qu'il n'y a pas de tendance réelle à l'expansion ni à la contraction chez les génomes de VGs. De fait, chaque groupe de virus est affecté par des forces évolutives spécifiques :

- principalement des duplications chez les Megavirus et les Mimivirus,
- principalement des duplications et des mouvements d'éléments mobiles chez les *Chlorella* Phycodnavirus,
- principalement des transferts horizontaux chez les *Micromonas* et les *Ostreococcus* Phycodnavirus.

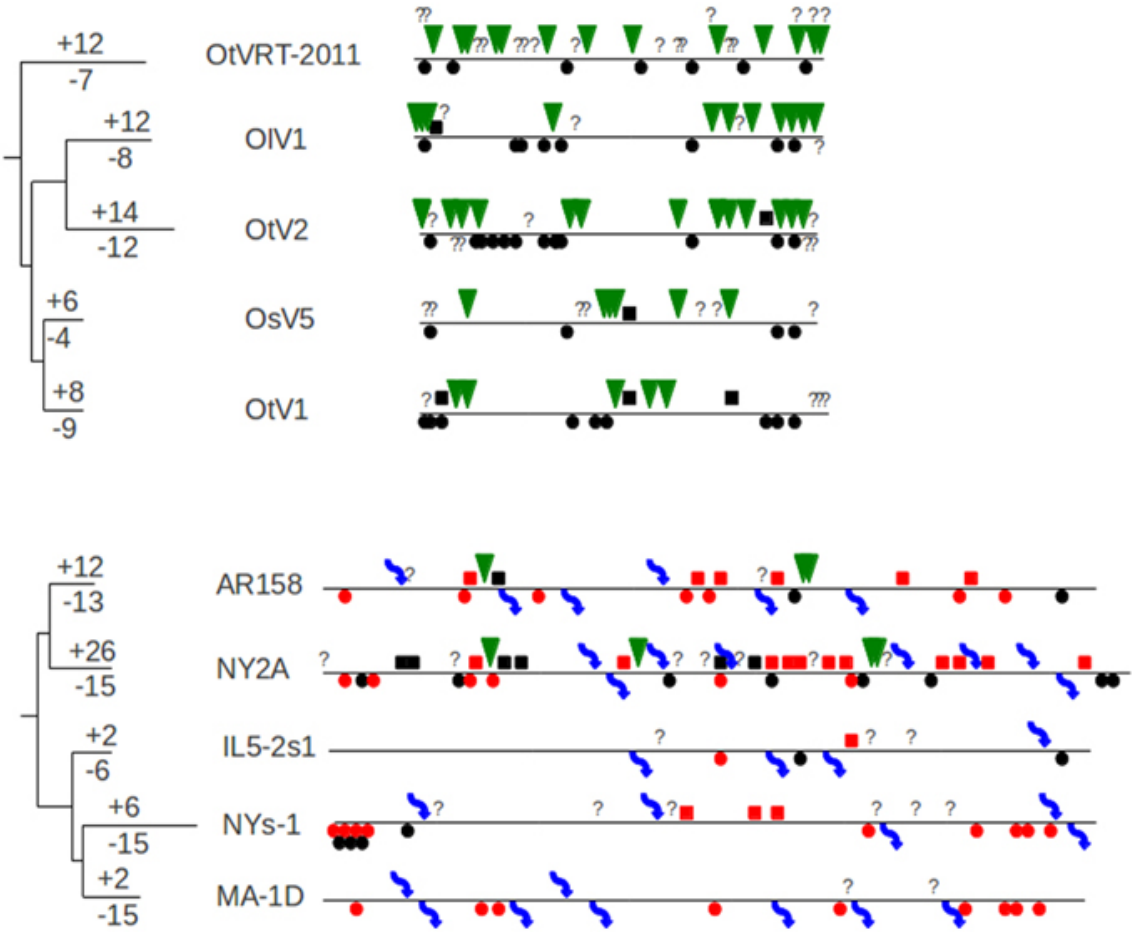

Figure 15: Carte de génomique des *Ostreococcus* Phycodnavirus (haut) et des *Chlorella* Phycodnavirus (bas) décrivant les différents événements génomiques qui ont eu lieu depuis la divergence de leurs dernière ancêtres commun respectif. Les symboles en dessous de la ligne génomique indiquent des pertes de gènes (points noirs = perte du gène, flèche = excision d'un élément mobile), au dessus des gains de gènes (carré = duplication, triangle = transferts horizontal, flèche = insertion d'un élément mobile). Les points d'interrogation indiquent un événement ambigu ou l'apparition d'un gène orphelin. Les événements en rouge correspondent aux familles de gènes répétées. Les arbres a gauche indiquent la phylogénie basée sur le génome complet et les chiffres indiquent les nombres de pertes et de gains de gène par lignée.

En conclusion, nos travaux confortent l'hypothèse selon laquelle le gigantisme des génomes chez les GV est une caractéristique apparue a plusieurs reprises mais probablement antérieurement à la diversification récente des différents rameaux évolutifs observables aujourd'hui. Ces différentes branches ont ensuite évolué en accordéon, les gains de gènes compensant finalement les pertes de gènes. Chaque rameau est donc sous l'action de forces évolutives spécifiques :

- plutôt les transferts de gènes chez les virus à spectre d'hôte étroit,
- plutôt l'expansion et la contraction de familles de gènes très répétées, dont les protéines sont notamment impliquées dans les interactions hôtes-virus chez les virus à spectre d'hôte large.

c. Les eucaryotes ont acquis des gènes de VGs : est-ce la partie émergée de l'iceberg?

La question de l'importance des transferts de gènes entre VGs et leurs hôtes eucaryotes est un sujet particulièrement polémique qui a fait l'objet de très nombreuses publications ces dix dernières années. Le point central de la controverse porte sur l'origine des nombreux gènes de VGs ayant des homologues cellulaires : ont-ils été hérités de leur dernier ancêtre commun, ou ont-ils été acquis plus récemment par transfert?

Il est hors de propos ici de lister les arguments *pro domo* des deux parties. Cependant, il apparaît de plus en plus clairement que les phylogénies de ces gènes de VGs contiennent peu de signal, entraînant des biais systématiques dans les arbres et rendant leurs interprétations très délicates. Toutefois, un certain nombre de ces gènes viraux sont réputés ne pas avoir d'homologue cellulaire proche. Ce sont les «gènes cœurs», au nombre de 23, qui auraient été hérités du dernier ancêtre viral des VGs. Ces gènes sont essentiellement des gènes de structure (codant par exemple la capside virale) et des gènes dont les protéines sont impliquées dans la réplication de l'ADN viral.

Le point de départ de ce travail est l'observation que certains de ces gènes cœurs ont pourtant bien des homologues cellulaires parmi les nombreux génomes eucaryotes récemment séquencés (tout ou partie) (Tableau 1).

Genes	**NCVOG**	**Taxa**	**Remarks**
Major Capsid Protein	NCVOG0022	*Acanthamoeba castellanii* *Hydra magnipapillata*	6 different genes 3 different genes
D5-like helicase/primase	NCVOG0023	*Guillardia theta* *Physcomitrella patens subsp. patens*	
Helicase II UL9	NCVOG0024	*Ectocarpus siliculosus*	Localized outside the pro-virus
VLTF3 transcription factor	NCVOG0262	*Guillardia theta* *Hydra magnipapillata*	
RNA helicase (COG1061)	NCVOG0076	*Guillardia theta* *Emiliania huxleyi*	3 different genes
Uracil DNA glycosylase	NCVOG1115	*Hydra magnipapillata*	
mRNA capping enzyme	NCVOG1117	*Guillardia theta*	
Nudix Hydrolase	NCVOG0236	*Dictyostellium sp.* *Polysphondylium pallidum*	

Tableau 1 : Liste des gènes coeurs de VG présent dans des génomes eucaryotes.

Les 8 hôtes concernés sont très hétérogènes: des amibes (*Acanthamoeba castellanii, Dictyostellium sp, Polysphondylium pallidum*) et des “algues” (*Emiliania huxleyi, Ectocarpus siliculosus*, *Guillardia theta*) organismes connus pour être infectés par des VGs mais aussi des hôtes dont aucune infection par des VGs n'a été reportée comme la mousse *Physcomitrella patens* ou l'hydre (cnidaire) *Hydra magnipapillata.*

Les phylogénies des gènes cœurs concernés confirment sans ambiguïté que les gènes trouvés dans les génomes eucaryotes sont étroitement apparentés à des gènes de VGs et très éloignés des homologues cellulaires (lorsqu'il y en a, exemple en Figure 16). Ces phylogénies prouvent que se sont bien les génomes hôtes eucaryotes qui ont acquis le gène en provenance d'un VG.

La totalité de ces gènes de GV acquis dans des génomes eucaryotes sont localisés dans des grands « scaffolds » ou chromosomes entiers, ces gènes sont entourés de gènes typiquement rencontrés chez

les eucaryotes. De plus, certains de ces gènes viraux ont été envahis par des introns. Ces éléments permettent d'exclure qu'il ne s'agit pas d'une contamination de l'ADN génomique de l'hôte par de l'ADN viral au cours du processus de séquençage. Ces gènes viraux sont effectivement bien intégrés dans le génome hôte et les données transcriptomiques disponibles pour certains taxa montrent qu'ils sont bien exprimés. Il est intéressant de constater que les 5 gènes viraux présents dans le génome de l'hydre co-localisent sur un fragment de 400 kb. L'analyse de ce fragment montre qu'il s'agit très vraisemblablement d'un morceau d'un génome entier de VG proche du groupe des Mimivirus. Dans ce cas précis, on ne peut donc pas complètement écarter une contamination mais la présence de très nombreux pseudogènes (densité en gène intact 17 fois inférieure aux Mimivirus séquencés) et sa composition en GC% rigoureusement identique à celle du génome de l'hydre tend à prouver qu'il s'agit d'une intégration probablement partielle d'un génome de VG dans le génome de l'hydre.

Ces gènes ont-ils une fonction cellulaire ? Il est difficile de répondre a cette question sur la base des seules données génomiques et transcriptomiques mais le cas de la D5 primase présente dans le génome du protiste *Guillardia theta* est intriguante. En effet les eucaryotes utilisent pour l'initiation de la réplication de l'ADN une ARN polymérase ADN dépendante nommée “primase”. Chez les Eucaryotes et les Archea la primase est une enzyme a deux sous-unités appelée Archeo Eukaryotic primase (gene PriS et PriL). Tous les eucaryotes sans exception utilise cette primase alors que beaucoup de virus et de phages utilisent une primase non homologue : la D5 primase. Or dans le génome de *Guillardia*, le gène codant la sous-unité catalytique de l'Archeo-Eukaryotic primase PriL est introuvable par des méthodes bioinformatiques. Il est donc possible que la D5 primase d'origine virale (issue d'un VG par transfert) ait remplacés la primase eucaryote d'origine, procurant un exemple supplémentaire de remplacement non-homologue d'une enzyme cellulaire par une enzyme virale au cours de l'évolution.

En conclusion, ce travail permet de mettre en évidence un rôle insoupçonné des VGs : celui de fournir à leurs génomes hôtes de nouveaux gènes aptes à remplir des fonctions cellulaires et peut-être remplacer des analogues fonctionnels d'origine cellulaire. Il est remarquable de noter que ce travail sur les gènes cœurs des VGs ne concernent qu'une minuscule fraction des gènes codés par un VG (23 pour plus de 1000 gènes chez les Mimivirus par exemple). On peut donc faire l'hypothèse que nous avons à faire ici à la partie émergée de l'iceberg et que le flux de transfert de gènes polarisé des VGs vers les hôtes est bien plus important.

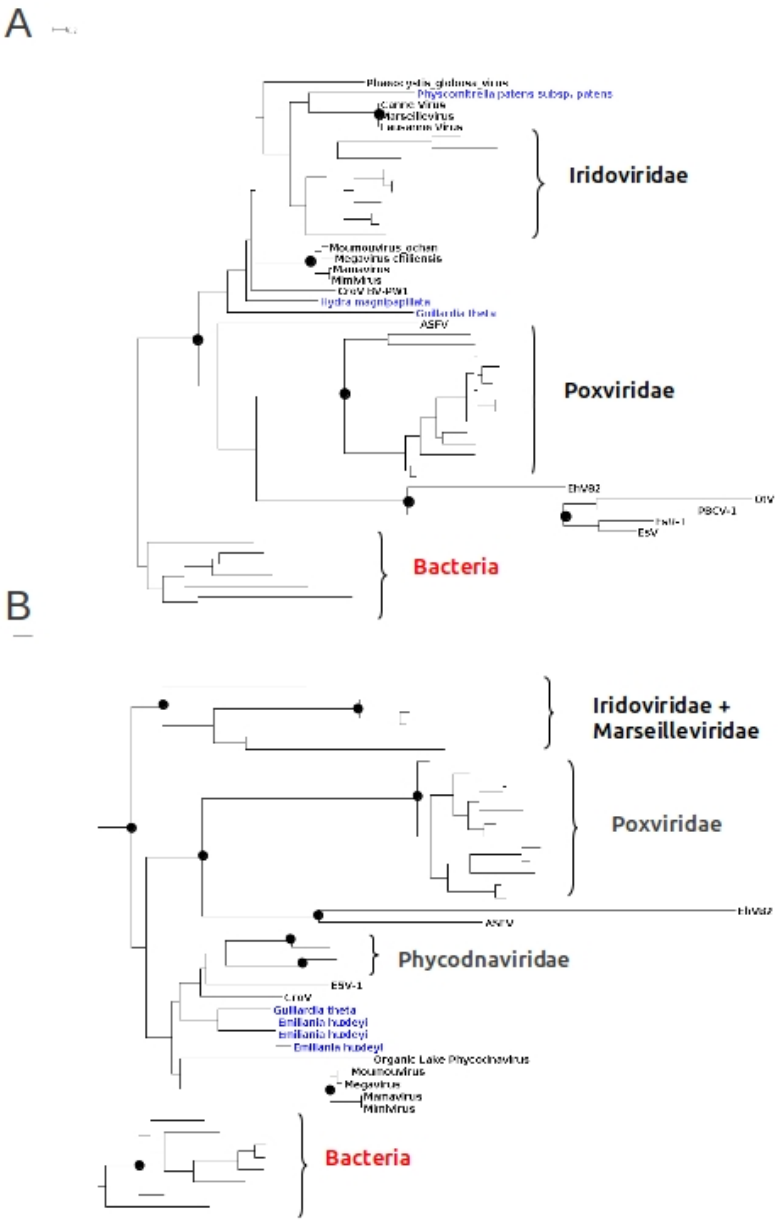

Figure 16: Phylogénie de maximum de vraisemblance de la D5 primase (A) et de la RNA helicase (B). Bleu = Séquence virale insérée dans un génome eucaryote, Noir = VG , Rouge = Bactérie, cercle noir = boostrap >95%

C. Le rôle des éléments génétiques mobiles : l'hypothèse d'une symbiose moléculaire :

Une des spécificité frappante des VGs au sein des virus repose sur l'abondance d'éléments génétiques mobiles (EGMs) dans leurs génomes. Au cours de ce travail j'ai tenter d'identifier d'une manière exhaustive les éléments répétés de tout les génomes de VG disponibles. On y retrouve une large variété d'éléments principalement d'origine procaryote : différentes familles de transposons, des introns, des intéïnes, des système de restriction-modification se propageant par transferts horizontaux ainsi que des familles d'éléments mobiles atypiques et n'ayant pas ou très peu de ressemblance avec des éléments trouvés dans les génomes cellulaires (Table 2).

Mobile Genetic Elements		Distribution	Origin	Biological roles
Insertion Sequence	IS4-like	Phycodnaviridae EsV-1 and FirrV-1	Prokaryote	Genome integration into the host/viral genome
	IS630	Aquavirinae and Mesomimivirinae	Prokaryote	
	IS607	Phycodnaviridae, Mimiviridae, Asfarviridae	Prokaryote	
	IS5	Aquavirinae	Prokaryote	
	IS3-like	Virophage Mavirus, Transpoviron	Prokaryote	
Restriction-Modification system		Phycodnaviridae, Mimiviridae, Marseilleviridae / virophage OLV	Prokaryote	Host DNA degradation, Competing virus / virophage DNA degradation
Intein and Intron-associated endonuclease	HNH	Mimiviridae, Phycodnaviridae, Asfarviridae, Marseilleviridae	Prokaryote	Cleave competing virus DNA
	GIY-YIG	Iridoviridae, Phycodnaviridae, Mesomimivirinae, Asfarviridae	Prokaryote	
	LAGLIDADG	Mesomimiviridae, Iridoviridae	Prokaryote	
Toxin/Antitoxin		*Bodo saltans* virus	Prokaryote	
DNA transposon	Mariner	Pandoraviridae, Iridoviridae	Eukaryote	Unkown but several copies have been domesticated
	Mutator	(Baculovirus)	Eukaryote	
	PiggyBac	Iridoviridae, (Herpesviridae)	Eukaryote	Endogenization
Virophage/Polinton/Transpoviron		Mimiviridae, Phycodnaviridae	Eukaryote	Host anti-viral protection
Major Interspersed Genomic Element (MIGE)		Mesomimivirinae, Aquavirinae	Prokaryote (?)	

Tableau 2 : Liste des EGMs trouvé chez les VGs et rôles biologiques potentiels

Ces éléments sont parfois trouvé en abondance jusqu'à plus de 50 copies dans certains génomes mais leurs distributions au sein de la phylogénie des VGs est erratique ce qui suppose des histoires évolutives complexes incluant des gains probablement par transferts horizontaux éventuellement suivis d'expansion ainsi que des événements de pertes. Comment expliquer cette apparente permissivité des génomes de GV aux EGMs alors que les autre génomes viraux n'en ont pas ou rarement ? Une hypothèse possible est que dans certains cas, ces EGMs puissent conférer un avantage adaptatif a leurs hôtes. En analysant la littérature, ces EGMs pourraient en effet être bénéfique aux virus en remplissant deux types de fonctions:

1) Détruire l'ADN des cellules infectées ou dégrader l'ADN de virus compétiteurs. Ce cas de figure est connu chez certains virus bactériens qui ont domestiqué des systèmes de restriction-modification (RM) de l'ADN pour dégrader les génomes de virus compétiteurs. Ces systèmes sont composés d'un tandem enzymatique avec une enzyme qui modifie chimiquement le génome qui les codent d'une manière site-spécifique et une enzyme qui dégrade à ces même sites l'ADN non-modifié. Ce système constitue ainsi un système immunitaire rudimentaire qui permet de dégrader tout ADN étrangers. Chez les VGs de la famille des *Chlorella* Phycodnavirus, il a été démontré que ces systèmes RM sont utilisés pour dégrader spécifiquement le génome de leurs hôtes cellulaires. Il s'agit donc d'une domestication d'un EGM pour accomplir des fonctions virales. De la même manière, il est connu chez certains virus bactérien que les homing endonucleases codées par des introns et des intéïnes sont utilisées pour détruire des virus compétiteurs co-infectant la même cellule.

En effet, les gènes non-protégés par l'intéïne ou l'intron sont spécifiquement clivés par l'endonucléase, conférant aux virus porteurs de ces EGMs une arme de destruction des autres virus. Toutefois, la domestication de ces systèmes RM et des introns/intéines par les GV a une contrepartie, en effet, la perte de ces éléments devient impossible: toute perte est suivie d'une dégradation de l'ADN du GV par les enzymes de clivage qui sont connues comme ayant un temps de demi-vie beaucoup plus long que les enzymes de protections. Ces EGMs « protecteur » auraient donc aussi un effet addictif.

2) L'autre grande famille de domestication d'EGM par les VGs reposent sur la capacité des VGs a intégrer les génomes cellulaires hôtes pour initier un cycle de latence (ou de dormance). C'est notamment le cas des nombreuses intégrases/transposases qui semblent jouer un rôle clé dans ce processus d'endogénisation. Plusieurs exemples existent dans la littérature : chez des Phycodnavirus d'algue qui ont un cycle biologique complexe comme les *Ectocarpus*, chez des Herpesvirus de poisson qui ont généré des formes virales hybride entre virus et transposon capable de se dupliquer dans les génomes cellulaire et enfin chez les Virophages, qui sont des formes miniatures de VGs ayant généré des formes non-infectieuses qui se propagent essentiellement verticalement dans les génomes hôtes. Ces cas de domestications semblent très nombreux et offrent d'excellent exemples de coopération bi-partites entre les VGs et les EGMs dans le but de maximiser la fitness du système ainsi composé.

En conclusion, ces travaux permettent de proposer l'hypothèse que l'abondance des EGMs au sein des génomes de VGs s'expliquent par une sorte de symbiose moléculaire entre ces éléments égoïstes dans laquelle les EGMs sont utilisés pour accomplir des fonctions virales (destruction de l'ADN hôte, de compétiteurs ou processus d'endogénisation). Toutefois, les EGMs impliqués dans la dégradation de l'ADN sont addictifs car leur perte est délétère pour le VGs (destruction de l'ADN du VG). On peut aussi penser qu'il existe une course aux armements entre VGs, favorisant les virus ayant un arsenal de système de dégradation le plus divers possible. Ces EGM auraient ainsi tendance a s'accumuler par un effet de cliquet.

Publications associées a ce chapitre :

Filée J (2018) Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. Curr Opin Virol. 33, 81

Filée J (2015). Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Front Microbiol.* Jun 16;6:593

Filée J (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology*. Oct;466-467:53-9

Filée J. (2013) Route of NCLDV evolution: the genomic accordion. *Curr Opin Virol.* 3(5):595-9

Filée J, Chandler M. (2010) Gene exchange and the origin of giant viruses. *Intervirology.* 2;53(5):354-61

Filée J (2009) Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. *J Invertebr Pathol.* 101(3):169-71

Filée J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic Large DNA Viruses. *BMC Evol Biol.* 26;8:320.

Filée J, Chandler M (2008) *Convergent Mechanisms of Genome Evolution of Large and Giant DNA Viruses.* *Res. Microbiol.* 159(5):325-31

Filée J, Siguier P, Chandler M (2007) *I am what I eat and I eat what I am: acquisition of bacterial genes by Giant Viruses.* *Trends Genet.* 23(1):10-5

Filée J, Bapteste E, Susko E, Krisch HM (2006) *A Selective Barrier to Horizontal Gene Transfer in the T4-Type Bacteriophages that Has Preserved a Core Genome with the Viral Replication and Structural Genes.* *Mol. Biol. Evol.* 23(9):1688-1696

Filée J, Comeau AM, Suttle CA, Krisch HM (2006) *T4-type bacteriophages.* *Med Sci.* 22(2):111-2.

Filée J, Tetart F, Suttle CA, Krisch HM (2005) *Marine T4-type bacteriophages, a ubiquitous component of the dark Matter of the biosphere.* *Proc.Natl.Acad.Sci.USA.* 102(35):12471-6

3. Rôle des Éléments Transposables dans la structure et l'évolution des génomes.

J'ai commencé a m'intéresser aux transposons lors de mon postdoc chez Michael Chandler (LMGM/Toulouse) et ce compartiment des génomes, le «mobilome», a certainement été depuis le sujet le plus structurant dans ma carrière scientifique. J'ai débuté par l'étude des Séquences d'Insertions (ISs) chez les prokaryotes, notamment chez les Archées où nos travaux avaient permis de montrer que le mobilome des archées était très semblables aux mobilomes des bactéries en raison des très important flux latéraux d'IS inter-domaines qui avaient aboutis a une homogénéisation du mobilome chez les procaryotes. Depuis mon recrutement au CNRS a Gif-sur-Yvette, mes travaux se sont focalisés sur une large variété d'organismes eucaryotes a la fois pour étudier la diversité, le rôle des Éléments Transposables (ETs) dans la structure et la taille des génomes mais aussi pour étudier la réponse de l'hôte en terme de régulation. Plusieurs de ces travaux ont été détaillé dans les parties suivantes.

a. Expansion des génomes chez les Vertébrés: rôle joué par les transposons dans la taille du génome de la Dipneuste.

La Dipneuste d'Australie *(Neoceratodus forsteri*) a un des plus grands génomes connus : 105 000 Mb soit 15 fois la taille du génome humain. La thématique générale de ces travaux, menés en collaboration avec l'équipe de Didier Casane (EGCE/Gif sur Yvette), est de comprendre pourquoi ce génome est si grand et notamment de déterminer si des ETs peuvent ou non expliquer cette taille. L'enjeu principal est de préciser la part occupée par les transposons dans ces génomes, la diversité des éléments présents et la dynamique évolutives de ceux ci.

L'analyse de trois librairies mini-génomiques montrent qu'environ 38% du génome serait composé de séquences répétées : 31% de rétrotransposons non-LTR, 6% de rétrotransposons à LTR (DIRS principalement) et moins de 1% de transposons ADN (Mariner principalement). Parmi les rétrotransposons non-LTR se sont les éléments de la famille CR1 qui sont majoritaires et qui représentent à eux seuls environ 15% du génome de la Dipneuste. En utilisant des expériences de "genome walking", plusieurs éléments CR1 entiers ont été séquencés complètement. L'échantillonnage de séquences codant pour la Reverse Transcriptase et l'Endonuclease a été complété par des expériences de PCR. La figure 17 montre une phylogénie de la famille CR1 obtenue avec ces séquences. Deux lignées de séquences sont discernables. En comparant avec les séquences obtenues avec les librairies mini-génomiques, nous pouvons estimer que 70% des séquences CR1 des librairies appartiennent au clade 1 et 30% au clade 2. Les séquences du clade 1, fortement apparentées entre elles, résultent probablement d'une expansion récente et explosive d'éléments CR1 dans le génome de la Dipneuste. Les séquences du clade 2 résultent probablement d'expansions plus anciennes, même si la datation des événements reste plus spéculative pour ce groupe de séquences.

En conclusion il est possible de proposer un scénario évolutif qui privilégie une expansion ancienne des éléments de type de CR1 qui ont ainsi occupé une place très importante dans le génome de la Dipneuste. Ensuite, des événements plus récents d'expansion ont eu lieu, permettant soit (i) au génome de maintenir une taille importante par un jeu d'équilibre entre des pertes par délétion compensées par des transpositions soit (ii) une augmentation progressive de la taille du génome au cours du temps.

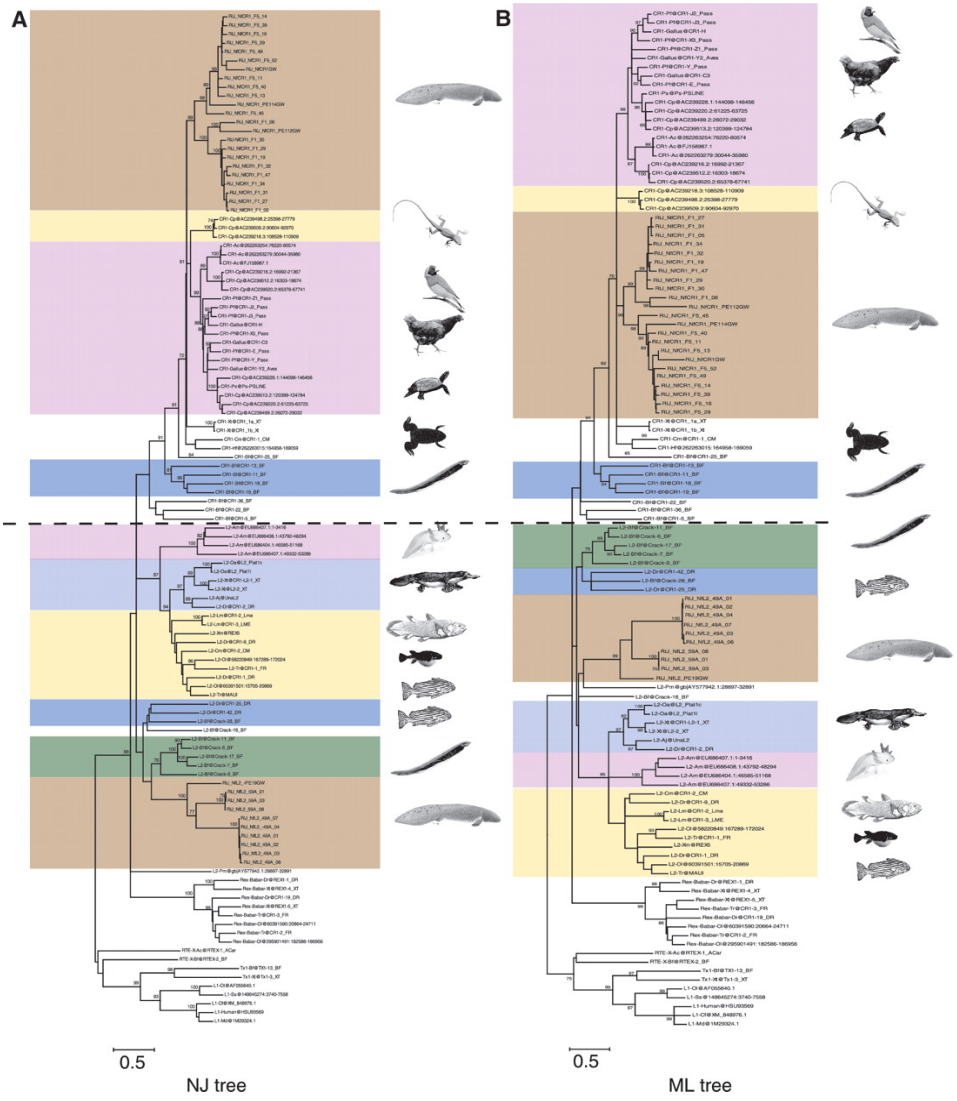

Figure 17 : Phylogénie Neighbour joining (A) et Maximum de Vraisemblance (B) de la famille CR1. Les séquences de dipneustes sont indiquées en marron.

b. Etude des ETs dans le génome du Triatomine Rhodnius prolixus.

La maladie de Chagas, l'une des principales maladies tropicales avec environ 10 millions de personnes infectées, est transmise par les punaises hématophages du genre *Rhodnius*. Ces punaises sont les vecteurs du parasite *Trypanosoma cruzi*, endémique en Amérique Latine. L'épidémiologie de la maladie de Chagas semble être liée a un processus de domiciliation des punaises, impliquant l'adaptation de populations sauvages sylvatiques aux environnements humains. Les déterminants génétiques de l'adaptation de ces punaises a l'homme est mal connu mais nous avons identifié deux

types de facteurs qui pouvait jouer un rôle : les éléments transposables et les symbiotes bactériens. Les travaux sur les symbiotes seront détaillés plus tard dans le manuscrit et nous allons nous focaliser ici sur le rôle des éléments transposables. Ce projet est mené dans le cadre d'une collaboration avec l'équipe de Carlos Almeida (Université de Campina, Brésil), de Myriam Harry (EGCE-IRD, Gif sur Yvette), Aurélie Hua-Van (EGCE-CNRS, Gif sur Yvette) et de plusieurs équipes partenaires impliquées dans la collecte des insectes (MNHN-Paris, IESE-Créteil, URMITE-Marseille...).
En utilisant la combinaison de méthodes de détection *de novo* et de méthodes basées sur la ressemblance avec une librairie d'ETs, nous avons déterminé la composition globale du génome en transposons (Figure 18). Le génome comprend environ 6% d'ETs, un chiffre plutôt modeste si on le compare au pourcentage observé chez des insectes ayant des tailles de génomes proches. De plus, les transposons ADN sont très largement dominants avec une sur-représentation d'ETs de la superfamille *mariner* (11000 copies, soit plus de 75% du total des ETs).
Nous avons montré qu'il y a une large variété d'éléments *mariner* dans le génome (90 familles environ) dont plusieurs groupes représentant de nouveaux clades, très divergents des autres clades connus au sein de la superfamille. Plusieurs familles sont de plus composées d'éléments non-autonomes utilisant en *trans* une transposase codée par un élément autonome. Ces familles d'éléments semblent avoir été générées classiquement par des délétions internes successives à partir d'éléments complets. Toutefois deux familles non-autonomes présentent un patron original : elles ont vraisemblablement été générées à partir de recombinaisons entre deux éléments placés cote à cote en position inverse.

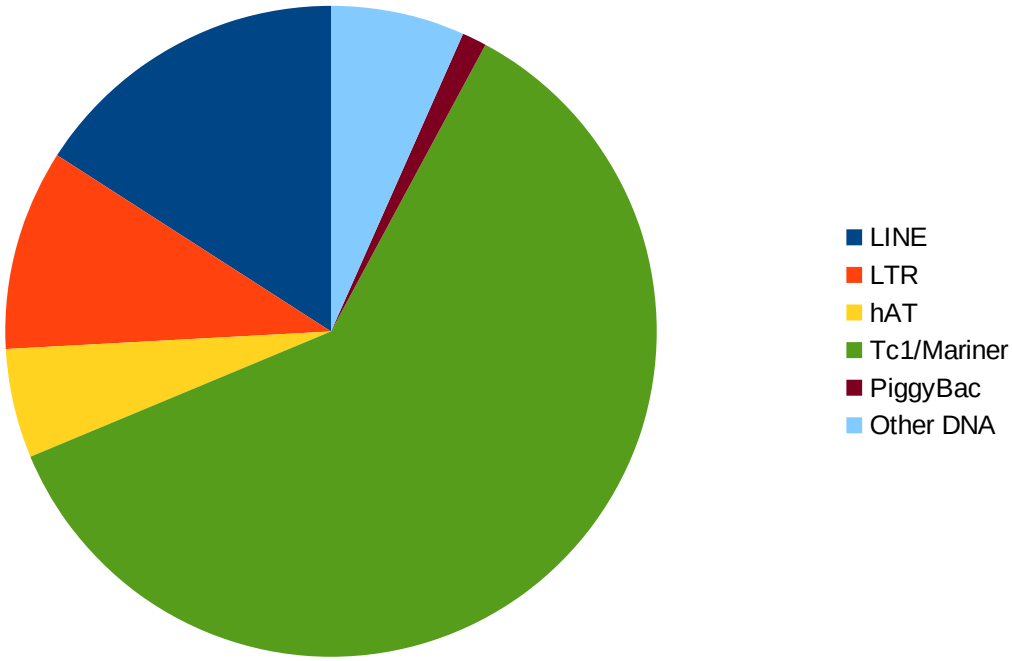

Figure 18: Répartition en superfamille des ETs dans le génome de *R. prolixus*

L'étude de la dynamique évolutive des différentes familles d'éléments mariner présents dans le génome de *R. prolixus* montre que 5 d'entre elles ont connu des expansions très récentes dont 4 semblent encore actives aujourd'hui (Figure 19a). C'est notablement vrai pour les familles les plus abondantes comme les familles Rpmar0 (>8000 copies) ou Rpmar1 (800 copies). Toutes les autres familles semblent aujourd'hui éteintes. Une analyse plus fine des événements de transposition en fonction du temps (Figure 19b) montre qu'au moins 4 patrons de transpositions sont discernables. La

famille Rpmar1 présente une courbe en « S », signifiant une augmentation lente du nombre de copies, puis une augmentation plus forte et enfin un ralentissement récent du niveau de transposition pouvant indiquer une perte progressive d'activité (et/ou une régulation par le génome hôte). La famille Rpmar0 présente une courbe «concave» compatible avec une expansion initiale explosive suivie d'un ralentissement progressif sans que la famille s'éteigne. Des patrons intermédiaires sont aussi discernables pour des familles aujourd'hui inactives. Prises ensemble, ces données montrent que le génome de *R. prolixus* a été régulièrement envahi au cours du temps par différentes familles d'éléments *mariner*. Certaines d'entre elles ont atteint un nombre de copies important et sont encore capables de transposer, sans que le génome hôte ne les régule totalement.

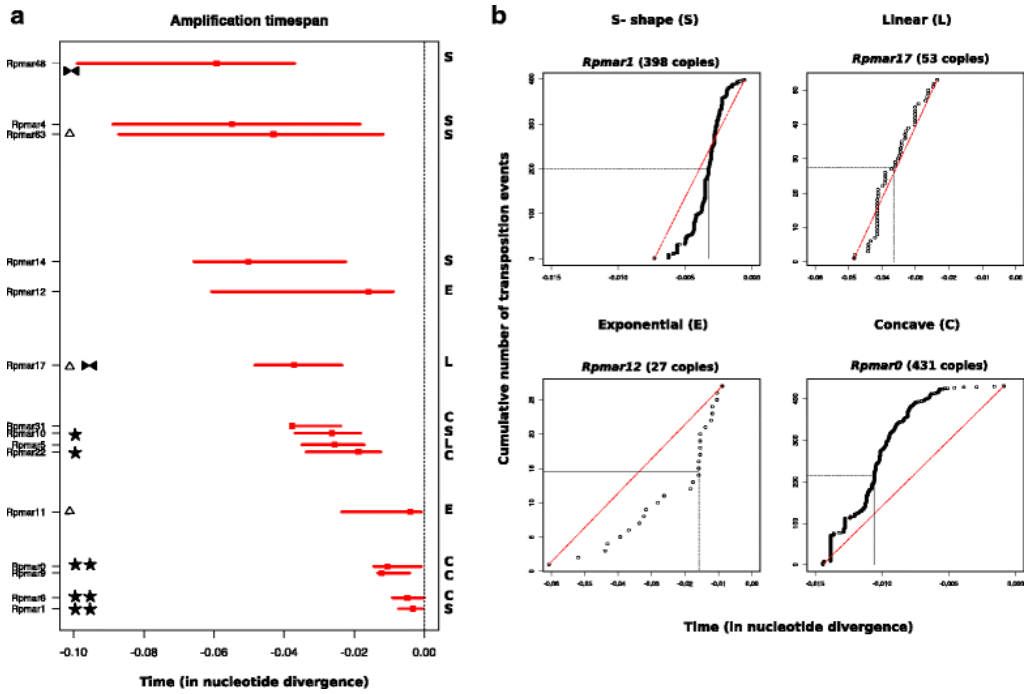

Figure 19: Dynamique d'amplification des familles d'élément *mariner*. (a) Laps de temps de l'amplification de chaque famille (barre rouge) de la plus ancienne (en haut) à la plus récente (en bas). Le carré rouge indique l'événement de transposition médian. Δ: délétion interne ; *: peu de copies entières ; ** : beaucoup de copies entières ; >< :famille issue de recombinaison interne. (b) Courbes décrivant le nombre cumulatif de transposition en fonction du temps. La droite en rouge indique un taux constant de transposition. La position de l'événement médian de transposition est indiquée par les 2 lignes verticales et horizontales en pointillé.

Un point important de ce travail a été de déterminer si ces invasions récurrentes d'ETs pouvaient provenir de transferts horizontaux. Nous avons détecté au moins 10 cas de transferts où l'élément présent chez *R. prolixus* présentait plus de 95% de similarité nucléotidique avec un élément présent dans le génome d'une espèce phylogénétiquement distante, excluant ainsi la possibilité d'un héritage vertical depuis un ancêtre commun. Les espèces incriminées sont remarquablement diverses : des insectes, des vers parasites, une guêpe parasitoïde, une chauve-souris sud américaine etc. Pour la chauve-souris et la guêpe parasitoïde, un transfert direct entre *R. prolixus* et ces taxa semble possible. Pour les autres cas de transferts avec les autres insectes (drosophile, bourdon etc...), la guêpe parasitoïde aurait pu jouer un rôle d'intermédiaire. Enfin, pour les vers se nourrissant de sang, deux transferts indépendants provenant du même hôte inconnu sont envisageables..

En conclusion, nos travaux montrent que les éléments *mariner* dominent le compartiment mobile du génome de *R. prolixus*. Cette prépondérance s'explique par 3 facteurs :

- des invasions anciennes et d'autres plus récente dont certaines sont encore en cours, une seule famille encore active aujourd'hui totalisant plus de 8000 copies (soit 2/3 à elle seule du total).

- une grande diversité d'éléments, dont des éléments non-autonomes, pour certains générés par des mécanismes de recombinaison originaux.

- de fréquent événements de transferts horizontaux d'éléments *mariner* ont lieu avec une grande diversité d'organismes, permettant de compenser l'inactivation ou la perte d'autres familles de transposons.

b. Génomique comparatives des rétrotransposons de type LTR chez les Spiraliens marins (mollusques et annélides).

L'objectif de ces travaux mené en collaboration avec Eric Bonnivard (Station biologique de Roscoff) est d'étudier la diversité des ET chez une assez large collection d'organismes marins appartenant aux clades des *Spiralia* où fort peu de données génomiques étaient disponibles. Nous avons utilisé plusieurs approches d'analyses combinant et comparant des données obtenues par séquençage complet des génomes, par assemblage des transcriptomes et enfin par assemblage d'une faible quantité de « reads » génomiques pour cible spécifiquement les séquences répétées.

Que se soit chez les mollusques ou chez les annélides (Figure 20), nos travaux ont montré que ces génomes contenaient une grande quantité d'ET, jusqu'à près de 60% du génome, et que ce mobilome étaient dominés par des rétrotransposons LTR même si beaucoup de séquences répétées ne peuvent pas être assignées a une superfamille connue . Il existe de plus une corrélation positive entre la taille des génomes et la quantité d'ET présent. Nos travaux ont aussi montré que l'assemblage des transcriptomes a partir d'un niveau assez faible de couverture (a partir de 40 millions de reads par transcriptome) étaient une approche économique qui permettait de détecter l'essentiel de la diversité des familles des ETs présent dans les données génomiques.

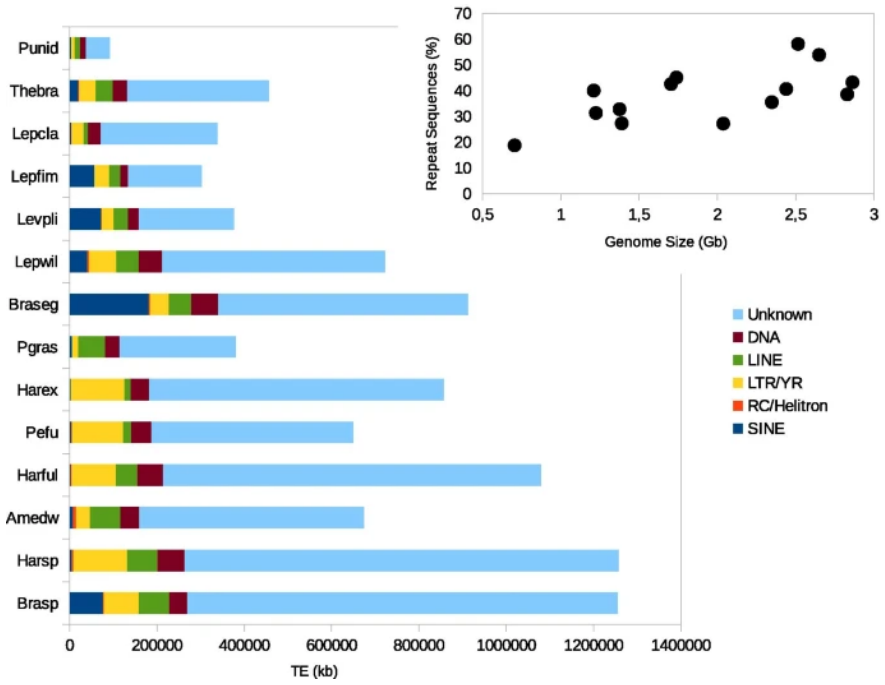

Figure 20: Proportions occupées par les différentes superfamilles d'ET au sein des génomes d'Annélides et contenu total (en %) des ETs dans ces génomes en fonctions de leurs tailles.

L'analyse phylogénétique combinée de ces données génomiques et transcriptomiques a permis de classifier plus de 1000 rétrotransposons dans 26 génomes d'annélides. Si les familles de Copia et BEL/ Pao reste rares chez les annélides, les rétrotransposons a LTR appartenant aux éléments Gypsy sont

beaucoup plus divers permettant de définir 4 nouveaux clades dans ce groupe.

c. Étude du mécanisme de défense contre les transposons par la voie des piARN chez deux espèces sœurs de Drosophile.

Chez les métazoaires, afin de se protéger des effets délétères des ETs dans les lignées cellulaires germinales, les espèces ont développé un mécanisme de régulation basé sur l'interférence à ARN.
En effet, il a été récemment mis en évidence que les génomes de ces espèces possèdent des locus appelé les « clusters à *pi*ARN » produisant des petits ARN de 23-29,nt qui ciblent spécifiquement les ARNm produits par les ETs. Ces ARNm sont alors spécifiquement dégradés par la machinerie cellulaire. En fonction du type de clusters et des protéines de défense impliquées, un processus d'amplification secondaire des *pi*ARN peut se mettre en place, appelée boucle d'amplification «ping-pong». Il reste à comprendre les mécanismes et le tempo de fonctionnement de la voie des *pi*ARN mais il est désormais admis que la régulation se met en route à partir du moment où un ET s'insère fortuitement dans un cluster, initiant la production des *pi*ARN qui vont le cibler spécifiquement et en réguler l'activité. L'objectif de cette étude est de mieux comprendre le lien entre l'histoire évolutive des ETs, leurs niveaux d'activité et la manière dont la régulation par la voie des piARN se met en place. Nous avons choisis de nous intéresser à deux espèces sœurs de drosophile (*D. melanogaster* et *D. simulans*) qui ont des contenus et une histoire évolutive des ETs complètement différente. Nous avons étudié comment et avec quelle dynamique la voie des piARN a pu réguler l'expansion des ETs dans ces deux génomes.

Avec Bastien Saint-Léandre que j'ai coencadré lors de sa thèse avec Aurélie Hua-Van (EGCE- Gif sur Yvette), nous avons montré que le génome *D. melanogaster* connaît une invasion récente - et toujours en cours - par les ETs. Cette invasion est caractérisée par une forte abondance d'éléments très similaires et de taille entière. Par opposition, le génome de *D. simulans* est caractérisé par de très nombreux éléments délétés, très divergents entre eux, traduisant des invasions anciennes et des ETs désormais régulés ou inactivés. En séquençant le transcriptome global ainsi que des banques de piARN germinales pour deux populations de chaque espèce, nous avons établi que pour les deux espèces le mécanisme de régulation par les piARN est seulement effectif dans la lignée germinale femelle. En effet, les piARN ciblant les ETs dans les testicules sont pratiquement absents pour les deux espèces alors qu'ils sont abondants dans les ovaires (Figure 21) avec en plus une importante réponse ping-pong (non montré).

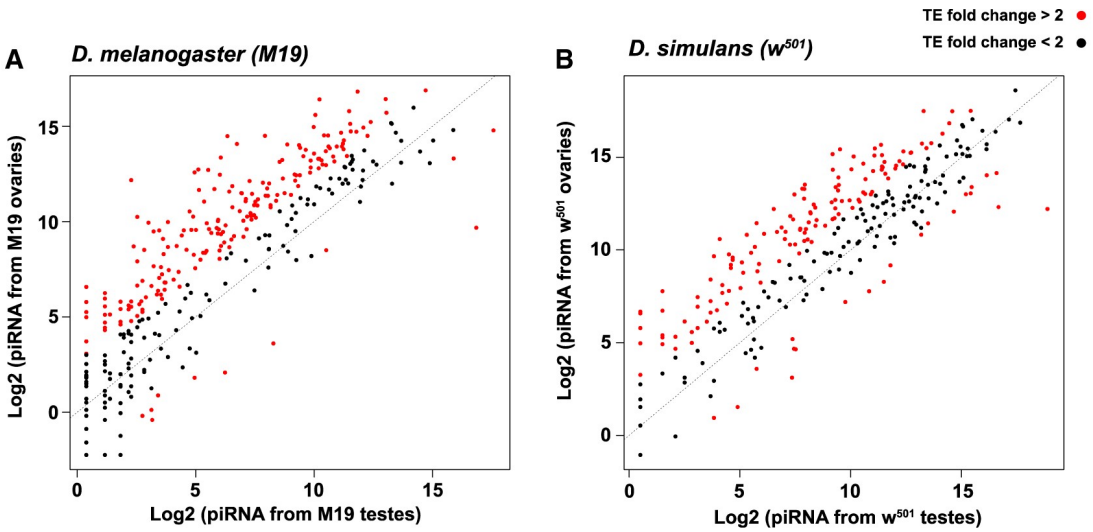

Figure 21: Expression des *pi*ARN dans les ovaires et les testicules ciblant chaque famille d'ETs chez *D. melanogaster* (A) et chez *D. simulans* (B)

Nous proposons que cette dissymétrie de la voie des *pi*ARN dans les cellules germinales est liée au coût métabolique global très important du processus qui nécessite un réseau complexe de protéines fortement exprimées.
Pour étudier la dynamique des TEs au sein des loci producteurs de piARN et les conséquences sur l'expression biaisée par le sexe, nous avons comparé la densité des piARN des TEs en fonction de leur biais d'expression. Chez les deux espèces, nous avons localisé les piARN et comparé leur distribution génomique dans les testicules et les ovaires (fig. 22). Le cluster 42AB, connu sous le nom de «master locus»» de piARN chez *D. melanogaster* est actif sur le plan de la transcription à la fois dans les testicules et les ovaires (fig. 22A) Néanmoins, nous pouvons également identifier d'autres clusters qui ne sont exprimés que dans les ovaires (par exemple, le cluster de piARN péricentrique du chromosome 2R, fig. 22B), et des clusters qui ne sont actifs sur le plan de la transcription que dans les testicules (fig. 22C).
De plus nous avons effectué un screening global du génome pour les loci producteurs de piARN dans une fenêtre de 1 kb et avons montrés que la plupart des clusters piARN sont actifs dans les ovaires (non montré). Chez *D. melanogaster*, 4 % des locus piARN sont spécifiques aux testicules, 5 % ont été trouvé exprimé dans les deux lignées germinales et 91 % n'ont été exprimé que dans les ovaires. Chez *D. simulans*, les clusters de piARN spécifiques aux femelles représentent également 91% de tous les clusters de piARN. Ces données renforcent clairement l'idée selon laquelle la lignée germinale féminine est le principal tissu impliqué dans le silencing des TEs et expliquent également la tendance globale à produire moins de piRNA dans les testicules.

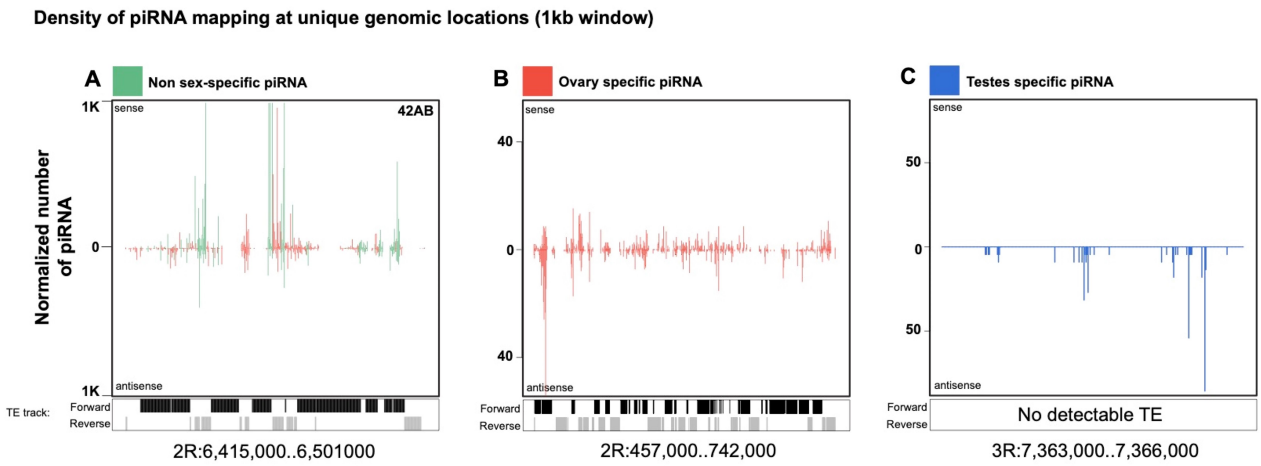

Figure 22: Cartes génomiques montrant le nombre de piRNA unique le long des régions du chromosome 2R (A, B) et 3R (C) chez *Drosophila melanogaster*. Les barres colorées indiquent le nombre de piARN en fonction de leur mode d'expression : les barres vertes affichent les piARN exprimés à la fois dans les testicules et les ovaires, les barres rouges les piARN exprimés exclusivement dans les ovaires et les barres bleues les piARN exprimés exclusivement dans les testicules. La composition en ET est indiquée en dessous : le noir montre les ETs insérés en sens direct et les ETs en gris sont ceux insérés en sens inverse.

Par conséquent, il semble que lorsqu'une espèce connaît une expansion intense d'ET (comme chez *D. melanogaster*), une accumulation de fragments de ETs se produit dans les régions génomiques dédiées à la production spécifique de piARN ovariens, conduisant à un fort contraste d'expression des ETs dans les deux lignées germinales. Ensuite, lorsque l'expansion des ET est sous contrôle (comme chez *D. simulans*), une perte progressive des ETs dans ces clusters de piARN femelle-exclusif se produit, ce qui rééquilibre le pattern d'expression des ETs entre les testicules et les ovaires. Outre le fait de montrer pour la première fois que la régulation des ETs par la voie des piARN est un mécanisme femelle-spécifique chez les drosophiles, ce travail permet de proposer un modèle d'action de la voie

des piARN dans lequel des effets à court terme et à long terme se conjuguent pour limiter l'expansion des ETs dans les génomes.

d. Développement de nouvelles méthodes bioinformatiques de détections des ETs.

Le développement de nouvelles méthodes d'identification des ETs dans les génomes, en particulier dans un contexte d'évolution permanente des technologies de séquençage, est un enjeu important de ce domaine d'études. Bien que je ne me sois jamais réellement définis comme un bioinformaticien, la mise au point de pipelines automatisant et enchaînant des programmes écris en langage Python a occupé une partie notable de mon temps travail et ont fait l'objet de deux publications.

La première que je ne détaillerais que très brièvement ici, a consisté à développer une méthode d'identification des ISs procaryotes avec des profils d'alignements de transposases/recombinases et des méthode de recherche par similarité de séquence basé sur des «modèles de Markov cachés». Cette approche étant plus sensible (et plus rapides) que les méthodes classiques de recherches basé sur l'algorithme BLAST contre une base de données de séquences de références et donc capable d'identifier des IS tres divergents et peu ressemblant a tout ce qui était connu. Testé sur des métagénomes et une collection de génomes assemblés, la méthode s'est avéré bien plus performantes que les approches de types BLAST (+50% et +20% respectivements d'ISs détectés).

La seconde méthode que j'ai développé est un soft qui permet de détecter les mouvements d'ETs entre un génome de référence et les données de séquençage directement avec les reads « longs » de type PacBio ou Minion a bas niveau de couverture (donc sans passer des étapes d'assemblage de ces reads). Ce programme nommé LoRTE est un programme Python qui fonctionne avec deux modules. Brièvement, le premier module utilise comme fichier d'entrée une liste d'ETs dans un génome de référence et regarde la présence ou l'absence de chaque ET dans les données de séquençage de type « long read ». Le deuxième module recherche les nouvelles insertions d'ET dans les reads qui ne sont pas présentes dans le génome de référence. Les fichiers de sortie du programme comprennent notamment une liste avec les coordonnées génomiques de chaque nouveaux événements, leurs niveaux de couvertures pour estimer la fiabilité de chaque prédiction et enfin les séquences en question.

Nous avons testé l'outil sur deux jeux de données PacBio provenant de la drosophile *D. melanogaster*. Un jeu de donnée « simulé » en découpant le génome de *D. melanogaster* en fragment de taille similaire aux séquences Pacbio (5 a 20kb et 10% de taux de mutation) dans lequel nous avons insérer 100 nouveaux ETs et délété 250 autres. L'autre jeu de donnée est un jeu de donnée PacBio réel de *D. melanogaster* provenant de la même souche que le génome de référence. Nous avons d'abord testé la capacité de LoRTE a faire des prédictions sur une liste de 4000 ETs annoté en fonction de la couverture des reads (Figure 23A). Pour les deux jeux de données, le programme est capable de faire une prédiction pour >99% des locus a partir d'un bas de niveau de couverture (a partir de 9x, soit une quantité totale de lecture 9 fois équivalente au génome de la drosophile). Nous avons ensuite testé la capacité de l'outil a détecter les insertions/déletions que nous avons artificiellement généré dans les données PacBio simulées (Fig 23B). Même avec un niveau de couverture faible (>10x) le programme détecte presque 100% des événements. De plus, aucun faux positif n'est détecté. Enfin nous avons utilisé les données PacBio réelle de drosophile (Fig 23C). Le nombre de délétions observées est relativement constant quelques soit la quantité de lecture utilisée, alors que les prédictions d'insertions ont tendance a augmenter proportionnellement. Nous avons pu valider les prédictions des délétions en utilisant un assemblage génomique a haut niveau de couverture des données PacBio. Par contre dans le cas des nouvelles insertions, seul une petite moitié des prédictions sont validées. Sachant que les prédictions non vérifiables sont soutenues par un très faible de séquence PacBio, généralement 1 seule même avec avec des jeux de donnée équivalent a 40 fois le génome total, on peut penser qu'il s'agit de

polymorphisme somatique a basse fréquence. Toutefois la présence de quelques faux positifs ne peut pas être exclue.
En conclusion, LoRTE est un outil fiable qui permet d'étudier l'activité et l'impact des ETs dans les populations naturelles en utilisant des bas niveau de couverture de lectures PacBio.

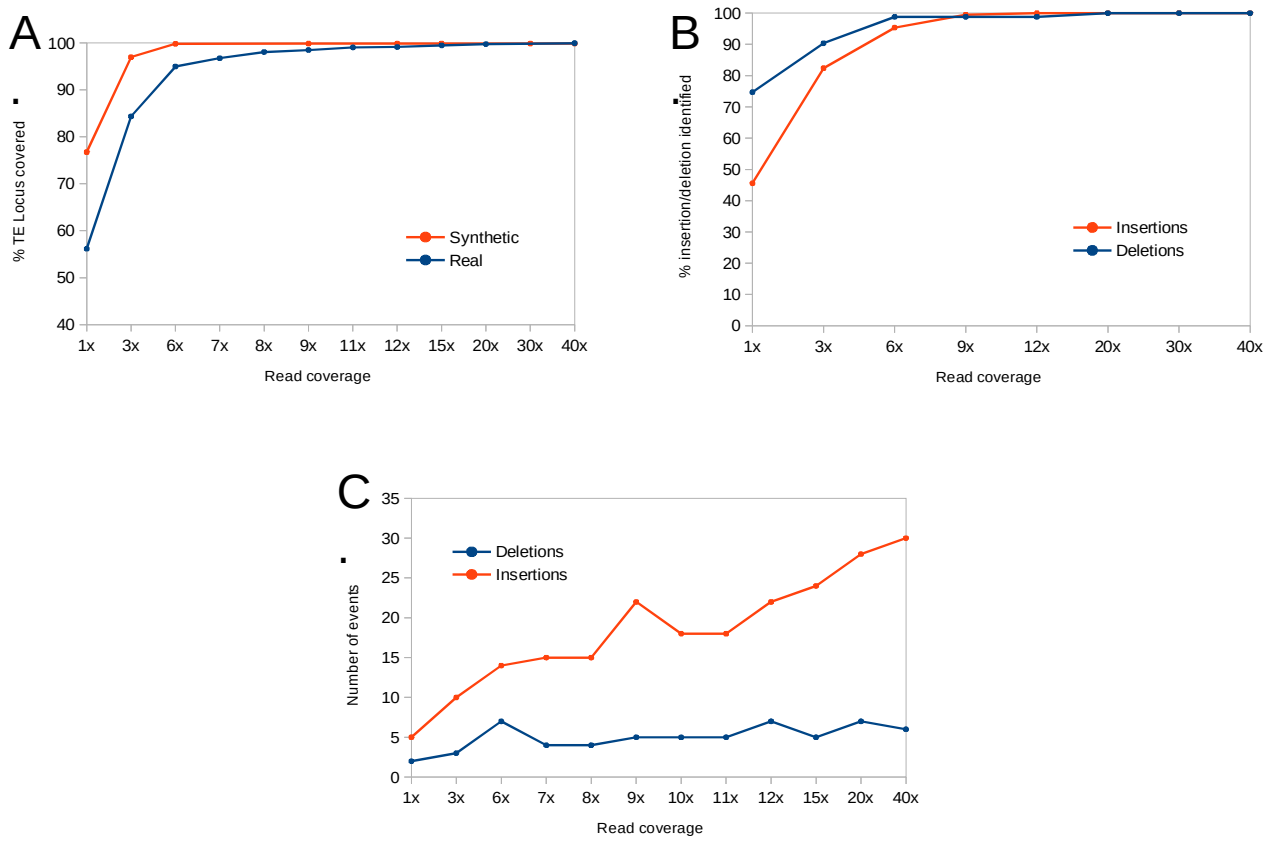

Figure 23. Performance de LoRTE en fonction de la quantité de read utilisée. A Pourcentage des TE annotés dans le génome de *D. melanogaster* qui ont été identifié par le programme. B Pourcentage d'insertions/délétions dans les lectures synthétiques identifiées. C Nombre de nouvelles délétions et insertions d'ETs trouvées dans les reads et absentes dans le génome de référence.

Publications associées a ce chapitre:

Siguier P, **Filée J**, Chandler M (2006) Insertion Sequences in Prokaryotic Genomes. *Curr. Opin. Microbiol.* 9(5):526-31

Filée J, Siguier P, Chandler M (2007) Insertion Sequence diversity in Archaea. *Microbiol Mol Biol Rev.* 71(1):121-57

Rouault JD, Casse N, Chénais B, Hua-Van A, **Filée J**, Capy P (2009) Automatic classification within families of transposable elements: application to the mariner Family. *Gene.* 15;448(2):227-32.

Hua-Van A, Le Rouzic A, Boutin T, **Filée J**, Pierre Capy P (2011) The struggle for life of the genome's selfish architects. *Biol. Direct.* 17;6:19

Metcalfe CJ, **Filée J**, Germon I, Joss J, Casane D. (2012) Mode and tempo in inflation of size of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1-like elements. *Mol. Biol. Evol.* 9(11):3529-39

Kamoun C, Payen T, Hua-Van A, **Filée J.** (2013) Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* 11;14:700

Filée J, Rouault JD, Harry M, Hua-Van A (2015) Mariner transposons are sailing in the genome of the blood-sucking bug Rhodnius prolixus. *BMC genomics.* 15;16(1):1061

Disdero E, **Filée J** (2017) LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*. 8 (1), 5

Bouallègue M, **Filée J**, Kharrat I, Mezghani-Khemakhem M, Rouault JD, Makni, Capy P (2017) Diversity and evolution of mariner-like elements in aphid genomes. *BMC Genomics* 18:494

Thomas–Bulle C, Piednoël M, Donnart T, **Filée J**, Jollivet D, Bonnivard E (2018) Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. *BMC Genomics* (15;19(1):821

Saint-Leandre B, Capy P, Hua-Van A, **Filée J** (2020). piRNA and Transposon Dynamics in Drosophila: A Female Story. *Genome Biol. Evol.* 12(6),2203-2207.

Filée J, Farhat S, Higuet D, Teysset L, Marie D, Thomas-Bulle C, Hourdez S, Jollivet D , Bonnivard E. (2021) Comparative genomic and transcriptomic analyses of transposable elements in polychaetous annelids highlight LTR retrotransposon diversity and evolution. *Mobile DNA* 12 (24).

4. Génomique évolutive de l'holobionte chez les insectes.

Ce dernier chapitre correspond chronologiquement a mes travaux les plus récents et constitue les voies de recherches principales que je compte explorer au cours des prochaines années. Ils peuvent se diviser en deux grandes thématiques l'une autour de la génomique évolutives des Triatomines et l'autre autour de la génomique des Mouches Soldats pour l'entomoculture.

a. Phylogénie des Triatomines du genre Rhodnius

Je participe a l'encadrement de Marie Merle avec Myriam Harry (EGCE a Gif) pour sa thèse sur l'évolution des génomes de *Rhodnius*, hémiptère vecteur de la maladie de Chagas en Amérique latine. Son premier travail a consisté a obtenir une phylogénie robuste du genre *Rhodnius* qui sera ensuite utilisée comme arbre de référence pour comprendre l'évolution des génomes du genre. Une importante partie de ce travail a été de rassembler et de séquencer une large collection d'espèce du genre *Rhodnius*, la plus représentative de la diversité du genre. Notre stratégie a consisté a séquencer le génome de 36 espèces (15 espèces sur 17 connues) avec un niveau modéré de couverture (20-40X) pour aboutir a un jeu de donnée mitochondrial (génome complet : 13 gènes), deux gènes ribosomaux nucléaires, et 51 gènes nucléaires. Ces données ont permis d'obtenir une phylogénie solide du genre, de comprendre les relations phylogénétiques entre les espèces est-Andines et ouest-Andine, de clarifier le statut taxonomique de plusieurs espèces débattues et d'une manière plus inattendue de relever plusieurs incongruences phylogénétiques entre gènes nucléaire et gènes mitochondriaux. Les test statistiques a l'échelle du génome entier de type ABBA/BABA montrent que ces incongruences sont très probablement le résultat d'introgression/hybridation. Par exemple si on prends les phylogénies du groupe *pictipes* (figure 24) qui inclut 3 espèces (*R. pictipes* sensus stricto, *R. stali* et *R. brethesi*), la position phylogénétique de l'espèce *R. stali* (populations StaWY et StaWZ) diffère entre gènes mitochondriaux et gènes nucléaires : groupe frère des populations de *R. brethesi* en mitochondrial mais groupe frère des populations de l'espèce *R. pictipes* en nucléaire. Une explication possible est l'introgression de la lignée *stali* par la lignée *brethesi* au niveau des gènes mitochondriaux (par exemple par hybridation avec une femelle de *R. brethesi*, les génomes mitochondriaux ayant une transmission essentiellement maternelle)

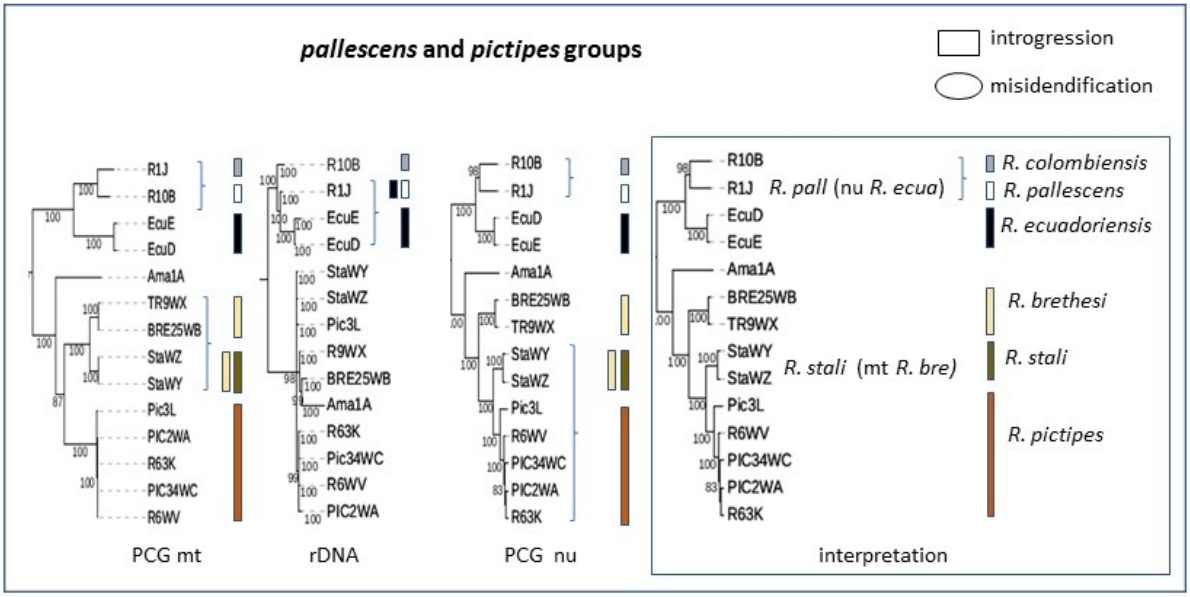

Figure 24 : Conflit de phylogénies chez les espèces du genre *Rhodnius*. A gauche la phylogénie mitochondriale (PCG mt), au centre phylogénie des gènes ribosomiques (rDNA), et a droite des gènes nucléaires (PCG nu). L'interprétation via des événements d'introgression est données dans l'encadré. Les rectangles de couleurs indiquent les différentes populations d'une même d'espèce.

b. *Evolution réductives des génomes du genre Rhodnius*

Au sein des Triatomines, les punaises du genre *Rhodnius* ont une singularité : les *Rhodnius* ont des tailles de génomes très inférieures. En effet, alors que la plupart des espèces de Triatomines ont des tailles de génomes comprises entre 1 et 1,5 Gb, nos mesures de taille des génomes en cytométrie de flux ainsi qu'en étudiant la distributions des *k-mers* des reads de séquençage génomique ont démontré que chez les *Rhodnius*, les génomes ont des tailles plus faibles, entre 600 et 700kb (Figure 25). De plus, chez certaines espèces comme chez *R. robustus* les génomes des femelles, quelque soit la population étudiée, étaient toujours 40 Mb plus réduits par rapport aux mâles.

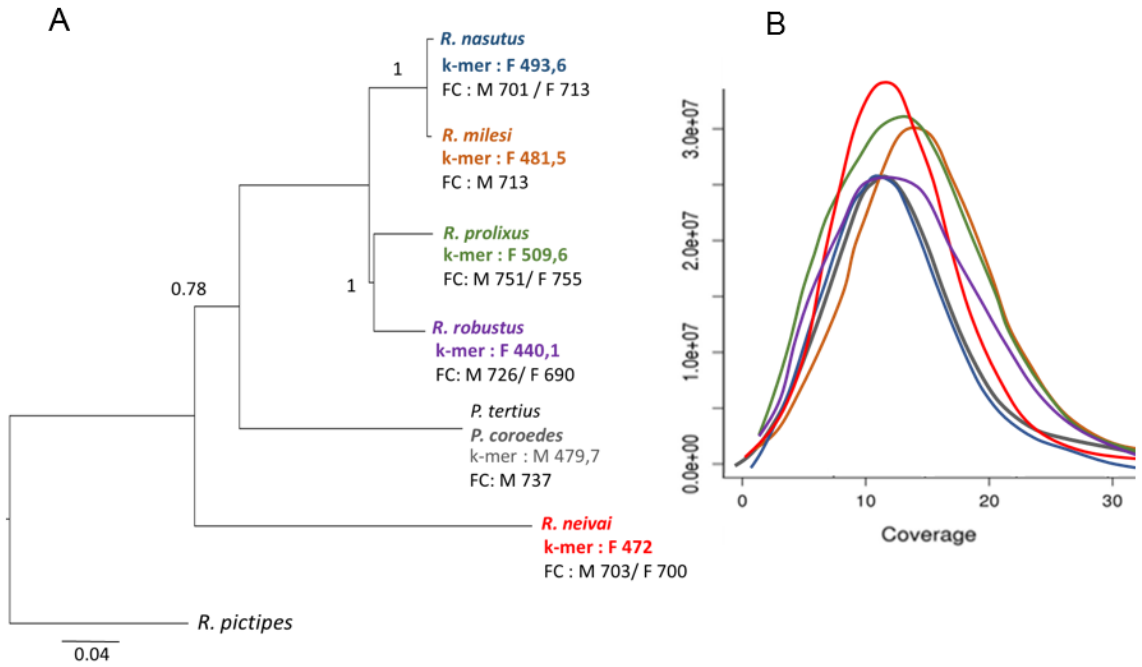

Figure 25: Estimation de la taille des génomes de différentes espèces du genre *Rhodnius* en utilisant la méthode des k-mers (A) et distribution des K-mer de taille 21 en fonction de la couverture (B).

Comment expliquer ces observations ?
Nous avons réalisé le séquençage hybride complet et l'assemblage génomique de 9 espèces de *Rhodnius* en combinant read court Illumina et reads long PacBio. Un des résultants marquant de ce travail repose sur l'étude du nombre de famille de gène orthologue au sein de ces espèces. En effet nos travaux indiquent une tendance globale a la diminution progressive des famille de gènes orthologues : le taux de perte de gène au cours du temps est 2,5 fois plus important que le taux de gain (Figure 26). Il en résulte une tendance globale des génomes de *Rhodnius* a la contraction. Nos données montrent aussi une relativement faible activité des éléments transposables, avec peu de familles actives et un niveau élevé de conservation des familles d'ET entre les génomes. Il en résulte un haut niveau de synténies et de conservation globale des génomes malgré une divergence temporelle assez ancienne entre les lignées que nous avons déterminé autour de 80 millions d'années.
L'ensemble de ces données apportent un éclairage original sur l'évolution des génomes de *Rhodnius* et converge vers l'idée que la petite taille des génomes de ces punaises au sein des Triatomines est le

résultat d'un processus global de réduction des génomes via des pertes de gènes et une faible activité des éléments transposables.

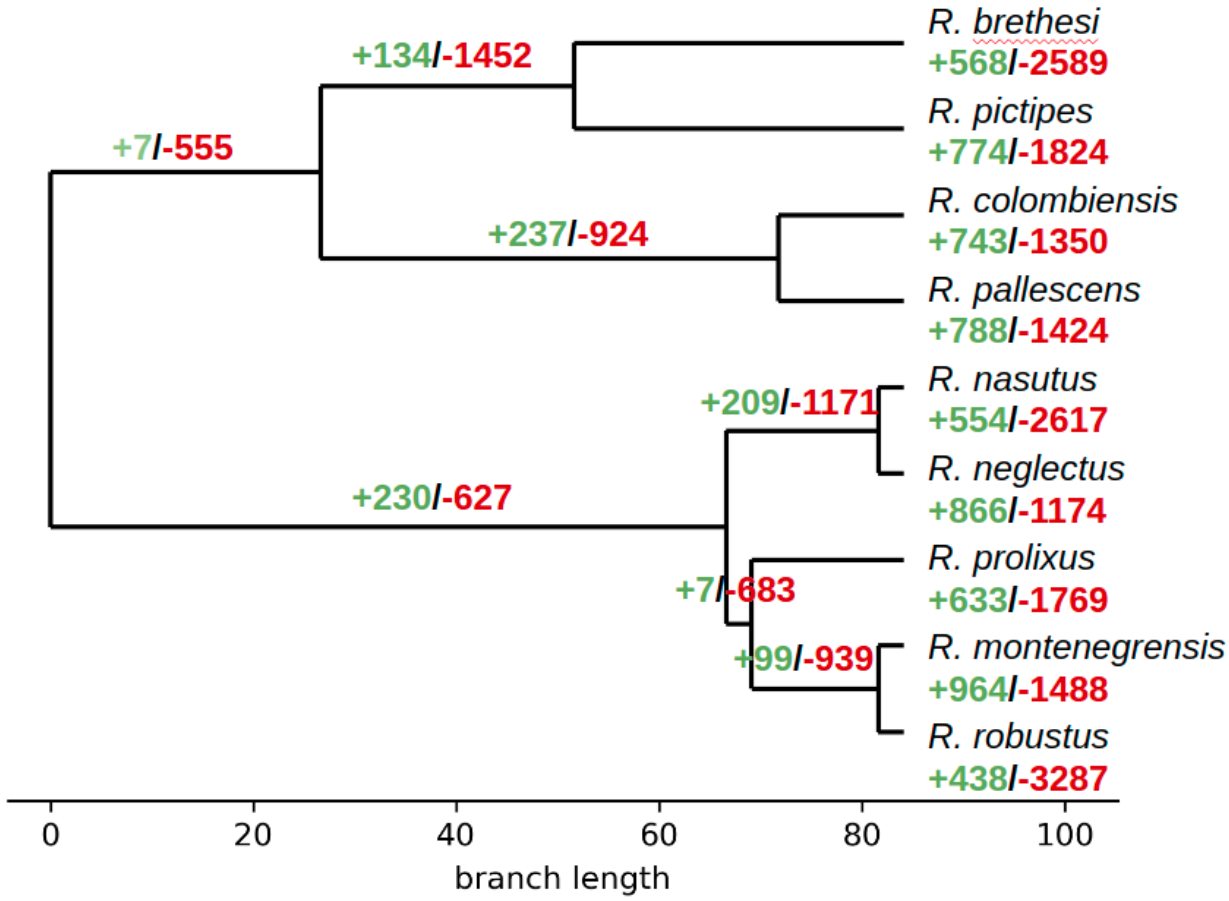

Figure 26: Evolution des familles de gènes au sein des génomes de *Rhodnius*. Sur chaque branche est indiqué le nombre de gain de gene (vert) et le nombre de perte de gènes (rouge). L'échelle indique le temps de divergence en millions d'années calculée avec un calibrage de données fossiles.

c. Evolution des symbioses nutritionnelles chez les Triatomines

Une grande diversité de bactérie symbiotique jouent un rôle clé dans l'adaptation des insectes hématophages a un régime alimentaire composé exclusivement de sang de vertébré. En effet, ce sang est trop pauvre en vitamine B pour permettre un développement des larves et/ou la survie des adultes. Les insectes hématophages ont donc développé des symbioses nutritionnelles avec des bactéries qui les approvisionnent en vitamine B. Par exemple, la punaise de lit *Cimex lectularius* dépends de symbionte du genre *Wolbachia* pour pouvoir se développer.
Les Triatomines du genre *Rhodnius* sont connues pour dépendre d'un symbiote intestinal du genre *Rhodococcus*. Toutefois, lorsque nous avons séquencé divers génomes de punaise du genre *Rhodnius*, une partie des contigs de plusieurs espèces avaient une très forte similarité avec des fragments génomiques de *Wolbachia* appartenant au groupe E et infectant la punaise de lit *Cimex* (Figure 27 a gauche). Par PCR en utilisant des amorces spécifiques de *Wolbachia*, nous avons de plus démontré que sur 120 individus appartenant a 17 espèces, 40% sont infectés par des *Wolbachia* (8 espèces). Les *Wolbachia* de *Rhodnius* ont donc une présence sporadique au sein du genre alors que les PCR avec des

amorces spécifique aux symbiotes intestinaux *Rhodococcus* ont montré une présence universelle chez toutes les populations.
Nous avons pu obtenir un assemblage métagénomique complet ou presque complet de 13 génomes de *Wolbachia* (Figure 27 a droite). Alors que les punaises du genre *Rhodnius* et du genre *Cimex* ont divergé depuis très longtemps (120Ma), leurs *Wolbachia* respectives sont très fortement apparentées . Leurs génomes sont d'ailleurs très semblables et presque complètement synténique avec seulement quelques délétions et mouvements de transposons (Figure 27 a droite).

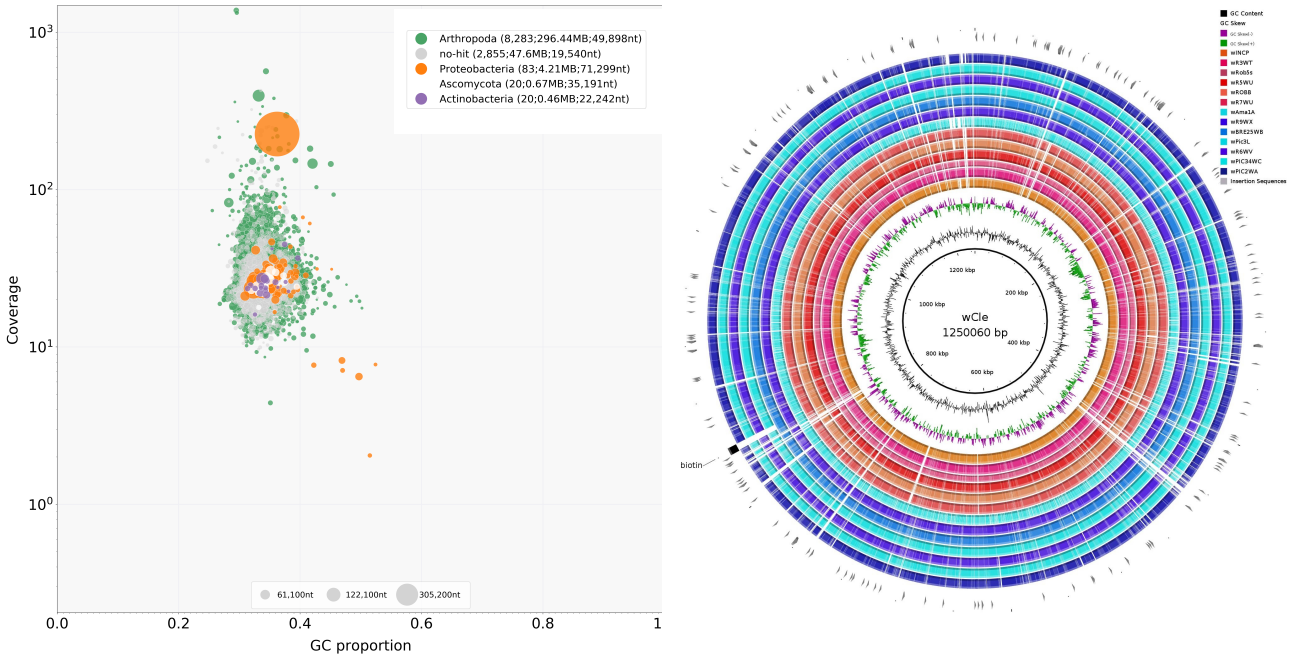

Figure 27: Blobplot des contigs du génome de *R. pictipes* (gauche) montrant la présence en sus de l'assemblage des contigs de *Rhodnius* (cercles verts) et d'un contig de *Wolbachia* (gros cercle orange en haut). La taille des cercles est proportionnelle a la taille des contigs. Et comparaison par alignement circulaire des génomes de *Wolbachia* infectant les espèces de *Rhodnius* par rapport a celui infectant la punaise de lit *Cimex lectularus* (droite). Les transposons sont indiqués par des flèches

La seule explication possible est ici un changement d'hôte, des *Wolbachia* infectant des *Rhodnius* ayant « sauté » chez *Cimex* (ou inversement). De plus, les génomes de *Wolbachia* infectant les *Rhodnius* possèdent presque tous un opéron complet de la biotine, qui est la voie métabolique emblématique de la symbiose nutritionnelle en vitamine B chez la punaise de lit *Cimex*. Seul une seule espèce de *Rhodnius* (3 populations) semble abriter des *Wolbachia* ayant perdu l'opéron biotine indiquant une probable perte de la capacité a entretenir une relation mutualiste avec les punaises (Figure 27 a droite). Nous avons de plus démontré que la quasi-totalité des génomes nucléaires de *Rhodnius* ont acquis latéralement des gènes de *Wolbachia* (jusqu'a plus de 200 kb) au cours du temps, incluant des populations où des infections actuelles n'ont pas pu être mise en évidence, suggérant une association universelle mais dynamique avec des pertes et des gains au cours du temps. L'ensemble de ces éléments indiquent que les punaise du genre *Rhodnius* maintiennent une symbiose tripartie avec d'une part des bactéries intestinales du genre *Rhodococcus* qui semblent être des symbiotes obligatoires transmis horizontalement et d'autre part avec des bactéries endocellulaires facultatives du genre *Wolbachia*, transmises horizontalement et potentiellement capable de suppléer a l'absence du symbionte obligatoire *Rhodococcus*. Ces découvertes qui constituent une histoire assez originale de saut d'hôte dans l'évolution de la symbiose nutritionnelle chez les punaises avec une symbiose tripartie dynamique impliquant du mutualisme obligatoire ou facultatif ainsi que des

réversions vers un stade parasite. A ce stade du travail, nous nous sommes demandé si la situation était identique chez le groupe frère des *Rhodnius* au sein des Triatomines, a savoir le genre *Triatoma*.
Ces travaux qui, sont en cours de finalisation, ont tout d'abord consisté a utilisé du métabarcoding utilisant l'ADNr 16S entier pour identifier les symbiotes intestinaux chez 7 espèces du complexe d'espèce *Triatoma (*64 individus*)*, en plus de différente d'espèce du genre *Rhodnius (*Figure 28) . Nos résultats confirment la prévalence des symbiotes du genre *Rhodococcus* chez les *Rhodnius* mais d'une manière surprenante, chez les espèces du complexe *Triatoma*, les symbiotes du genre *Rhodoccocus* sont très rares ou le plus souvent absent. Au seins des *Triatoma*, c'est un symbiote du genre *Arsenophonus* qui est le plus prévalent et le plus abondant. Les bactéries du genre *Arsenophonus* sont connues chez les insectes pour établir avec leurs hôtes divers type de relations qui évoluent du parasitisme a divers types de mutualisme incluant la symbiose nutritionnelle comme chez des Hémiptères appartenant aux groupes des Aleurodes.

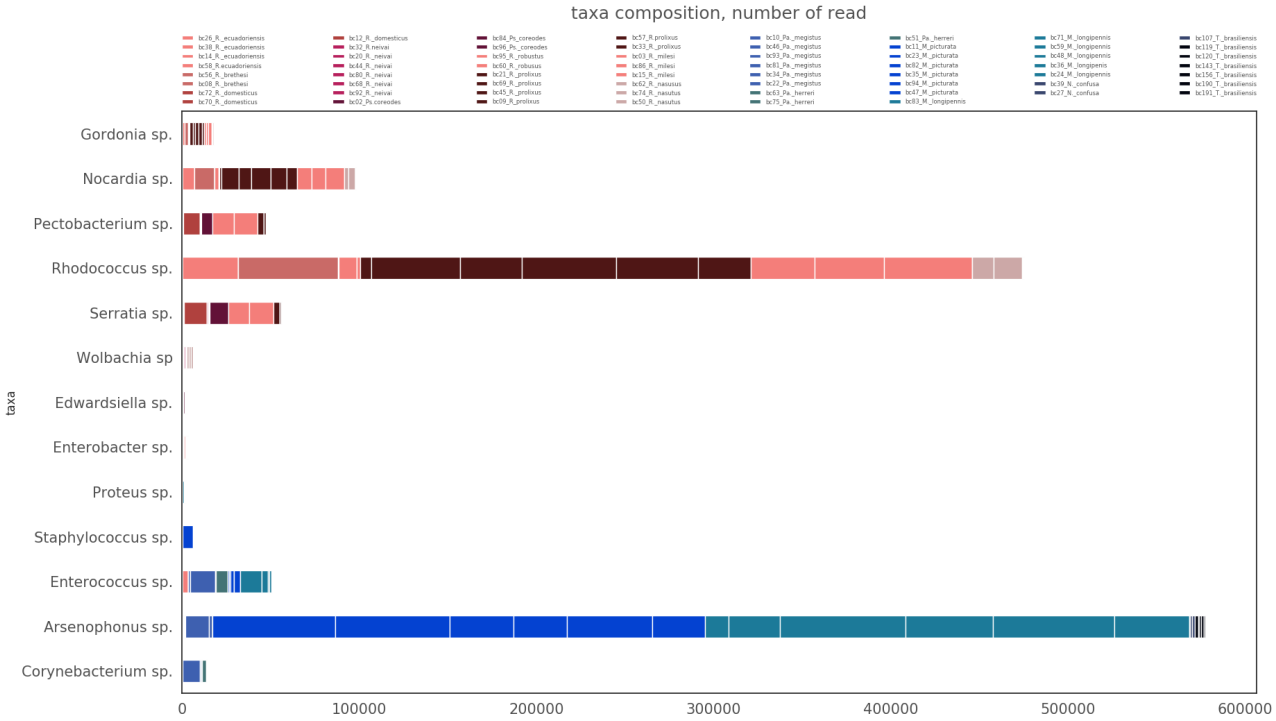

Figure 28: Nombre de reads brut mappant sur l'ADNr 16S de chaque espèces bactériennes identifiées. En rouge les espèces et individus du genre *Rhodnius*, en bleu les espèces et individus du complexe *Triatoma*

En séquençant le génome entier de l'espèce *Triatoma brasiliensis*, l'assemblage génomique nous as permis d'identifier la présence de plusieurs contigs que l'on peut assigner au (méta)génome entier d'*Arsenophonus (*3,8 Mb au total pour 20 contigs avec un N50 de 400 kb)(Figure 29 gauche). Et la comparaison génomique avec les autres génomes d'*Arsenophonus* infectant divers insectes montrent que son génome est très divergents des autres génomes connu, infirmant la possibilité de transferts récents entre espèces connues (Figure 29 droite). Nous comptons poursuivre ces travaux en étudiant plus en détails le génome de d'*Arsenophonus* chez *Triatoma* ainsi qu'en réalisant des phylogénies de l'ADNr 16S pour mieux comprendre le type de transmission avec des analyses de réconciliations avec l'arbre des espèces d'insectes.
Pris ensemble, ces travaux chez les symbiotes des Triatomines suggèrent une évolution non-linéaire avec des sauts d'hôtes et des probables remplacement fonctionnels entre les *Rhodnius* et les *Triatoma* par des bactéries appartenant a des groupes phylogénétiques très éloignés. Si l'ensemble de ces

prédictions sur des bases génétiques et génomiques mériteraient des validation fonctionnelles en curant les Triatomine de leurs symbiotes (par traitement antibiotique par exemple) pour observer l'effet sur la survie de ces bactéries, ces résultat ouvrent des perspectives en terme de santé publique pour lutter contre ces vecteurs, par exemple avec des symbiotes génétiquement modifiés.

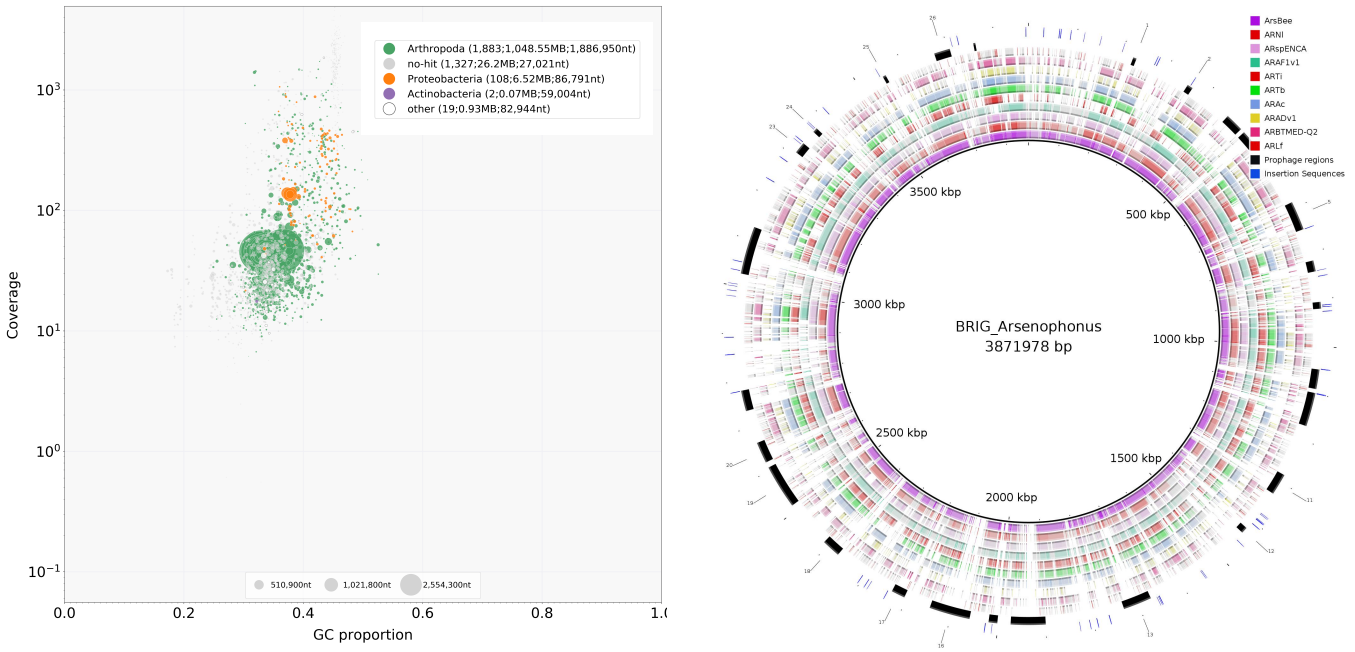

Figure 29: Blobplot et alignement circulaire génomiques des génomes d'*Arsenophonus*. A gauche, les spots oranges correspondent aux séquences d'*Arsenophonus*, en vert au génome de *Rhodnius*. A droite, les rectangles noirs indiquant les présences de pro-phages intégrés qui semblent jouer un rôle important dans les variations génomiques chez cette bactérie.

d. Histoire naturelle et génomique évolutive de la domestication de la mouche-soldat noire Hermetia illuscens

Cette collaboration avec l'entreprise CycleFarm basé dans la région d'Angers a pour objectif d'apporter une base de connaissance fondamentale sur la génétique de la mouche soldat *Hermetia illuscens*. Ce diptère cosmopolite et considéré comme invasif présente pourtant un intérêt agronomique et écologique important puisqu'elle est utilisée en élevage de masse afin de recycler des sous-produits de l'industrie agronomique (déchets en particulier) afin de générer des farines utilisées pour l'alimentation des poissons et des volailles. La filière a pris un essor tres important ces
dernières années avec l'industrialisation du processus de production et plusieurs PMEs Françaises ont avec succès développé des usines de productions dont l'objectif a terme est de génèrer plusieurs milliers de tonnes de farine d'insectes issues d'*Hermetia*. La récente ouverture du marché vers l'élevage de volailles et probablement bientôt celui des porcs vont considérablement accroître la demande et la filière commence a investir en R&D pour optimiser et mieux matiirseriliser sa production.
Le projet comporte plusieurs volets, les deux principaux étant la compréhension de l'histoire évolutive de l'espèce ainsi que l'étude de l'évolution des génomes de l'espèce. Un étudiant en thèse, Joseph Guilliet, dont j'assure l'encadrement, assure une partie de l'activité opérationnelle, il implique donc un partenaire industriel dans le cadre d'un contrat de collaboration incluant un financement CIFRE. Elle

implique aussi plusieurs chercheurs de l'EGCE a Gif (Frederic Marion-Poll, Nicolas Pollet, Dominique Joly) ainsi que de nombreux partenaires en Amérique du sud, Afrique et Asie pour la collecte des échantillons. L'objectif de la thèse de Joseph Guillet est de documenter et de comprendre la diversité génétique de l'espèce, d'étudier l'évolution des génomes de l'espèce et de commencer a établir des schémas et des méthodes de sélection en lien avec la domestication de l'espèce et les besoins de la filière.

Pour ce qui concerne le volet « histoire évolutive » nous avons procédé a un échantillonnage le plus large possible en France et dans le monde qui a aboutis a l'analyse de plus de 100 stations en combinant a la fois des données de cytochrome oxydase I sur plus de 400 individus (COI) et le séquençage complet du génome mitochondrial pour un nombre plus réduis d'individus (55 génomes complets). Les résultat principaux de ce travail font apparaître une très forte diversité génétiques cryptique de l'espèce qui nous as permis de définir au moins 10 haplotypes différents a la répartition cosmopolite. Une plus grande diversité d'haplotype se retrouve en Amérique du Sud, un résultat qui supporte l'idée que la mouche-soldat noire est originaire de ce continent avec au sein du continent américains un dégradé sud/nord de diversité haplotypique suggérant une colonisation naturelle (Figure 30).

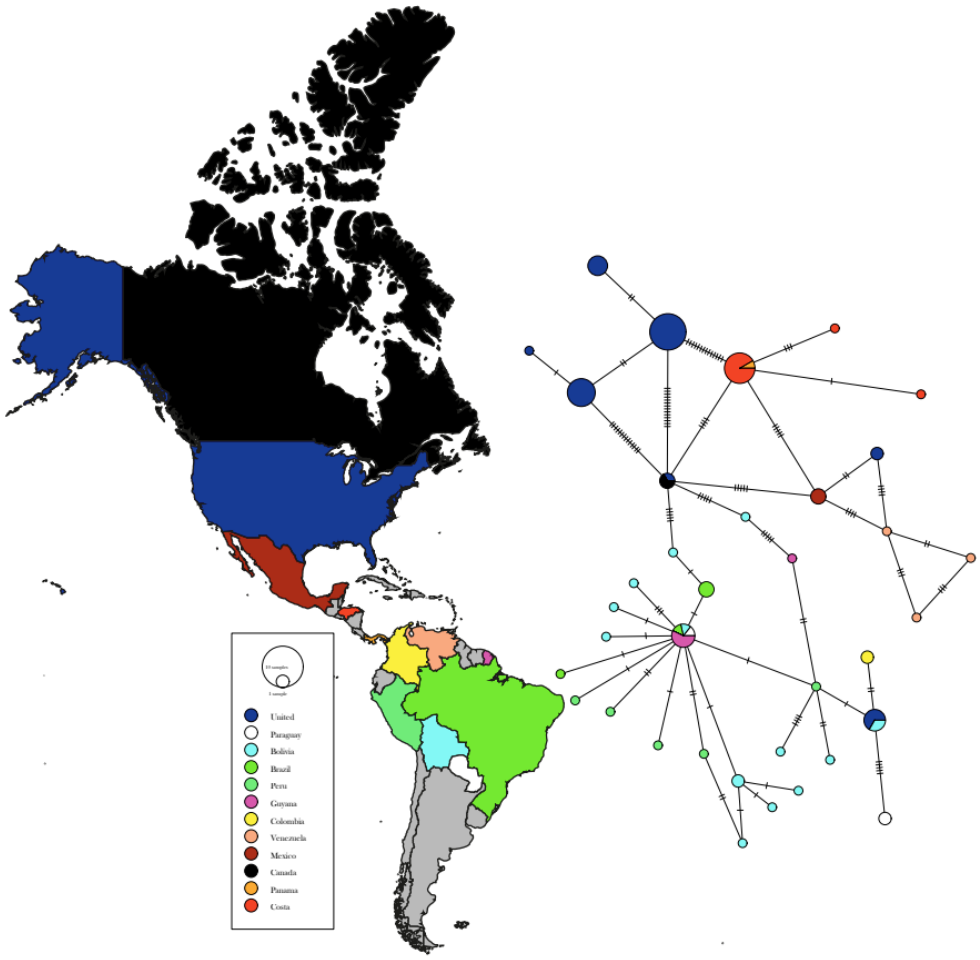

Figure 30: Gradient croissant sud/nord de la diversités haplotypique du gène COI chez la mouche soldat noire sur le continent Américain.

Néanmoins, a toutes les échelles géographiques (continent, pays et même localité) une grande diversité génétique est présente accréditant l'idée d'introductions multiples hors du continent Américains mais rendant difficile l'établissement d'un scénario de dispersion de l'espèce. Nos données génétiques indiquent que la totalité des souches d'élevages que nous avons étudiées dérive d'une seule et même

population d'origine nord-Américaine (groupe en rose sur la figure 31) avec des suspicions fortes de "fuite" de ces mouches dans l'environnement naturel immédiat des fermes d'élevages où elles auraient fondé des populations férales.

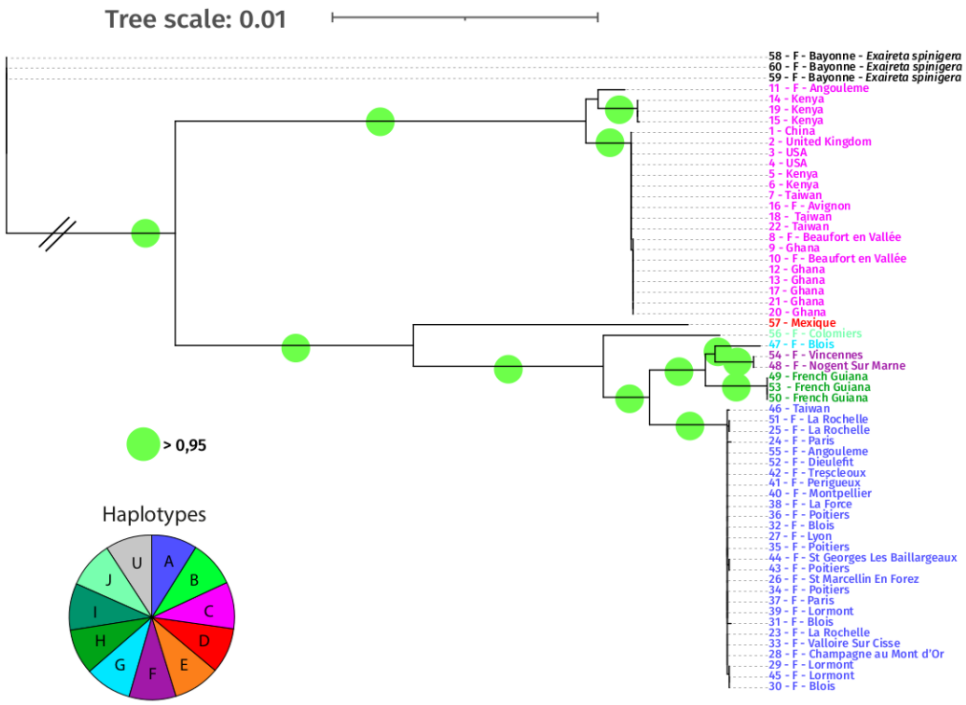

Figure 31: Phylogénie de la mouche soldat noire basée sur le génome mitochondrial complet

De plus, nos datations sur l'arbre mitochondrial indique un temps de divergences de 2 millions d'années entre le groupe en rose (lignée nors-Américaine et commerciale) sur la figure 28 et les autres groupes. Cette datation correspond a la fermeture de l'isthme de Panama ce qui accrédite l'hypothèse de la colonisation naturelle du continent nord-américain et de la progressive diversification de ce groupe par isolement avec les autres lignées au sud. Cette donnée ouvre la possibilité d'avoir en réalité un complexe d'espèces et non plus seulement une seule espèce.

Enfin a l'échelle de la France, nous avons collecté 4 haplotypes/groupes phylogénétiques différents et avons d'ors et déjà pu établir deux lignées d'élevage issu de parents sauvages. Nous avons de plus obtenus un génome complet et nous sommes en train de mener des travaux de génomiques comparative entre les génomes des différents groupes sur les compartiments connu pour évoluer rapidement. Ainsi nos données préliminaires sur les variations de taille des génomes indiquent des variations de tailles substantielles entre les différents groupes (entre 800Mb et 1Gb) causée par des niveaux d'amplifications très différents de quelques familles seulement de rétrotransposons LINE qui ont transposé a des rythmes très intense occupant jusqu'a 100Mb (plusieurs dizaines de milliers de copies par génome). Les conséquences génomiques et phénotypiques de ces insertions, si il y en a, sont inconnues et d'une manière générale, aucune information n'existe sur les caractéristiques zootechniques de ces lignées sauvages très divergentes d'*Hermetia.*

Nos futurs travaux sur l'espèce consisteront a la fois a domestiquer en élevage et a décrire ces lignées sauvages ainsi que d'apporter des informations en génomique comparative qui pourrait expliquer certains traits, caractéristiques particulières ou adaptations locales (par exemple a une climatologie nettement plus froide en hiver en Europe qu'en Amérique Latine, continent d'origine). Nous

prévoyons des études sur le microbiote de l'espèce et les éventuels parasites et maladies en utilisant des méthodes de barcoding ADNr 16S mais aussi de la métagénomique. L'ensemble de ces résultats décrits une diversité génétiques insoupçonnées de la mouche-soldat noire ce qui ouvre d'importante perspective a la fois appliquée (sélection) mais aussi plus fondamentale sur une espèce qui constitue un bon modèle d'étude sur la domestication des insectes.

Finalement, nos échantillonnages en France ont révélé que la mouche soldat noire vivait dans certain compost du sud de la France en sympatrie avec un autre stratyomide : la mouche soldat bleue (*Exaireta spinigera*). Très peu de chose sont connues sur cette espèce dont l'origine semble être l'Australie. Que se soit sa biologie, sa génétique/génomique et bien sur ces potentialités en entomoculture, tout reste a découvrir. Il y a pourtant un enjeu fondamental a diversifier les taxons et la diversité génétique en entomoculture pour éviter l'uniformisation et les multiples problèmes causés par la perte de diversité génétique en agronomie. L'étude multi-échelle de la mouche soldat bleue sera donc un de mes objectifs dans les prochaines années.

e. Autres travaux et projets en cours.

- Rôle des symbiotes du genre *Spiroplasma* chez les papillons du genre *Morpho*

Au cours d'un projet d'étude de génomique des populations de lépidoptère du genre *Morpho* en collaboration avec Violaine Laurens (MNHN-Paris) et Héloïse Bastide (EGCE- Gif) nous avons détecté la présence d'un symbiote du genre *Spiroplasma* dans un des 10 génomes (re)séquencés. Les bactéries du genre *Spiroplasma* sont connues comme étant des pathogènes de plantes et d'arthropodes et d'induire un distorsion du sex-ratio chez de plusieurs espèces d'insecte en tuant spécifiquement les embryons mâles dans les progénitures des individus infectés. Ce phénotype « tueur de mâles a été décris chez plusieurs espèces d'arthropode dont un lépidoptère. Chez la Drosophile, ce phénotype est lié a la production d'un toxine appelée Androcidin qui inhibe la machinerie de compensation du dosage de l'expression des gènes associés aux chromosomes X chez les embryons mâles. Chez les papillons du genre *Morpho*, le génome du Spiroplasme que nous avons identifié code pour une Androcidin très fortement apparentée a la toxine de Drosophile suggérant un transfert horizontal récent du gène et/ou un saut d'hôte. De plus, ce génome de *Spiroplasma* (4 Mb) est d'une taille inhabituelle, plus du double de tout les *Spiroplasma* séquencés a ce jour (1 a 2Mb). Nous comptons documenter la distribution et la phylogénie de ce Spiroplasme au seins des populations naturelles de *Morpho*, d'étudier son génome et d'essayer de mieux comprendre ses effets phénotypiques et notamment si il induits une distorsion du sex-ratio.

- Rôle des Eléments Transposables dans l'adaptation aux parasitismes chez l'huître plate.

Ce projet est une collaboration avec Eric Bonnivard et Arnaud Tanguy a la station biologique de Roscoff au seins du projet BivET (Génomique comparative et expression différentielle des éléments transposables de bivalves en lien avec l'adaptation et la réponse aux parasitoses.). L'huitre plate native d'Europe (*Ostrea edulis*) est un bivalve comestible dont la production avait été abandonnée suite aux pressions parasitaires de deux protistes parasites (*Marteilia refringens* et *Bonamia ostreae*). L'établissement récent par sélection de lignées d'huitre plate en partie résistante a ces parasites a ouvert une voie d'étude intéressante pour comprendre les déterminants génomiques de la résistance et d'une manière générale a mieux comprendre la parasitose et ses effets physiologiques en terme de stress. Dans ce projet, l'objectif est de mieux comprendre le rôle des ETs dans le processus d'adaptation en comparant les données génomiques et transcriptomiques des lignées résistantes et sensibles. Le projet inclus aussi l'obtention et l'étude des métagénomes des parasites.

- Evolution de l'inquilinisme chez les parasites d'hyménoptères.

L'inquilinisme est une forme de parasitisme où l'inquilin se sert du corps de son hôte comme abris physique. L'objectif de ce projet porté par Héloïse Bastide est de comprendre les mécanismes de l'adaptation convergente a l'inquilinisme convergente de deux mouches apparentée au genre Drosophila (*Braula sp.* et *Cacoxenus sp.*) a la vie parasitaire sur des abeilles sociales et solitaires. Il s'agit d'un travail de génomique comparative pour comprendre les déterminants génétiques et génomiques de l'adaptation par comparaison de plusieurs génomes des parasites récoltés sur différentes espèces d'abeilles. Je suis en particulier concerné dans ce projet par les compartiments ETs et les données métagénomiques sur les symbiotes.

Publications associées a ce chapitre:

1. **Filée J**,Agésilas-Lequeux K, Lacquehay L, Bérenger JM, Dupont L, Mendonça V, Aristeu da Rosa J, Harry M (2022) *Wolbachia genomics support a tripartite nutritional symbiosis in blood-sucking Triatomine bugs. BioRxiv.* 2022.09.06.506778

2. Legout H,Ogereau D**,Filée J**,Garnery L,Gilbert C,Requier F,Yassin A,Bastide H (2022) The genome of the bee louse fly reveals deep convergences in the evolution of social inquilinism. *BioRxiv.* 2022.11.08.515706

3. Guilliet J, Baudouin G, Pollet N, **Filée J** (2022) The natural history of the black soldier fly, *Hermetia illucens*: insights from complete mitochondrial genome sequences. *BMC Ecol Evo* 22,72

4. **Filée J**, Merle M, Bastide H, Mougel F, Beranger JM, Folly-Ramos E, Almeida EC, Harry M. (2022) Phylogenomics for Chagas disease vectors of the Rhodnius genus: what we learn from mito-nuclear conflicts and recommendations. *Front. Ecol. Evol.* 9:750317

5. **Filée J***, Merle M*, de Oliveira J, Almeida EC, Mougel F, Bastide H, Girondot M da Rosa JA, Harry M. (2022) Evidence for genome size variation in the Rhodniniі tribe of Chagas disease vectors. *Am. J. Trop. Med. Hyg*. 16;107(1):211-215
* : Authors contributed equally to the work

6. Hardwick KM, Bichang'a GB, Abtew AB, Awori RM, Cepko LCS, Chebon-Bore LJ, Darby A, […], **Filée J ,** […], Schaack S. (2020) Comprehensive transcriptome of the maize stalk borer, Busseola fusca, from multiple tissue types, developmental stages, and parasitoid wasp exposures. *Genome Biol. Evol.* 12 (12), 2554-2560.

7. Lima-Oliveira TM, Fontes FVHM, Lilioso M, Pires-Silva D, Teixeira MMG, Meza JGV, Harry M, **Filée J**, Costa J, Valença-Barbosa C, Folly-Ramos E, Almeida CE (2020) Molecular eco-epidemiology on the sympatric Chagas disease vectors Triatoma brasiliensis and Triatoma petrocchiae: Ecotopes, genetic variation, natural infection prevalence by trypanosomatids and parasite genotyping. *Acta Trop.* 201:105188.

8. Planes S, Allemand D, Agostini S, Banaigs B, Boissin E, Boss E, [...] **Filée J** [...] Zoccola D (2019) The *Tara* Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLoS Biol* 17(9).

9. Hardwick KM, Ojwang' AME, Stomeo F, Maina S, Bichang'a G, Calatayud PA, **Filée J**, Djikeng A, Miller C, Cepko L, Darby AC, Le Ru B, Schaack S (2019) Draft Genome of *Busseola fusca*, the Maize Stalk Borer, a Major Crop Pest in Sub-Saharan Africa. *Genome Biol. and Evol*. 11(8).

Conclusions & perspectives

Ces 20 années de recherche scientifique ont été marquées par les progrès du séquençage de l'ADN qui ont permis de révéler la diversité des interactions évolutives au sein des consortiums d'entité biologique composées de cellules eucaryotes, de divers procaryotes, de virus de toutes natures et d'une abondance encore plus grande d'éléments génétiques mobiles. Avant tout, ces entités collaborent entre elles, entretiennent parfois des relations conflictuelles, mais souvent aussi coopératives. C'est sans doute là un des points les plus saillants des travaux exposés ici. En effet, comment interpréter autrement ces échanges de gènes entre virus et leurs hôtes, comment expliquer autrement la coévolution entre transposons et génomes hôtes, comment comprendre les symbioses avec les bactéries intracellulaires? Souvent la distinction entre la compétition et la coopération est tenue, les deux processus n'étant pas exclusif, un symbiote pouvant a la fois procurer avantage sélectif a son hôte dans un contexte donné (par exemple si le symbiote nutritionnel est le seul a pouvoir procurer la vitamine en question) et par ailleurs avoir un coût en terme de fitness dans un autre contexte (le symbiote devenant parasite si un autre symbiote prends en charge la nutrition). Il est assez clair a mes yeux que nous commençons seulement a décrire la complexité de ces interactions au seins des consortiums biologiques. Voilà un des deux grandes directions fortes de mes travaux futurs, en particulier sur des modèles insectes qui ont le double avantage de composer un groupe doté d'une biodiversité naturelle phénoménale... et d'être largement étudié au sein du laboratoire.

La deuxième direction a été esquissée sur mes travaux en entomoculture. Je suis, pour des raisons qui tiennent a mes convictions qu'on pourrait qualifier de «citoyennes», particulièrement concerné par les crises globales du climat et de la biodiversité. J'ai longtemps cherché à comment être utile compte tenu de mes compétences, finalement pas si étendue dès que je les confronte a des problèmes de sociétés. Les insectes sont je crois un élément intéressant d'un processus de transition écologique réel qui ne tient pas d'une des entreprises de «green-washing» habituelles. Mon implication au sein de la filière depuis quelques années me font dire qu'il y a là un large spectre possible d'innovations en matière d'économie circulaire et de traitement des biodéchets. Les défis sont grands parce que la production actuelle est le fait d'acteurs qui ont une approche souvent naïve des insectes: tout reste a faire sur la domestication et la résilience des espèces cibles, les connaissances sur les pathogènes et la biosécurité sont quasi-nulles, même la biologie des deux espèces élevées en masse aujourd'hui est très mal connue etc... Il est assez évident pour moi que la génétique et la génomique sont des outils précieux pour essayer de répondre a ces questions et j'ai la volonté de continuer a pouvoir travailler dans cette direction en participant a structurer au sein du laboratoire un axe de travail centré sur la valorisation des insectes utiles pour la transition écologique.

References

1.Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265**, 687–695 (1977).

2.Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1–8 (2016).

3.Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

4.Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature News* **514**, 550 (2014).

5.Shendure, J., Findlay, G. M. & Snyder, M. W. Genomic Medicine–Progress, Pitfalls, and Promise. *Cell* **177**, 45–57 (2019).

6.Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences* **87**, 4576–4579 (1990).

7.Gilbert, J. A. & Dupont, C. L. Microbial metagenomics: beyond the genome. *Ann Rev Mar Sci* **3**, 347–371 (2011).

8.Guerrero, R., Margulis, L. & Berlanga, M. Symbiogenesis: the holobiont as a unit of evolution. *Int Microbiol* **16**, 133–143 (2013).

9.Scola, B. L. *et al.* A Giant Virus in Amoebae. *Science* **299**, 2033–2033 (2003).

10.Schulz, F. *et al.* Giant virus diversity and host interactions through global metagenomics. *Nature* **578**, 432–436 (2020).

11.Bevan, M. W. *et al.* Genomic innovation for crop improvement. *Nature* **543**, 346–354 (2017).

12.Biswas, S., Zhang, D. & Shi, J. CRISPR/Cas systems: opportunities and challenges for crop breeding. *Plant Cell Rep* **40**, 979–998 (2021).

13.Lazcano, A. & Miller, S. L. On the Origin of Metabolic Pathways. *J Mol Evol* **49**, 424–431 (1999).

14.Mushegian, A. Gene content of LUCA, the last universal common ancestor. *Frontiers in Bioscience-Landmark* **13**, 4657–4666 (2008).

15.López-García, P. & Moreira, D. The Syntrophy hypothesis for the origin of eukaryotes revisited. *Nat Microbiol* **5**, 655–667 (2020).

16.Spang, A. *et al.* Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* **521**, 173–179 (2015).

17.Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J. & Embley, T. M. Phylogenomics provides robust support for a two-domains tree of life. *Nat Ecol Evol* **4**, 138–147 (2020).

18.Cunha, V. D., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the

universal tree of life topology. *PLOS Genetics* **14**, e1007215 (2018).
19.Esser, C. *et al.* A Genome Phylogeny for Mitochondria Among α-Proteobacteria and a Predominantly Eubacterial Ancestry of Yeast Nuclear Genes. *Molecular Biology and Evolution* **21**, 1643–1660 (2004).
20.Martin, W. *et al.* Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences* **99**, 12246–12251 (2002).
21.Drew, G. C., Stevens, E. J. & King, K. C. Microbial evolution and transitions along the parasite–mutualist continuum. *Nat Rev Microbiol* **19**, 623–638 (2021).
22.Duron, O. & Gottlieb, Y. Convergence of Nutritional Symbioses in Obligate Blood Feeders. *Trends Parasitol* **36**, 816–825 (2020).
23.Nikoh, N. *et al.* Evolutionary origin of insect-Wolbachia nutritional mutualism. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 10257–10262 (2014).
24.Lynch, M. & Conery, J. S. The Origins of Genome Complexity. *Science* **302**, 1401–1404 (2003).
25.Sun, C. *et al.* LTR Retrotransposons Contribute to Genomic Gigantism in Plethodontid Salamanders. *Genome Biology and Evolution* **4**, 168–183 (2012).
26.Rogel, M. A. *et al.* Genomic basis of symbiovar mimosae in Rhizobium etli. *BMC Genomics* **15**, 575 (2014).
27.Hof, A. E. van't *et al.* The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105 (2016).
28.Lavialle, C. *et al.* Paleovirology of 'syncytins', retroviral env genes exapted for a role in placentation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**, 20120507 (2013).
29.Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089–1103 (2007).
30.Muller, S., Pandey, R. R. & Pillai, R. S. Les piARN forgent un système immunitaire pour le génome. *Med Sci (Paris)* **29**, 487–494 (2013).
31.Zhaxybayeva, O. & Doolittle, W. F. Lateral gene transfer. *Current Biology* **21**, R242–R246 (2011).
32.Nelson-Sathi, S. *et al.* Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* **517**, 77–80 (2015).
33.Tantoso, E. *et al.* To kill or to be killed: pangenome analysis of Escherichia coli strains reveals a tailocin specific for pandemic ST131. *BMC Biology* **20**, 146 (2022).
34.Brochier, C., Philippe, H. & Moreira, D. The evolutionary history of ribosomal protein RpS14:: horizontal gene transfer at the heart of the ribosome. *Trends in Genetics* **16**, 529–533 (2000).

35.Martin, W. F. Eukaryote lateral gene transfer is Lamarckian. *Nat Ecol Evol* **2**, 754–754 (2018).
36.Bapteste, E. & Huneman, P. Towards a Dynamic Interaction Network of Life to unify and expand the evolutionary theory. *BMC Biology* **16**, 56 (2018).
37.Dagan, T. & Martin, W. The tree of one percent. *Genome Biology* **7**, 118 (2006).
38.Xia, J. *et al.* Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**, 1693-1705.e17 (2021).
39.Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions | Science Advances. https://www.science.org/doi/10.1126/sciadv.aba0111.
40.Lang, B. F. *et al.* An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**, 493–497 (1997).
41.Dynes, J. L. & Firtel, R. A. Molecular complementation of a genetic marker in Dictyostelium using a genomic DNA library. *Proceedings of the National Academy of Sciences* **86**, 7966–7970 (1989).

Annexe

CV Jonathan Filée

Born in Charleroi, Belgium (20/12/1976)

Allée Louis de Villetain
Batiment A
91190 Gif sur Yvette
France
33 6 74 34 64 68
jonathan.filee@universite-paris-saclay.fr

Education :

2003 : PhD degree, Université Paris-Sud Orsay (France), Institut de Génétique et Microbiologie, Laboratoire de Biologie moléculaire des Extrémophiles.

PhD supervisor : Jaqueline Laurent.

Jury : P. Capy (president), M. Chandler, P. Lopez-Garcia, J. Laurent, D. Prangishvili

Subject : Phylogeny of viral protein involved in the replication and the metabolism of DNA.

2000 : DEA « Biodiversité : Génétique, Histoire et Mécanisme de l'Evolution », Université Paris XI Orsay (France).

Professional experience :

2008-Present : Permanent Researcher (CRCN CNRS) (EGCE, U. Paris-Saclay at Gif sur Yvette, France) : Evolutionary genomics of Viruses and Transposons.

2005-2007: 3-Year Post-Doctoral fellowship in the "Unité des Elements Génétiques Mobiles" (Mick Chandler : LMGM, U. Toulouse, France)

2004: 1-Year post-doctoral fellowship in the "Unité de Myovirologie" (Henry Krisch : LMGM, U. Toulouse, France)

2000: 6-months internship in the "Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles" (Patrick Forterre : IGM Orsay, U. Paris-Saclay, France):

1999: 7-months internship in the "Unité de Physiologie Microbienne" (Nicole Tandeau de Marsac : Institut Pasteur, France)

Editorial activities :

Ad hoc reviewer for a large variety of journals and institutions including *: Nature, Science, PNAS,*

Current Biology, Mol Biol Evol, Gen. Biol. Evol. , Gene, BMC Evol. Biol., Virology, NAR, Res. in Microbio., Intervirology, the National Scientific Foundation (NSF) etc...

Guest Editor 2016-2017*: PloS Genetics*

Academic Responsabilities :

Elected nember Comité National CNRS (2017-2021) (S29 – CID51 as nember of the board)

Main Recent Funding:

As PI:

2019-2021: Industrial Contract with the Cyclefarms company, *Étude de l'histoire évolutive d'une espèce d'insecte d'intérêt agronomique.* 165KE + CIFRE phD grant.

2018: IDEEV, *The microbiome of the Chagas Vector.* 12KE

2014-2015: APEGE-CNRS, *XenoMite* 20 KE

2014*:* IDEEV, *Etadapt.* 8KE

As participant :

2022-2023 : SAD/Brittany Region, BivET (PI : Eric Bonnivard)

2017-2020: FASEB-CNRS, *AIMMBRA* (PI: C. Almeida, M. Harry)

2017-2021: ANR *CoralGene* (PI: S. Planes)

2016-2017 : Labex BASC, *RADIANT* (PI: Myriam Harry)

2016 : IDEEV *PacBioTbra* (PI : M. Harry)

Recent Supervising :

phD student

2019-2023 : Marie Merle (25%), co-supervised with M. Harry and F. Mougel

2019-2022 : Joseph Guillet (100%)

2016-2019: Bastien Saint-Léandre (25%), co-supervised with P. Capy & A. Hua-Van

Master student

2021 : Lea Payen (M2, U. Paris Saclay) co-supervised with Myriam Harry

2020 :Layla Adil (M2, Sorbonne Université). Lea Payen (M1, U. Paris-Saclay) co-supervised with Clément Gilbert.

2019 : Marie Jeanne Selvam (M2, U. Paris Saclay)

2018: Emilie Aubin (M2,U. Versaille-Saint-Quentin)

2017: Kenny Agesilas-Lequeux (M2, U. Paris-Saclay), co-supervised with M. Harry

2017: Laurie Lacquehay (M2,U. Paris-Saclay)

2016: Eric Disdero (M2, U. Paris-Saclay)

2015: Leornado Lauriot (M1, U. Paris-Saclay)

2015: Eidji Bord (M1, U. Paris-Diderot)

2014: Raphaelle Trouslard (M2, U. Paris-Saclay)

2013: Choumouss Kamoun (M1, U. Paris-Diderot)

Publications: h index = 22 , 2769 citations (GS)

Publications with M2 Master and phD students have been underligned

Preprints:

1. **Filée J**,Agésilas-Lequeux K, Lacquehay L, Bérenger JM, Dupont L, Mendonça V, Aristeu da Rosa J, Harry M (2022) *Wolbachia* genomics support a tripartite nutritional symbiosis in blood-sucking Triatomine bugs. *BioRxiv.* 2022.09.06.506778

2. Legout H,Ogereau D**,Filée J**,Garnery L,Gilbert C,Requier F,Yassin A,Bastide H (2022) The genome of the bee louse fly reveals deep convergences in the evolution of social inquilinism. *BioRxiv.* 2022.11.08.515706

3. **Filee J***, Becker HJ*, Mellottee L, Zhihui L, Lambry JC, Liebl U, Myllykallio H (2022) Bacterial origin of thymydilate and folate metabolism in Asgard Archea. *BioRxiv*
   * : Authors contributed equally to the work

Journal Articles:

1. Guilliet J, Baudouin G, Pollet N, **Filée J** (2022) What complete mitochondrial genomes tell us about the evolutionary history of the black soldier fly, *Hermetia illucens*. *BMC Ecol. Evol.* 1;22(1):72

2. **Filée J**, Merle M, Bastide H, Mougel F, Beranger JM, Folly-Ramos E, Almeida EC, Harry M. (2022) Phylogenomics for Chagas disease vectors of the Rhodnius genus: what we learn from mito-nuclear conflicts and recommendations. *Front. Ecol. Evol.* 9:750317

3. **Filée J***, Merle M*, de Oliveira J, Almeida EC, Mougel F, Bastide H, Girondot M da Rosa JA, Harry M. (2022) Evidence for genome size variation in the Rhodninii tribe of Chagas disease vectors. *Am. J. Trop. Med. Hyg*. 107(1):211-215
   * : Authors contributed equally to the work

4. **Filée J**, Farhat S, Higuet D, Teysset L, Marie D, Thomas-Bulle C, Hourdez S, Jollivet D , Bonnivard E. (2021) Comparative genomic and transcriptomic analyses of transposable elements in polychaetous annelids highlight LTR retrotransposon diversity and evolution. *Mobile DNA* 12 (24).

5. Hardwick KM, Bichang'a GB, Abtew AB, Awori RM, Cepko LCS, Chebon-Bore LJ, Darby A, […], **Filée J ,** […], Schaack S. (2020) Comprehensive transcriptome of the maize stalk borer, Busseola fusca, from multiple tissue types, developmental stages, and parasitoid wasp exposures. *Genome Biol. Evol.* 12 (12), 2554-2560.

6. Saint-Leandre B, Capy P, Hua-Van A, **Filée J** (2020). piRNA and Transposon Dynamics in Drosophila: A Female Story. *Genome Biol. Evol.* 12(6),2203-2207.

7. Lima-Oliveira TM, Fontes FVHM, Lilioso M, Pires-Silva D, Teixeira MMG, Meza JGV, Harry M, **Filée J**, Costa J, Valença-Barbosa C, Folly-Ramos E, Almeida CE (2020) Molecular eco-epidemiology on the sympatric Chagas disease vectors Triatoma brasiliensis and Triatoma petrocchiae: Ecotopes, genetic variation, natural infection prevalence by trypanosomatids and parasite genotyping. *Acta Trop.* 201:105188.

8. Planes S, Allemand D, Agostini S, Banaigs B, Boissin E, Boss E, [...] **Filée J** [...] Zoccola D (2019) The *Tara* Pacific expedition—A pan-ecosystemic approach of the "-omics" complexity of coral reef holobionts across the Pacific Ocean. *PLoS Biol* 17(9).

9. Hardwick KM, Ojwang' AME, Stomeo F, Maina S, Bichang'a G, Calatayud PA, **Filée J**, Djikeng A, Miller C, Cepko L, Darby AC, Le Ru B, Schaack S (2019) Draft Genome of *Busseola fusca*, the Maize Stalk Borer, a Major Crop Pest in Sub-Saharan Africa. *Genome Biol. and Evol*. 11(8).

10. Hardwick KM, Ojwang' AME, Stomeo F, Maina S, Bichang'a G, Calatayud PA, **Filée J**, Djikeng A, Miller C, Cepko L, Darby AC, Le Ru B, Schaack S (2019) Draft Genome of *Busseola fusca*, the Maize Stalk Borer, a Major Crop Pest in Sub-Saharan Africa. *Genome Biol. and Evol*. 11(8).

11. Thomas–Bulle C, Piednoël M, Donnart T, **Filée J**, Jollivet D, Bonnivard E (2018) Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. *BMC Genomics* (15;19(1):821

12. **Filée J** (2018) Giant viruses and their mobile genetic elements: the molecular symbiosis hypothesis. *Curr Opin Virol.* 33, 81

13. Bouallègue M, **Filée J**, Kharrat I, Mezghani-Khemakhem M, Rouault JD, Makni, Capy P (2017) Diversity and evolution of mariner-like elements in aphid genomes. *BMC Genomics* 18:494

14. Disdero E, **Filée J** (2017) LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. *Mobile DNA*. 8 (1), 5

15. **Filée J**, Rouault JD, Harry M, Hua-Van A (2015) Mariner transposons are sailing in the genome of the blood-sucking bug Rhodnius prolixus. *BMC genomics.* 15;16(1):1061

16. **Filée J** (2015). Genomic comparison of closely related Giant Viruses supports an accordion-like model of evolution. *Front Microbiol.* 2015 Jun 16;6:593

17. **Filée J** (2014). Multiple occurrences of giant virus core genes acquired by eukaryotic genomes: the visible part of the iceberg? *Virology.* 2014 Oct;466-467:53-9

18. Kamoun C, Payen T, Hua-Van A, **Filée J.** (2013) Improving prokaryotic transposable elements identification using a combination of de novo and profile HMM methods. *BMC Genomics* 11;14:700

19. **Filée J.** (2013) Route of NCLDV evolution: the genomic accordion. *Curr Opin Virol.* 3(5):595-9

20. Metcalfe CJ, **Filée J**, Germon I, Joss J, Casane D. (2012) Mode and tempo in inflation of size of the Australian lungfish (Neoceratodus forsteri) genome: a major role for CR1-like elements. *Mol. Biol. Evol.* 9(11):3529-39

21. Hua-Van A, Le Rouzic A, Boutin T, **Filée J**, Pierre Capy P (2011) The struggle for life of the genome's selfish architects. *Biol. Direct*. 17;6:19

22. **Filée J**, Chandler M. (2010) Gene exchange and the origin of giant viruses. *Intervirology.* 2;53(5):354-61

23. Rouault JD, Casse N, Chénais B, Hua-Van A, **Filée J**, Capy P (2009) Automatic classification within families of transposable elements: application to the mariner Family. *Gene.* 15;448(2):227-32.

24. **Filée J** (2009) Lateral gene transfer, lineage-specific gene expansion and the evolution of Nucleo Cytoplasmic Large DNA viruses. *J Invertebr Pathol.* 101(3):169-71

25. **Filée J**, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic Large DNA Viruses. *BMC Evol Biol.* 26;8:320.

26. **Filée J**, Chandler M (2008) Convergent Mechanisms of Genome Evolution of Large and Giant DNA Viruses. *Res. Microbiol.* 159(5):325-31

27. **Filée J**, Siguier P, Chandler M (2007) Insertion Sequence diversity in Archaea. *Microbiol Mol Biol Rev. 7*1(1):121-57

28. **Filée J**, Siguier P, Chandler M (2007) I am what I eat and I eat what I am: acquisition of bacterial genes by Giant Viruses. *Trends Genet.* 23(1):10-5

29. Siguier P, **Filée J**, Chandler M (2006) Insertion Sequences in Prokaryotic Genomes. *Curr. Opin. Microbiol.* 9(5):526-31

30. **Filée J**, Bapteste E, Susko E, Krisch HM (2006) A Selective Barrier to Horizontal Gene Transfer in the T4-Type Bacteriophages that Has Preserved a Core Genome with the Viral Replication and Structural Genes. *Mol. Biol. Evol.* 23(9):1688-1696

31. **Filée J**, Comeau AM, Suttle CA, Krisch HM (2006) T4-type bacteriophages. *Med Sci.* 22(2):111-2.

32. **Filée J**, Tetart F, Suttle CA, Krisch HM (2005) Marine T4-type bacteriophages, a ubiquitous component of the dark Matter of the biosphere. *Proc.Natl.Acad.Sci.USA.* 102(35):12471-6

33. **Filée J**, Forterre P (2005) Viral proteins functioning in organelles: a cryptic origin? *Trends Microbiol*. 13(11):510-3

34. **Filée J**, Forterre P, Laurent J (2003) The role played by viruses on the Evolution of their cellular host: a view on informational proteins phylogenies. *Res. Microbiol.* 154:237-43

35. Myllykallio H, Leduc D, **Filée J**, Liebl U (2003) Life without dihydrofolate reductase FolA. *Trends Microbiol.* 11(5):220-3

36. Gadelle D*, **Filée J***, Buhler C, Forterre P (2003) Phylogenomics of type II DNA topoisomerases. *BioEssay* 25: 232-242.

* : Authors contributed equally to the work

37. Myllykallio H, Lipowsky G, Leduc D, **Filée J**, Forterre P, Liebl U (2002) An alternative flavin-dependent mechanism for thymidylate synthesis. *Science* 5578:105-7.

38. **Filée J**, Forterre P, Sen-Lin T, Laurent J (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J Mol Evol* 54:763-73.

Books:

- Francez-Charlot A, **Filee J**, Castanie-Cornet MP, Cam K (2004) Regulation of flhDC by the His-Asp phosphorelay RcsCDB. In “Global Regulatory Networks in Enteric Bacteria” ed. B. Press.

- Forterre P, **Filee J**, Myllykalio H (2003) Origin and evolution of DNA and DNA replication machineries. In "The Genetic Code and the Origin of Life" L. Ribas, ed. Landes Bioscience.

université PARIS-SACLAY

Titre: L'apport de la génomique comparative dans la compréhension des interactions évolutives au sein des holobiontes

Mots clés: Génétique et génomique évolutive, virus, transposon, symbiose

Résumé: Ce travail de recherche est centré sur la compréhension des interactions évolutives entre ce qu'il est convenu d'appeler un «holobionte», c'est a dire au sein d'un consortium d'entités biologiques tel qu'un hôte eucaryotes avec ces symbiotes, ces virus et les autres éléments génétiques mobiles. A travers une grande diversité d'organismes, ces travaux décryptent les relations de compétitions et de coopérations a l'œuvre au seins de ces consortiums. Une partie importante de ce manuscrit concerne l'évolution et l'origine des virus avec un important corpus de données sur l'influence réciproque des échanges latéraux de gènes entre les virus et leurs hôtes, notamment le rôle-clé des virus dans l'évolution de l'appareil de réplication des mitochondries et des chloroplastes. Plus généralement nos travaux éclairent la place centrale jouée par les remplacements non-homologues des gènes de la réplication et du métabolisme de l'ADN avec de notables conséquences pour l'origine des eucaryotes. L'étude des éléments transposables, de leurs influences sur la structure et l'évolution des génomes hôtes ainsi que de leurs régulations sera aussi détaillé. Enfin, l'importance des symbioses chez les insectes constitue un élément important de ce travail avec nottament des travaux sur l'évolution des symbioses nutritionelles chez les Triatomines vecteur de la maladie de Chagas. Ce travail débouche enfin sur des travaux plus appliqués sur la génétique et le génomique de diverses espèces d'insectes, dont la domestication de la mouche soldat noire pour la production de farine d'insecte.

Title : Understanding the evolutionary interactions in the holobionts through the contributions of comparative genomics.

Keywords : Evolutionary genetic and genomics, virus, transposon, symbiosis

Abstract : This work aims to better understand the evolutionary interactions ruling the «holobionts», that is, the consortia composed of an eukaryotic host and their symbionts, viruses and other mobile genetic elements. Through the study of a large variety of biological models, this work unravels many events of cooperation rather than competition between viruses and their hosts with the description of numerous events of lateral gene exchanges. Specifically, we document the role of virus in the evolution of the DNA replication apparatus, showing for instance that mitochondria and chloroplasts use viral genes to replicate their DNA. We provide many arguments that point out the key-role of non-homologous gene replacements during the evolution of the genes involved in the metabolism and the replication of the DNA with significant consequences on the origin of the eukaryotes. An important part of this manuscript is devoted to the study of the structural role play by the transposable elements during their host genome evolution. We describe many examples of the interplay between self-disseminating transposons and the regulation process implemented by the cellular genomes to control the bursts. We also describe several microbial symbioses in insects with a specific emphasis on the evolution of the nutritional symbiosis in Triatomine, the main vector of the Chagas disease. We finally conclude this work by more applied developments on insect genomics, with recent studies on the genetic and genomic of the black soldier fly for entomoculture and meal production.