Développement et applications d'outils d'analyse métagénomique des communautés microbiennes associées aux insectes

Cervin Guyomar

UNIVERSITE ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 600
Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation
Spécialité : *Génétique, génomique et bioinformatique*

Par
Cervin GUYOMAR

Développement et applications d'outils d'analyse métagénomique des communautés microbiennes associées aux insectes

Thèse présentée et soutenue à Rennes, le 7 Décembre 2018
Unité de recherche : IGEPP et IRISA

Rapportrices avant soutenance :

Laurence MOUTON	Maitre de conférences, Université Lyon 1
Hélène CHIAPELLO	Ingénieure de recherche, INRA

Composition du Jury :

Président :	Philippe VANDENKOORNHUYSE	Professeur, Université de Rennes 1
Examinateur·rice·s :	Didier BOUCHON	Professeur, Université de Poitiers
	Hélène CHIAPELLO	Ingénieure de recherche, INRA
	Laurence MOUTON	Maitre de conférences, Université de Lyon 1
	Éric PELLETIER	Directeur de recherche, CEA
Dir. de thèse :	Jean-Christophe SIMON	Directeur de recherche, INRA
Co-dir. de thèse :	Claire LEMAITRE	Chargée de recherche, Inria
Co-dir. de thèse :	Christophe MOUGEL	Directeur de recherche, INRA

Remerciements

Dans les derniers moments de la rédaction de ma thèse, c'est le moment de faire une courte pause, et d'adresser petits et grands merci à ceux qui ont croisé mon chemin pendant ces trois ans. Je vais donc dire merci à beaucoup de gens, et en oublier beaucoup également, milles excuses par avance.

Merci d'abord à ceux qui se sont autoproclamés ma *dream team* d'encadrants en début de thèse. Beaucoup diraient que 4 encadrants, ca fait beaucoup pour une seule thèse, mais je remercie chacun d'entre eux pour ce qu'ils m'ont appris et leur contribution à ce travail. Je pense avoir pas mal changé pendant ces trois ans, et vous n'y êtes pas pour rien.

Merci à Claire, pour beaucoup de choses, à commencer par sa confiance, ses conseils, et le temps passé à relire cette thèse alors qu'elle aussi a été bien occupée ces derniers mois.

Merci à Jean-Christophe qui s'est efforcé pendant 3 ans de faire de moi un spécialiste du puceron (et ce n'était pas facile !). Merci pour ces relectures tardives, et cette belle escapade dans le Negev dont je me souviendrai.

Merci à Fabrice pour ses questions et projets pleins d'imagination. Ça a été un plaisir de travailler avec toi.

Merci enfin à Christophe pour sa vision et ses conseils avisés, qui ont toujours été les bienvenus pour guider ma thèse.

Merci ensuite à l'ensemble du jury, qui a accepté d'évaluer mon travail, et en particulier aux deux rapportrices qui vont s'atteler à la lecture de ce manuscrit.

Je ne sais pas exactement comment j'en suis arrivé là, mais voilà quelques personnes grâce à qui mon avis sur la thèse est passé de "Mais pour quoi faire" à "Je commence Lundi prochain". Merci à Ian Clark qui m'a vu faire mes premiers pas maladroits de bioinformaticien, tout en me faisant prendre conscience de ce qu'est l'humour anglais. Merci à David Causeur et Franck Picard qui m'ont donné la bonne idée de faire une thèse et soutenu dans ce projet. Merci enfin à Dominique Lavenier, qui a reçu un ingénieur innocent, et lui a fait assez confiance pour lui donner l'opportunité de rentrer dans son équipe. Merci aussi pour ton aide bienvenue durant ces derniers mois de thèse.

Merci ensuite à tous les symbiotes des équipes GenScale, Dyliss et GenOuest, pour l'ambiance inégalée qui y règne, et que je risque d'avoir du mal à retrouver. La tradition veut qu'on y "vient pour se faire des collègues, et qu'on repart avec des amis", et je n'y dérogerai pas. Quelques mentions spéciales (non exhaustives, et parfois redondantes) pour :

Camille d'abord pour quelques mois passés dans un bureau "pépinière" de Symbiose, et pour être la co-inventrice du traditionnel du bain de mer du séminaire au vert.

Marie et Joseph ensuite, pour leurs prénoms qui vont si bien ensemble évidemment, mais aussi pour 2 ans passés dans le même bureau, dont je n'ai JAMAIS eu l'occasion de confondre la porte avec celle de la photocopieuse. Mon sens de la décoration a été ébranlé pendant ces deux ans, mais j'ai tenu le choc grâce à votre bonne humeur permanente. Merci surtout de m'avoir laissé râler en toutes circonstances (et je continue même ici !).

Les 4C, une équipe de choc pour un C-minaire organisé à la perfection.

Clémence, avec qui j'ai co-réalisé la plus belle contribution de ma thèse en première année, ainsi qu'à toutes les personnes ayant prêté leurs talents d'acteurs ou d'artiste à ce projet. J'en

profite pour coller un lien, quelques vues supplémentaires sont toujours bonnes à prendre : `https://youtu.be/3DRgLLITKUc`. Merci aussi d'avoir partagé mon sort de doctorant en phase finale. Ton aide m'a été précieuse et je ne te remercierai jamais assez ! Et non, je ne te souhaite pas une bonne année cette fois ci.

Wesley, qui a fait semblant de me prendre pour un maitre de stage crédible, et les étudiants qui on fait semblant de me prendre pour un enseignant crédible.

L'équipe dite "des mangeurs de graines", aussi connue sous le nom de "gang des punks à chat".

Les dévouées relectrices, traqueuses de fautes et pourfendeuses de répétitions : Chloé, Clémence et Stéphanie. Merci pour ces fautes que j'aurais pu relire mille fois sans apercevoir.

Au tour de quelques autres symbiotes, cités sans autre raison particulière que leur sympathie et leur humour et qui ont rendu pauses et missions si agréables. Je citerai Chloé, Matéo et Renaud, spécialistes des pauses interminables sur des sujets improbables. Les doctorants (docteurs maintenant !) de la Triforce dont le nombre de private jokes dépasse l'entendement (mais j'ai bien aimé celles que j'ai comprises). Clémence qui a le bon goût d'avoir un sens de l'humour très similaire au mien (et pour une certaine pile d'assiettes). Anthony pour son soutien psychologique constant quand à ma condition de doctorant.

Tous ceux qui font que fare une thèse n'est pas (si) difficile : Marie pour son aide et sa gentillesse de tous les instants, l'équipe GenOuest pour maintenir le cluster, en particulier Joseph, qui m'a appris que c'était une bonne idée d'avoir un *admin sys* dans son bureau.

Dépenser de l'énergie pour me déplacer d'un point A à un point A par différents moyens a sans doute été la chose la plus importante après la présente thèse (quand même) pendant 3 ans, alors voilà les remerciements obligatoires aux sportifs et collègues avec qui j'ai couru/nagé/roulé. Merci pour ces pauses bienvenues dans mes journées. Mention spéciale aux nombreux coureurs du midi avec qui j'ai partagé un bon paquet de foulées. Bon courage à vous pour les échéances futures.

Les anciens de l'agro. En particulier le club des doctorants rennais qui inclut Pierre (félicitations pour ta thèse !), Florence (bon courage pour la tienne !) et Océane. Et Solène Duaut, à qui j'avais promis qu'elle serait dans mes remerciements si je faisais une thèse, voilà c'est fait.

Merci à ma famille, et surtout à mes parents. Ca fait un petit moment que vous ne savez plus trop ce que je fais, ni pourquoi je le fais, mais je bénéficie toujours de votre soutien sans faille, et ça me fait très plaisir.

Merci à Anaëlle, pour m'avoir supporté avec douceur, patience (il en a fallu !) et encouragements pendant ces 3 ans.

Enfin merci à toi, lecteur qui ne se reconnait dans aucune des personnes listées plus haut. Je suis content que tu viennes jeter un coup d'œil à ma thèse, j'espère que ça te plaira et que tu ne comptes pas t'arrêter aux remerciements ;)

Table des matières

Résumé

Ces travaux de thèse reposent sur le double objectif de proposer des approches innovantes pour l'étude des relations entre un hôte et son microbiote, et de les appliquer à la description fine de l'holobionte du puceron du pois par des données métagénomiques. Les relations symbiotiques façonnent le fonctionnement et l'évolution de tous les organismes, mais restent décrites de manière imparfaite, notamment à cause de la difficulté à caractériser la diversité génomique des partenaires microbiens constitutifs des holobiontes. L'essor des technologies de séquençage métagénomique révolutionne l'étude de ces systèmes, mais pose également des problèmes méthodologiques pour analyser les jeux de données métagénomiques. La métagénomique est ici appliquée au puceron du pois, un modèle d'étude des relations symbiotiques qui abrite une communauté bactérienne d'une complexité modérée, idéale pour développer de nouvelles approches de caractérisation de la diversité microbienne. Cette thèse vise à mieux décrire la communauté de symbiotes qu'abrite cet holobionte, notamment en distinguant les différents niveaux de variabilité génomique en son sein. Nous présentons une démarche pour l'analyse métagénomique d'holobiontes, qui repose d'abord sur l'assignation taxonomiques des lectures par alignement à des génomes de référence ou préalablement assemblés, puis sur la recherche de variants génomiques. L'étude de génotypes complets de symbiotes permet de retracer l'histoire évolutive des relations hôte-microbiote avec une résolution élevée. Chez le puceron du pois, nous mettons en évidence des niveaux et structures de diversité génomique différents selon les symbiotes, que nous proposons d'expliquer par les modalités de transmission ou l'histoire évolutive propre à chacun des partenaires microbiens. Cette approche repose sur la disponibilité d'un génome de référence adapté, qui est souvent difficile à obtenir en métagénomique. Dans un second temps, nous présentons donc une méthode d'assemblage guidé par référence en contexte métagénomique. Cette méthode se déroule en deux temps : le recrutement et l'assemblage de lectures par alignement sur un génome de référence distant, puis l'assemblage *de novo* ciblé des régions manquantes, permis par des développements complémentaires apportés au logiciel *MindTheGap*. Comparativement à un assembleur métagénomique, cette méthode permet l'assemblage d'un seul génome en un temps réduit, et permet de détecter d'éventuels variants structuraux sur le génome ciblé. Appliqué au puceron du pois, *MindTheGap* a réalisé l'assemblage du symbiote obligatoire *Buchnera* en un seul contig, et a permis d'identifier différents variants structuraux du bactériophage APSE. Ces travaux ouvrent la voie à la fois à une caractérisation plus précise des relations hôte-microbiote chez le puceron du pois par des approches fonctionnelles et métaboliques, ainsi qu'à l'application des outils présentés à des systèmes plus complexes.

Chapitre 1

Introduction

Les microorganismes sont une forme de vie dont l'omniprésence et le rôle considérable ont été constamment réévalués à la hausse avec le progrès des technologies permettant leur caractérisation biologique et fonctionnelle. L'une des découvertes les plus récentes à leur sujet est le fait qu'ils sont impliqués dans la plupart des fonctions biologiques des organismes qui les hébergent. Ainsi, dans un corps humain, les bactéries sont dix fois plus nombreuses que les cellules humaines, et elles impactent profondément son métabolisme. L'objectif de cette thèse a été d'explorer les possibilités des technologies de séquençage métagénomique pour décrire les relations entre un organisme et les microorganismes qu'il abrite. Cette partie introductive vise à présenter successivement la nature de ces relations, les modalités des technologies permettant de les explorer, et le modèle d'étude utilisé au cours de cette thèse.

1.1 Caractéristiques des associations symbiotiques

1.1.1 De la symbiose à l'holobionte

Étymologiquement, le terme symbiose provient du grec "symbiôsis" qui signifie "vivre ensemble". Le terme est utilisé pour la première fois en 1877 par Albert Bernhard Frank pour décrire l'interaction entre des algues et des champignons formant des lichens, puis généralisé par Anton de Bary [De Bary, 1879]. Initialement, le terme ne pose pas d'hypothèse sur la nature de l'association entre ces deux espèces. Il regroupe donc à la fois les associations favorables aux deux partenaires (mutualisme), favorables à l'un et sans effet sur l'autre (commensalisme), et favorables à l'un et délétère pour le second (parasitisme). Toutefois, une certaine confusion règne sur cette définition, avec par exemple un emploi plus restrictif de ce terme pour décrire uniquement les relations mutualistes [Martin and Schwab, 2012]. Le consensus actuel sur ce sujet est que la frontière peut être mince entre ces trois modes d'association. Le coût net pour l'hôte d'héberger un symbiote peut aisément varier, entre autres en fonction de conditions environnementales [Sachs and Simms, 2006, Douglas, 2008]. On préfère alors parler d'un continuum parasitisme-commensalisme-mutualisme, qui donne tout son sens à une définition élargie de la symbiose, et c'est cette définition qui semble finalement s'imposer dans la communauté scientifique. On retiendra que ce débat est révélateur à la fois de l'importance du concept de symbiose, du large éventail d'interactions qu'il regroupe, et du poids de l'environnement sur le signe des associations.

Au fil des travaux scientifiques, ce qui apparaissait comme une exception devient une

règle dans le vivant. Ainsi, l'immense majorité des organismes est l'hôte d'organismes plus petits appelés symbiotes. Les premiers travaux décrivent un certain nombre de bactéries pathogènes présentes dans le corps humain. Ces communautés sont reconnues dans leur intégralité par le terme de microflore, progressivement remplacé par celui de microbiote. Avec l'introduction du terme de microflore, puis de microbiote, on reconnait l'étendue et la diversité de ces communautés symbiotiques, qui désigne l'ensemble des microorganismes associés à une plante ou à un animal. Ce concept est reconnu progressivement comme une composante fondamentale du fonctionnement de son hôte. Les avancées technologiques du séquençage permettent progressivement d'accéder au microbiome, soit le contenu génomique d'un microbiote, et permettent de répondre à certaines questions sur le rôle de ces communautés symbiotiques (défini par exemple par [Marchesi and Ravel, 2015]). En 2010, la revue *Nature* publie les premiers résultats du *Human Microbiome Project*, qui vise à décrire les principaux microbiomes humains. La couverture affiche comme titre "Our other genome", en référence à l'importance considérable du microbiome dans le fonctionnement d'un organisme humain [Zhao, 2010]. Ainsi, le seul microbiome intestinal contiendrait 150 fois plus de gènes que le génome humain [Zhu et al., 2010], et pourrait être impliqué dans des maladies telles que le diabète ou l'obésité [Cho and Blaser, 2012].

Ce constat amène à reconsidérer certains éléments du fonctionnement et de l'évolution des organismes. Ainsi, l'acquisition de symbiotes dans une population d'hôtes à des effets comparables ou supérieurs à l'apparition d'un nouvel allèle sur le génome de l'hôte. Le concept du phénotype étendu, défendu par Richard Dawkins, propose que le phénotype d'un organisme puisse avoir des effets sur son environnement [Dawkins, 1999]. Dans le cadre de relations symbiotiques, l'environnement du symbiote est l'organisme de son hôte. Suivant ce principe, l'explication classique du phénotype comme le produit du génotype et de l'environnement prend une autre mesure : l'environnement, et donc le phénotype de l'hôte, est directement influencé par le génotype de ses symbiotes.

Dans le prolongement de cette idée, Lynn Margulis définit pour la première fois le concept d'holobionte en 1991 [Margulis et al., 1991]. Ce terme désigne l'ensemble constitué par un hôte eucaryote et ses symbiotes microbiens. Cette idée remet en question la notion même d'individualité. Un individu n'est pas seulement défini par son génome, mais également par son microbiome. Rosenberg propose en complément le concept d'hologénome [Rosenberg and Zilber-Rosenberg, 2011], qui regroupe les génomes d'un holobionte et les place au sein d'une même unité de sélection, conférant ainsi à l'holobionte des propriétés proches de celles d'un "super-organisme". Ce concept redéfinit la théorie synthétique de l'évolution, en ne considérant pas comme critère de sélection la valeur adaptive d'un organisme, mais bien celle d'un holobionte. Si l'existence et la pertinence du concept d'holobionte ne sont pas remises en question, la place de l'hologénome comme unité de sélection est cependant sujette à débats [Douglas and Werren, 2016, Moran and Sloan, 2015].

L'importance des relations entre eucaryotes et microorganismes ne fait plus aucun doute, mais la nature de ces relations reste méconnue. Trois points en particulier sont à étudier lors de l'étude d'un holobionte :

1. Quels sont les microorganismes impliqués dans ces relations, et comment se structure leur diversité ?
2. Quel est le lien fonctionnel qu'ils partagent avec leur hôte ?
3. Quelle est l'histoire évolutive de ces associations symbiotiques ?

1.1.2 Qui sont-ils ? : Diversité multi-échelle des microbiomes

La multiplication des études portant sur différents holobiontes révèle une grande diversité dans le contenu de ces communautés. Tout d'abord, les associations symbiotiques n'épargnent aucune part du vivant. D'une part, tous les macroorganismes observés jusqu'à ce jour sont associés à des microorganismes. De l'autre, toutes les branches du vivant contiennent des microorganismes symbiotiques, et en particulier une grande diversité de taxons bactériens.

Un être humain est porteur d'un microbiote, qui contient plusieurs centaines de taxons bactériens. Les microbiotes de différents humains sont très variables entre eux, et contiennent chacun une grande diversité génomique [Huttenhower et al., 2012]. Ils évoluent également en fonction de facteurs endogènes comme exogènes. À cela s'ajoute une variabilité inter-individuelle : des individus différents sont porteurs de microbiotes différents [Arumugam et al., 2011]. En comparant la similarité des microbiotes avec la proximité phylogénétique de plusieurs espèces, on se rend compte qu'ils sont fréquemment corrélés. On désigne sous le nom de phylosymbiose cette corrélation entre similarité microbienne et phylogénétique [Brooks et al., 2016].

1.1.3 Que font-ils ?

Les effets fonctionnels les plus évidents des relations hôte-symbiotes sont les pathologies causées par des microbes pathogènes, dont les exemples ne manquent pas. Les manifestations des relations mutualistes avec des microbes ont quant à elles été décrites bien plus récemment. L'article de [Rosenberg and Zilber-Rosenberg, 2018] dresse un inventaire de ces bénéfices.

Une première fonction à laquelle répondent les microbes mutualistes est trophique : les symbiotes peuvent fournir à leur hôte des nutriments indisponibles dans son régime alimentaire. C'est notamment le cas de nombreux insectes dont le régime alimentaire peut être peu varié, comme les fourmis, la mouche *tsé-tsé* ou les pucerons [Zientz et al., 2004]. Le régime alimentaire de ces insectes ne leur permet pas d'obtenir certains acides aminés ou vitamines indispensables, que des symbiotes produisent pour eux. Dans le cas des lichens ou des mycorhizes des plantes, des organismes hétérotrophes s'associent à des organismes autotrophes pour se fournir en énergie. Chez l'humain, si le microbiote est associé à des maladies telles que le diabète ou l'obésité [Carding et al., 2015], mais est aussi impliqué dans des modifications métaboliques importantes durant la grossesse [Koren et al., 2012].

Le microbiote a également à rôle dans la défense immunitaire des organismes. Ainsi, le microbiote intestinal est impliqué dans la défense contre certains pathogènes, par exemple

par la production d'antibiotiques [Donia et al., 2014]. Chez les insectes, on peut également citer des symbiotes dont la présence est associée à des niveaux accrus de résistance à des parasitoïdes [Oliver et al., 2005].

On peut également mentionner d'autres conséquences du microbiote sur la vie de son hôte, comme des effets comportementaux [O'Mahony et al., 2015], sur le développement des organismes [Sommer and Bäckhed, 2013] ou la production de chaleur [Rosenberg and Zilber-Rosenberg, 2016].

1.1.4 Comment évoluent-ils ?

Si les relations symbiotiques, et plus précisément mutualistes, sont omniprésentes dans le vivant, il convient de s'interroger sur les modalités de cette réussite évolutive, et les forces qui agissent sur ces interactions. Dans cette partie, nous nous intéressons aux modalités de la persistance de ces relations symbiotiques, et à la manière dont elles façonnent les différents partenaires.

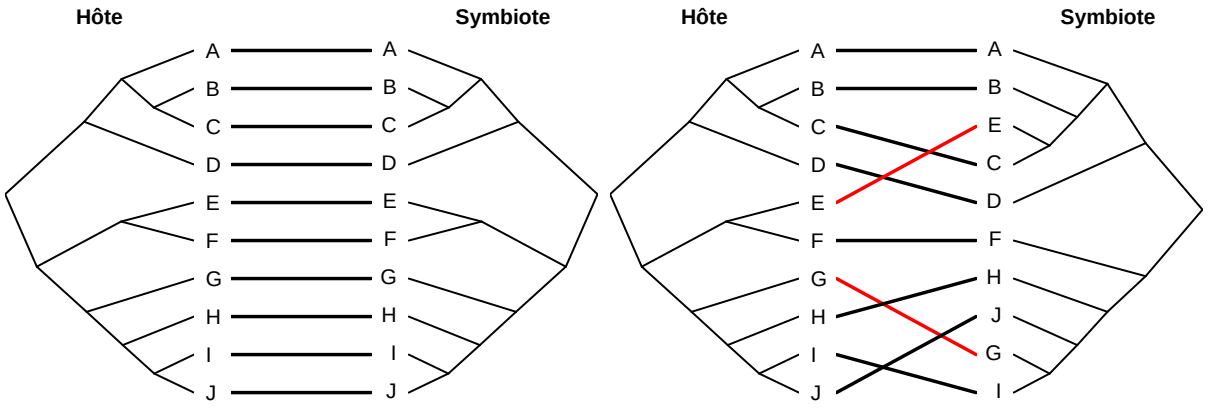
Modes de transfert des symbiotes

La façon dont les symbiotes persistent dans leur hôte au fil des générations est un élément central de leur évolution réciproque. Deux principaux modes d'association symbiotique peuvent être distingués en fonction de ce mécanisme. Ils sont détaillés dans la revue [Bright and Bulgheresi, 2010].

Dans le premier cas, l'acquisition symbiotique se fait de manière horizontale. L'hôte vit de manière aposymbiotique dans les premiers stades de son développement, puis est colonisé par différentes bactéries environnementales, capables de vivre sous forme libre. Ce mode d'acquisition nécessite une série de mécanismes présents à la fois chez l'hôte et le symbiote qui permettent l'acquisition, la sélection et la coopération entre ces entités. Le transfert vertical correspond lui à la transmission de microbes par l'un des parents. Dans ce cas, l'hôte ne connait pas ou peu de phase asymbiotique, et l'association entre hôte et symbiotes est durable. Le nombre de bactéries transmises à la descendance peut être faible. Ce phénomène a d'importantes conséquences évolutives et est qualifié de bottleneck ou goulot d'étranglement. Entre ces deux archétypes existent une multitude d'intermédiaires. Des symbiotes majoritairement hérités verticalement peuvent par exemple ponctuellement être transférés horizontalement entre individus [Moran and Dunbar, 2006], voire espèces différentes [Sandström et al., 2001]. De même, chez l'humain, le microbiote est partiellement hérité mais peut subir de profonds changements en fonction des conditions environnementales [David et al., 2014].

Déterminer le mode de transfert de symbiotes s'avère difficile. Une première manière de le faire consiste à observer les mécanismes de la transmission symbiotique [Moran and Dunbar, 2006]. Quand il est difficile d'observer les manifestations de ces transferts, une alternative est de recourir à des études phylogénétiques. L'hypothèse sous-jacente est que la transmission symbiotique verticale se traduit par une codiversification de l'hôte et de son symbiote, qui se traduit par la congruence des phylogénies des deux organismes. Des incongruences entre ces

Figure 1 – *Relation entre cophylogénie et modes de transmission symbiotique.* *La stricte congruence entre la phylogénie de l'hôte et du symbiote (à gauche) peut traduire un transfert exclusivement vertical du symbiote. À droite, des incongruences entre les deux phylogénies traduisent des événements de transfert horizontal.*



deux phylogénies sont un indice d'un changement d'hôte ou de perte du symbiote. À l'inverse, les holobiontes dont l'acquisition de symbiotes se fait de manière horizontale ne devraient pas montrer cette particularité. Dans certain cas, la co-diversification peut cependant être observée en l'absence de transfert vertical, notamment à cause de l'évolution des systèmes de sélection symbiotique, et de différences environnementales. On peut notamment observer ce cas de figure chez le calamar et la bactérie *Vibrio fischeri* [Nishiguchi et al., 1998].

Symbioses obligatoires et facultatives

On distingue couramment des symbiotes obligatoires (ou primaires) et des symbiotes facultatifs (ou secondaires). Les premiers sont indispensables à la survie de leurs hôtes, et sont donc retrouvés systématiquement en leur sein. Les seconds au contraire ne sont pas nécessaires à la survie de leur hôtes, mais peuvent néanmoins leur apporter certains avantages selon les conditions environnementales. Leur abondance dans une population d'hôtes est variable, et peut être vue comme analogue à la fréquence allélique d'un gène. Cette fréquence intermédiaire résulte de différentes forces écologiques et évolutives. Des effets de sélection permettent à un équilibre de s'établir en fonction du coût que représente l'hébergement du symbiote par l'hôte, et des bénéfices qu'il retire de cette association. Les holobiontes peuvent être sujets à des transferts, qui sont source d'innovation dans leur niche écologique, et sont soumis à la sélection naturelle qui mène à des fréquences symbiotiques propres à un milieu donné. Par conséquent, la fréquence de ces symbiotes facultatifs peut varier dans le temps et l'espace en fonction des conditions environnementales, et peut être étudiée sous l'angle de la génétique des populations, en considérant la présence ou l'absence d'un microbe comme un allèle de son hôte.

Histoire évolutive de la symbiose

De la bactérie environnementale au symbiote obligatoire

Si les relations parasitaires, symbiotiques facultatives et obligatoires semblent très différentes, elles résultent généralement d'une même histoire. Ces différents types de relations s'inscrivent en réalité dans un continuum entre bactéries environnementales et symbiotes obligatoires. Face à une infection par un parasite, une réponse fréquemment apportée par les hôtes est une domestication, menant à l'acquisition d'un symbiote secondaire [Moya et al., 2008]. Les symbiotes transmis verticalement entretiennent une relation particulière avec leur hôte. En effet, par plusieurs aspects, le transfert vertical permet l'alignement des intérêts évolutifs de l'hôte et de ses symbiotes mutualistes [Herre et al., 1999]. Tout d'abord, la reproduction d'un symbiote nécessitant celle de son hôte, la sélection va favoriser les symbiotes augmentant la valeur sélective de leur hôte. Par ailleurs, en réduisant la diversité génomique au sein de la population symbiotique par *bottleneck*, le transfert vertical limite également la compétition intra-spécifique pour le symbiote, qui peut être néfaste à l'hôte, et contribue à l'établissement d'une inter-dépendance entre le symbiote et son hôte. Après des générations de transfert vertical et de vie sous forme symbiotique, un microorganisme perd certaines fonctions devenues inutiles, ainsi que la capacité à vivre sous forme libre. À ce stade, l'intégration fonctionnelle entre l'hôte et le symbiote est complète, et on assiste à l'émergence de symbioses intra-cellulaires, ou endosymbioses, et de structures dédiées telles que les bactériocytes chez les insectes. Le stade ultime de cette association entre un hôte et un microbe est la théorie endosymbiotique, qui décrit la domestication de microorganismes comme à l'origine des organites que sont la mitochondrie et le chloroplaste [Ward, 1883, Margulis, 1976].

Modifications génomiques

Durant cette évolution, les génomes bactériens subissent différentes transformations. La taille des génomes, ainsi que leur proportion en bases G et C diminuent. Ainsi, certains symbiotes obligatoires possèdent des génomes parmi les plus petits connus [Moran et al., 2008]. Ce phénomène s'explique à la fois par le nombre restreint de fonctions nécessaires au symbiote, et par la dynamique des populations symbiotiques. En passant de l'état de bactérie libre à celui de symbiote, beaucoup de fonctions génomiques deviennent inutiles ou redondantes avec celles de l'hôte. Les mutations peuvent donc s'accumuler dans les régions génomiques liées à ces fonctions sans constituer un handicap, ce qui mène à leur disparition. Parmi ces fonctions, on peut trouver des mécanismes de réparation de l'ADN, ce qui entraine la diminution du taux de GC dans ces génomes [Wernegreen, 2005]. Par ailleurs, lors du transfert vertical des symbiotes, seul un faible nombre d'individus est transféré. Cet effet de *bottleneck* s'accompagne d'une augmentation de la dérive génétique et accélère l'évolution de ces génomes.

Les effets de la dérive génétique se traduisent aussi sur la structure des génomes, qui voient leur taille diminuer. Dans certains cas, ce processus peut se poursuivre, jusqu'à rendre

le symbiote obligatoire dépendant d'une autre bactérie de la communauté. Cela mène à sa disparition et à son remplacement par la nouvelle bactérie en tant que symbiote obligatoire [Pérez-Brocal et al., 2006].

Bilan des problématiques concernant les holobiontes

Les holobiontes sont donc des systèmes complexes qui posent de nombreuses questions que nous pouvons regrouper sous les trois questions : "Qui sont-ils ?", "Que font-ils ?" et "Comment évoluent-ils ?". Ces troix niveaux de lecture ne peuvent être compris que de manière restreinte par les méthodes classiques de la biologie. La caractérisation taxonomique des communautés bactériennes est fortement limitée par l'impossibilité de cultiver la plupart des microorganismes. De même, les études fonctionnelles des holobiontes sont difficiles et ne prennent en compte qu'une partie des fonctions de l'holobionte. En conséquence, il est difficile d'étudier les trajectoires évolutives que peuvent prendre les associations symbiotiques, qui sont diverses, et orientées par une multitude de forces évolutives qui s'exercent sur les symbiotes, dans et en dehors de leurs hôtes. Dans cette thèse, nous étudions le potentiel des données métagénomiques à faire progresser les connaissances sur ces trois questions.

1.2 Séquençage à haut-débit et métagénomique

1.2.1 De la microbiologie à la métagénomique

Bien que l'étude des communautés microbiennes soit ancienne, elle a longtemps été restreinte à l'utilisation de techniques d'imagerie permettant simplement d'observer des caractères morphologiques. Dans ce cadre, seul des organismes susceptibles d'être mis en culture pouvaient être étudiés. Ainsi, avant l'essor des technologies de biologie moléculaire, seul une étude à faible résolution d'une mince fraction des microbes existants était possible.

Les progrès des techniques de biologie moléculaire ont permis de contourner ces obstacles et ont contribué à révolutionner la microbiologie. Le séquençage Sanger, développé en 1977, a permis d'accéder à la structure et à la fonction des génomes bactériens, en donnant leur séquence sous forme de fragments appelés lectures et mesurant quelques centaines de bases. La même année, l'ARN ribosomique est décrit comme un marqueur permettant de classer taxonomiquement les espèces. Sur les bases de ces travaux, Pace propose en 1985 de séquencer l'ARN ribosomique directement dans l'environnement, sans passer par une étape de culture bactérienne. Cette idée permet de s'affranchir du biais de cultivabilité, qui rendait jusqu'alors invisible une large fraction de la diversité microbienne [Rappé and Giovannoni, 2003]. Proposé en 1998 par Handelsman [Handelsman et al., 1998], le terme de métagénomique désigne le séquençage direct de l'ADN dans un milieu, qui permet potentiellement d'accéder aux génomes de tous les membres d'une communauté.

	Technologie	Longueur des lectures (bases)	Taux d'erreur	Coût moyen par gigabase
Première génération	Sanger	400-900	<0.1%	NA
Seconde génération	Illumina	150 300 (lectures pairées)	<0.1%	$7 (HiSeq X)
	Roche 454	400	1%	$9,500
	ABI SOLiD	75	<0.1%	$70
Troisième génération	Pacific Bioscience	10k en moyenne	∼5%	$1,000
	Oxford Nanopore	10k en moyenne	∼5%	$750

Table 1 – ***Caractéristiques des principales techniques de séquençage.*** *D'après [Goodwin et al., 2016]*

1.2.2 Séquençage de deuxième génération

La technologie Sanger est toutefois limitée par le faible nombre de lectures qu'elle génère, et par conséquent le coût élevé que représente un grand projet de séquençage. De nouvelles technologies, qualifiées de séquençage de nouvelle génération (NGS) ont émergé successivement dans les années 2000 et ont permis le séquençage simultané d'un grand nombre de molécules. Ces technologies ne permettent pas d'accéder directement à la séquence complète d'un génome. À la sortie du séquenceur, on récupère des lectures dont la longueur n'excède pas quelques centaines de bases, et qui peuvent provenir de n'importe quelle région du génome. Par ailleurs, ces séquences ne sont pas totalement fidèles, et peuvent contenir des erreurs, comme des substitutions nucléotidiques, des insertions ou des délétions.

Les différentes technologies de séquençage haut-débit actuellement disponibles se démarquent sur trois points essentiels : la longueur des lectures obtenues, leur quantité, et leur fidélité. Leurs principales caractéristiques sont résumées dans le tableau 1.2.2. La technologie qui s'est imposée est commercialisée par la société Illumina. Par rapport aux méthodes concurrentes, les lectures issues de la technologie Illumina ne se distinguent ni par leur longueur, ni par leur taux d'erreur. En revanche, elles génèrent un volume de données largement supérieur, avec plusieurs gigabases par expérience, ainsi qu'un coût de séquençage bien moindre. L'émergence des technologies à haut-débit a révolutionné la génomique. Le premier séquençage d'un génome humain, permis par la technologie Sanger, a coûté près de 3 milliards de dollars [Venter et al., 2001], et a duré près d'une quinzaine d'années. En l'état actuel de la technologie Illumina, une telle tâche prend désormais une journée et moins d'un millier de dollars [Goldfeder et al., 2017].

1.2.3 La métagénomique moderne

Étant donné la complexité et la diversité des communautés microbiennes, la métagénomique nécessite des efforts de séquençage important, et son essor n'a été possible

qu'avec la généralisation des techniques de séquençage haut-débit.

Métagénomique ciblée et métagénomique plein-génome

On distingue deux principaux types de données métagénomiques. La métagénomique ciblée ou *metabarcoding* consiste à l'amplification puis au séquençage d'une région particulière du génome. Une région fréquemment utilisée est l'ADN ribosomique 16S des bactéries, qui est un excellent marqueur phylogénétique. À l'inverse, la métagénomique plein-génome ou shotgun consiste au séquençage de tout l'ADN contenu dans l'échantillon. Plutôt que d'amplifier une région spécifique du génome, tout l'ADN du génome ou du métagénome est découpé en fragments analysés avec les techniques haut-débit classiques. L'ensemble des génomes des membres de la communauté peut donc être séquencé.

Un certain flou existe sur la dénomination donnée à la métagénomique ciblée. Selon de nombreux auteurs, le terme de métagénomique est peu adapté, car il s'agit d'une technique ciblée sur une petite portion du génome [Esposito and Kirschberg, 2014]. Aussi, dans le reste du manuscrit, qui se concentre sur les techniques *shotgun*, nous emploierons le terme de métagénomique pour désigner la métagénomique plein-génome.

La métagénomique ciblée est utilisée pour caractériser taxonomiquement un échantillon. Parce que seule une petite fraction du génome est séquencée, elle est moins coûteuse que la métagénomique plein-génome, et permet d'identifier des organismes plus rares pour un effort de séquençage équivalent. Pour ces raisons, elle a été prioritairement utilisée dans les première étapes de la métagénomique, ainsi que pour de grands projets de catalogage de la diversité bactérienne.

Par rapport à la métagénomique ciblée, l'information obtenue en métagénomique plein-génome est à la fois plus volumineuse, plus difficile à interpréter et plus riche. L'information plein génome permet d'atteindre une meilleure résolution que les marqueurs utilisés en métagénomique ciblée. La métagénomique *shotgun* souffre moins des biais liés à l'amplification des séquences ciblées, ce qui la rend plus à même de représenter quantitativement des communautés. Enfin et surtout, alors que la métagénomique ciblée informe uniquement sur la composition taxonomique des communautés, la métagénomique plein génome renseigne sur le potentiel fonctionnel des communautés, à travers le répertoire de gènes de ses membres. Ainsi, la méthode *shotgun* est capable de répondre aux questions classiques posées par métagénomique sur les membres des communautés microbiennes qui sont "Qui-sont-ils ?" et "Que sont-ils capables de faire ?", tandis que la métagénomique ciblée est difficilement capable de répondre à la seconde.

En contrepartie, le séquençage d'un échantillon par la métagénomique plein génome est nettement plus coûteux, car une importante quantité de lecture doit être séquencée pour atteindre une couverture suffisante. Toutefois, grâce aux progrès continuels des technologies de séquençage, cette technique se démocratise largement, et les grands projets se sont multipliés. On peut citer le projet MetaSoil [Delmont et al., 2011] pour l'étude du microbiome du sol, ou le projet HMP pour *Human Microbiome Project* [Turnbaugh et al., 2007] qui vise à étudier les différents microbiotes humains. Métagénomiques *shotgun* et ciblée peuvent

également être utilisées conjointement. Elles sont complémentaires, notamment dans le cas des communautés les plus complexes, où les organismes rares ne peuvent être identifiés que par des méthodes ciblées. On peut citer le projet TARA Océans [Bork et al., 2015], qui vise à explorer la diversité microbienne des océans, et combine ces deux approches.

1.2.4 Défis bioinformatiques pour la métagénomique

Si les données métagénomiques permettent en principe de porter un nouveau regard sur des communautés jusqu'alors mal connues, ces nouvelles données possèdent certaines particularités qui justifient et nécessitent le développement de méthodes dédiées.

Volume de données

Les séquençages de données métagénomiques peuvent générer des volumes de données importants. Dans des écosystèmes complexes tels que le sol ou l'eau de mer, un important effort de séquençage est nécessaire pour caractériser les organismes rares [Welch and Huse, 2011, Roesch et al., 2007]. À titre d'exemple, un échantillon du projet TARA Océans peut contenir près de 300 millions de séquences. Par ailleurs, de nombreuses études nécessitent le séquençage et l'analyse conjointe de plusieurs dizaines ou centaines d'échantillons, afin de comparer des écosystèmes différents.

Ce volume de données important pose des problèmes à toutes les étapes de l'analyse. Des outils et bases de données dédiés existent pour le stockage, l'indexation et le catalogage de telles données (IMG/MER, CAMERA, MG-RAST, et EMG). D'un point de vue algorithmique, la recherche de solutions performantes permettant de gérer une multitude de jeux de données de taille importante est une priorité.

Diversité génomique

Les communautés bactériennes présentent généralement un continuum de diversité, découpé en plusieurs niveaux taxonomiques. Classiquement, on distingue au sein de ces communautés plusieurs espèces microbiennes différentes, dont les génomes peuvent présenter des régions homologues, par exemple suite au transfert horizontal d'un gène. Les abondances de ces espèces sont mesurées par leur couverture par les lectures métagénomiques, et elles peuvent être très déséquilibrées.

Par ailleurs, chaque espèce est représentée par un nombre variable d'individus, qui peuvent présenter des génotypes différents, incluant courts variants et variations structurales. Contrairement aux données génomiques dont la ploïdie est connue, les données métagénomiques abritent donc un nombre très important de variations. Ces différents variants peuvent également être quantifiés en mesurant leur couverture. En fonction de l'effort de séquençage fourni, certains variants rares pourront facilement être confondus avec les erreurs de séquençages, ou avec des régions provenant d'une autre espèce de la communauté.

Appréhender cette diversité est donc complexe, en particulier lorsqu'il s'agit de comparer différentes communautés pouvant abriter des espèces distinctes. Les tâches classiques de la

génomique telles que l'assemblage où la recherche de variants, sont rendues difficiles par ce polymorphisme [Sczyrba et al., 2017], et nécessitent le développement d'algorithmes dédiés. Bien souvent, une manière d'appréhender ce problème est de restreindre l'analyse à des unités taxonomiques opérationnelles distinguables par des critères arbitraires (par exemple un seuil de similarité), en ne considérant pas la variabilité au sein de ces unités qui peut pourtant avoir des impacts fonctionnels.

1.2.5 Séquençage de troisième génération et métagénomique

La principale restriction du séquençage de seconde génération est la longueur limitée des lectures, qui rend par exemple difficile voire impossible certains problèmes d'assemblage. Les technologies de séquençage les plus récentes peuvent être qualifiées de "longue portée" [Sedlazeck et al., 2018]. Elles permettent par exemple de séquencer des lectures plus longues, pouvant atteindre le million de bases (technologies PacBio ou NanoPore), ou bien de relier de courtes lectures provenant de la même région génomique (technologies 10X genomics et Hi-C). Ces techniques ont un potentiel élevé pour répondre à certains des problèmes posés par la métagénomique et exposés dans le paragraphe précédent. Ainsi, les plus longues lectures permettent d'améliorer les assemblages métagénomiques. Les techniques telles que le Hi-C peuvent être appliquées afin de distinguer les lectures provenant d'organismes différents [Burton et al., 2014]. Cependant, le débit proposé par les technologies de séquençage de longues lectures ne rivalise pas à l'heure actuelle avec le séquençage NGS. Ce paramètre étant critique dans les applications métagénomiques qui nécessitent un important effort de séquençage pour appréhender la diversité des communautés, l'usage des nouvelles méthodes reste conditionné à leurs progrès technologiques futurs.

1.3 Le modèle biologique de la thèse : le puceron du pois

Le puceron du pois, *Acyrthosiphon pisum*, est un insecte de l'ordre des Hémiptères, qui s'avère être un modèle privilégié au sein de la communauté scientifique travaillant sur la symbiose. Comme de nombreux insectes, il est associé de manière durable à plusieurs bactéries symbiotiques, qui contribuent à son métabolisme et à son phénotype, et il forme un parfait exemple d'holobionte constitué autour d'un insecte [Mandrioli and Manicardi, 2013]. Ce système présente par ailleurs un certain nombre de caractéristiques biologiques qui le rendent à la fois aisé à étudier, et riche d'enseignements sur sa diversité et son histoire évolutive.

1.3.1 Quelques éléments de biologie sur le puceron

Un insecte d'intérêt agronomique

Les pucerons tiennent une place importante parmi les insectes ravageurs de culture. Ils s'attaquent à une grande variété de plantes cultivées. Bien qu'ils soient très variables et

difficiles à estimer, on chiffre les dégâts causés en centaines de millions de dollars [Morrison and Peairs, 1998, Oerke et al., 1994]. Un exemple célèbre est *Daktulosphaira vitifoliae*, le phylloxera, qui a ravagé les vignes françaises au XIX^{e} siècle. Le puceron du pois en particulier, vit et se nourrit sur de nombreuses espèces cultivées de fabacées parmi lesquelles le pois, la luzerne et le trèfle.

Les pucerons causent principalement deux types de dégâts à leur hôte [Dedryver et al., 2010]. En se nourrissant de la sève élaborée de leur plante-hôte, ils détournent ainsi une partie des nutriments nécessaires à leur croissance. De plus, ils sont les principaux insectes vecteurs de virus pathogènes pour les plantes [Nault, 1997].

Le cycle de reproduction

Le cycle de reproduction des pucerons comprend une phase de reproduction sexuée et une phase asexuée. Durant le printemps et l'été, les femelles se multiplient par parthénogénèse. Avec l'arrivée de l'automne, elles donnent naissance aux formes sexuées, ce qui initie la phase de reproduction sexuée. Les œufs issus de cette reproduction sexuée resteront en diapause jusqu'au printemps suivant, où le cycle reprendra. Durant la phase parthénogénétique, la colonie se reproduit très rapidement, ce qui lui permet d'envahir très vite une nouvelle plante-hôte [Miura et al., 2003], tandis que la reproduction sexuée et les œufs permettent la survie pendant l'hiver en résistant aux températures négatives. En laboratoire, ce cycle de vie se traduit par la possibilité de conserver facilement un génotype donné, en maintenant les conditions de l'été, et donc la parthénogénèse.

Structure en biotypes

Les pucerons peuvent être assimilées à des parasites de leur plante-hôte. Pour cette raison, ils entretiennent des relations fortes avec les plantes. Le puceron du pois porte mal son nom, car il ne se nourrit pas uniquement du pois mais d'une multitude d'espèces de la famille des Fabacées (légumineuses). Cependant, cette espèce de puceron forme un complexe de biotypes, chacun de ces biotypes étant une population de pucerons adaptée à une ou plusieurs espèces de fabacées. À ce jour, on dénombre au moins 15 biotypes [Peccoud et al., 2009a, Peccoud et al., 2015], qui ont été décrits selon leur profil génétique et leur gamme d'hôte. Des analyses génétiques et phylogénétiques ont révélé un *continuum* de divergence entre ces biotypes. Certains continuent à échanger fréquemment du matériel génétique, tandis que d'autres sont isolés reproductivement et pourraient constituer de nouvelles espèces. Ce complexe s'est diversifié il y a environ 10000 ans, concomitamment à la domestication par l'Homme des plantes pour l'agriculture [Peccoud et al., 2009b]. Ces biotypes sont spécialisés : leur performance (le taux de multiplication) est accrue sur la plante-hôte à laquelle ils sont adaptés [Peccoud et al., 2009a]. Ils partagent néanmoins la possibilité de pouvoir se reproduire sur un hôte universel (la féverole *Vicia faba*), ce qui est à nouveau un atout pour l'élevage de ces populations en laboratoire.

1.3.2 Le cortège symbiotique du puceron du pois

Les insectes, qui représentent une part importante de la diversité du vivant et sont présents dans une multitude de niches écologiques, sont très fréquemment associés à des symbiotes. On estime ainsi que près de 10% d'entre eux ont besoin de symbiotes pour leur survie [Moran and Baumann, 2000]. Parmi les associés possibles des insectes, on peut citer des levures, des virus ou des bactéries. Le puceron du pois est un organisme modèle pour l'étude des symbioses chez les insectes, et abrite plusieurs symbiotes bactériens.

***Buchnera aphidicola* : le symbiote obligatoire typique des pucerons**

Certains symbiotes échangent des nutriments avec leur hôte, en synthétisant des molécules que l'hôte ne peut pas se procurer dans son milieu naturel. Du fait de son régime alimentaire, le puceron ne peut se procurer certains acides aminés indispensables. *Buchnera aphidicola* est un symbiote les produisant pour son hôte. Il s'agit d'un des premiers exemples connus de symbiote d'insecte [Buchner, 1965], et il présente toutes les caractéristiques des symbioses mutualistes présentées dans la première section.

Il s'agit d'un symbiote endocellulaire, situé dans des cellules spécialisées appelées bacteriocytes, présentes dans le corps de l'insecte. Ce symbiote est dit obligatoire ou primaire, car la croissance de son hôte est fortement ralentie en son absence [Mittler, 1971, Douglas and Prosser, 1992]. Réciproquement, il ne peut plus assurer de nombreuses fonctions nécessaires à la vie en dehors de son hôte [Shigenobu et al., 2000]. *Buchnera* est transmis verticalement à la fois durant la reproduction sexuée et la parthénogénèse, en colonisant les oeufs ou les embryons asexués [Miura et al., 2003]. Le caractère exclusivement vertical du transfert de *Buchnera* au cours de la reproduction a des conséquences sur la dynamique évolutive de *Buchnera* et de son hôte. *Buchnera* est ainsi le premier cas décrit de co-diversification d'un insecte avec son symbiote bactérien [Moran et al., 1993], et ce à différents niveaux phylogénétiques [Funk et al., 2001]. L'origine de cette relation symbiotique est l'infection d'un puceron par une gamma-protéo-bactérie libre qui remonterait à entre 100 et 200 millions d'années [Moran et al., 1993]. Les études phylogénétiques confirment une origine monophylétique de *Buchnera*, qui aurait ensuite coévolué avec son hôte, permettant à cette association d'être retrouvée actuellement dans la quasi-totalité des espèces de pucerons.

Au cours de la transmission verticale de *Buchnera*, le nombre de bactéries transmises à la descendance est significativement inférieur à la population présente dans un hôte [Mira and Moran, 2002]. La transmission exclusivement verticale et le "goulot d'étranglement" par lequel passent les populations symbiotiques à chaque génération ont d'importantes conséquences évolutives. Ils accélèrent la dérive génétique et la fixation de mutations au sein de la population [Moran, 1996], ce qui accélère le processus de réduction du génome fréquemment observé chez les symbiotes obligatoires. Ainsi, le génome du *Buchnera aphidicola* fait seulement 640 kilobases, et son taux de bases GC dépasse rarement les 25% [Degnan et al., 2011]. En comparant les principaux génomes disponibles de *Buchnera*, et un génome proche de leur ancêtre commun tel que celui d'Escherichia coli (4.6 Mb, 50% GC), on se rend compte d'une remarquable conservation de la synténie, sans réarrangement chromosomique ou acquisition

de gènes [Tamas et al., 2002]. Ce phénomène s'explique entre autres par la perte de gènes responsables de la recombinaison de l'ADN, et par le faible nombre de transferts horizontaux potentiels dans l'environnement qu'est l'hôte [Tamas et al., 2002, Nikoh et al., 2010]. D'un point de vue fonctionnel, cette réduction du génome s'accompagne par la perte d'un grand nombre de fonctions devenues inutiles ou redondantes au sein de l'hôte. L'étude du répertoire des gènes de *Buchnera* montre à la fois la prédominance des fonctions de biosynthèse d'acides aminés, et la complémentarité entre les voies de synthèse d'acides aminés du puceron et de *Buchnera*.

Le génome nucléaire de *Buchnera* est complété par deux plasmides, présents en de multiples copies, et dédiés à la synthèse d'acides aminés. Le premier porte un opéron de synthèse de la leucine (pLeu) [Silva et al., 1998], tandis que le second permet la synthèse du tryptophane (pTrp) [Panina et al., 2001]. Ces deux acides aminés font partie des acides aminés indispensables à l'hôte, et ces deux plasmides jouent donc un rôle clé dans la relation symbiotique. La présence de multiples copies de ce plasmide pourrait être un moyen d'augmenter les capacités de synthèse de ces acides aminés. Ce nombre de copies est également variable entre différentes espèces de puceron [Thao et al., 1998].

Des symbiotes secondaires aux effets phénotypiques variés et peu connus

En complément de *Buchera*, plusieurs autres symbiotes ont été décrits chez le puceron du pois. Il s'agit de symbiotes secondaires, qui sont donc retrouvés en fréquences variables dans les populations de pucerons. Les symbiotes décrits à ce jour pour le puceron du pois sont au nombre de huit : *Hamiltonella defensa*, *Regiella insecticola*, *Serratia symbiotica*, *Rickettsiella viridis*, *Fukatsuia symbiotica* , *Spiroplasma sp.*, *Rickettsia sp.* and *Wolbachia sp.*. Il est possible de retirer ou ajouter ces symbiotes en laboratoire, et de mesurer ainsi les effets sur leur hôte. Des travaux ont ainsi démontré une multitude d'effets (résumés dans le tableau 2), tels l'adaptation à certaines plantes-hôtes [Tsuchida et al., 2004], la résistance à des températures élevées [Montllor et al., 2002], le changement de couleur [Tsuchida et al., 2010], la résistance contre des ennemis naturels [Oliver et al., 2003], ou la manipulation du mode de reproduction [Simon et al., 2011].

Un exemple très étudié de ces symbiotes secondaires est *Hamiltonella defensa*, qui apporte à l'hôte une résistance à la guêpe parasitoïde *Aphidius ervi*, principal ennemi naturel du puceron du pois [Oliver et al., 2003]. Cette protection résulte en réalité de la présence d'un bactériophage, nommé APSE, qui produit différentes toxines pouvant empêcher le développement du parasitoïde [Oliver et al., 2014].

Dynamique évolutive

Les symbiotes secondaires sont majoritairement transmis verticalement. Bien que le transfert soit le plus souvent d'origine maternelle, certains cas de transfert vertical d'origine paternelle ont été décrits chez le puceron [Moran and Dunbar, 2006], vraisemblablement à des taux très faibles [Peccoud et al., 2014]. Par ailleurs, et contrairement au symbiote obligatoire *Buchnera*, des cas de transferts horizontaux ont été observés entre des individus de même

Symbiote	Effets phénotypiques
Buchnera aphidicola	Symbiote obligatoire Fournit des acides aminés indispensables [Buchner, 1965]
Hamiltonella defensa	Protection contre des parasitoïdes [Oliver et al., 2003]
Regiella inseticola	Protection contre des pathogènes fongiques [Łukasik et al., 2013] Adaptation à la plante-hôte [Tsuchida et al., 2004]
Fukatsuia symbiotica	Protection contre des parasitoïdes [Guay et al., 2009] Résistance à la chaleur [Guay et al., 2009]
Serratia symbiotica	Résistance à la chaleur [Russell and Moran, 2006] Protection contre des parasitoïdes [Oliver et al., 2003]
Rickettsia sp.	Protection contre des pathogènes fongiques [Łukasik et al., 2013] Résistance à la chaleur [Montllor et al., 2002]
Rickettsiella viridis	Changement de couleur [Tsuchida et al., 2010] Protection contre des pathogènes fongiques [Łukasik et al., 2013] Protection contre des parasitoïdes [Leclair et al., 2017]
Spiroplasma sp.	Protection contre des pathogènes fongiques [Łukasik et al., 2013] Manipulations reproductives [Simon et al., 2011]

Table 2 – ***Effets phénotypiques connus des symbiotes du puceron du pois***

espèce ouu d'espèces différentes [Sandström et al., 2001, Henry et al., 2013, Russell et al., 2003]. La fréquence et le mécanisme de ce phénomène sont mal connus, mais il reste minoritaire par rapport à la transmission verticale. Les modes de transfert proposés à ce jour comprennent des parasitoïdes vecteurs de symbiotes [Gehrer and Vorburger, 2012], et le transfert via la plante-hôte ou le contact entre congénères [Chrostek et al., 2017]. La possibilité d'un transfert horizontal a des conséquences sur l'évolution des populations d'insectes et de symbiotes. Les symbiotes secondaires transmissibles horizontalement possèdent un répertoire génétique, qui peut favoriser l'adaptation à des nouvelles niches écologiques. En ce sens, ils sont des éléments mobiles au rôle comparable à celui des plasmides dans les populations bactériennes. Les symbiotes peuvent également échanger du matériel génétique avec leur hôte. Ainsi, certains gènes d'origine symbiotique ont pu s'intégrer au génome de l'hôte par transfert horizontal, et dans certains cas lui apporter de nouvelles fonctions [Nakabachi, 2015].

Une des principales limites de la compréhension actuelle du rôle des symbiotes secondaires réside dans la caractérisation de leurs effets. Dans l'ensemble, l'effet phénotypique de ces bactéries sur leur hôte est décrit de manière incomplète, notamment car il peut varier pour une même espèce de symbiote et selon le contexte environnemental [Leclair et al., 2016]. Ainsi, des effets phénotypiques observés dans une population de pucerons infectés par un symbiote peuvent ne pas être reproduits dans une population porteuse d'une souche différente de ce symbiote.

Ressources génomiques existantes

L'holobionte du puceron du pois étant un modèle pour l'étude des relations symbiotiques, une quantité relativement élevée de données génomiques le concernant est disponible. Outre le génome de l'hôte, dont un premier assemblage a été publié en 2010 [Richards et al., 2010], certains de ses symbiotes (ou leurs équivalents chez d'autres espèces de pucerons) ont également été séquencés et assemblés. C'est le cas du symbiote obligatoire *Buchnera aphidicola*, ainsi que de *Hamiltonella defensa*, *Regiella insecticola*, *Serratia symbiotica*, *Fukatsuia symbiotica*. Ce bilan peut toutefois être contrasté, d'abord car les génomes de certains symbiotes ne sont pas disponibles à l'heure actuelle, puis parce qu'un unique génome de référence est insuffisant pour représenter la diversité de ces symbiotes. Un des écueils de la caractérisation des effets phénotypiques des symbiotes est la variabilité de ces effets entre différentes souches symbiotiques. Ainsi, plusieurs études montrent des effets phénotypiques différents en fonction des génotypes symbiotiques considérés [Brandt et al., 2017, Leclair et al., 2016, Oliver et al., 2005]. À ce jour, seuls les symbiotes *Hamiltonella defensa* et *Serratia symbiotica* a été étudiée par une approche de génomique comparative de plusieurs génomes [Chevignon et al., 2018, Manzano-Marín and Latorre, 2016]. Cette étude illustre les importantes différences entre les génomes de différentes souches d'*Hamiltonella* et de *Serratia* au niveau de la structure du génome, et du répertoire d'éléments transposables présents. Une meilleure compréhension du rôle des symbiotes du puceron passerait par l'application d'une telle approche aux autres symbiotes du complexe.

1.4 Objectifs de la thèse

Au travers de cette thèse, nous nous proposons d'explorer la diversité génomique du microbiote d'individus et de populations du puceron *A. pisum* sous l'angle nouveau des données de métagénomique haut-débit. Si les relations symbiotiques sont connues de longue date, leur étude est restée limitée tout d'abord à des observations, puis à des approches moléculaires. Ces techniques ont révélé l'importance des relations symbiotiques au sein du vivant, et permis d'élucider certains éléments de leur histoire évolutive. Elles souffrent cependant de trois limitations principales : leur résolution dépasse difficilement celle de l'espèce, elles ne permettent souvent pas de reconstruire les histoires évolutives de l'ensemble des associations hôte-microbiote, et elles sont souvent incapables de décrire finement l'interaction fonctionnelle entre un hôte et ses symbiotes.

La métagénomique apporte une information nouvelle, à haute résolution et permettant d'accéder au potentiel fonctionnel des symbiotes. Cette approche plein-génome permet de distinguer des souches bactériennes différentes, et d'amorcer la description de leurs implications fonctionnelles [Zhang and Zhao, 2016]. Pourtant, cette promesse donnée par cette nouvelle technologie est encore loin d'être réalisée. Il s'agit d'une discipline jeune, qui nécessite de nouveaux développements pour tirer le plein potentiel de ces données à haute-résolution.

Au cours de cette thèse, nous nous proposons de présenter à la fois des méthodologies innovantes pour l'étude des relations hôte-microbiote, et des résultats nouveaux sur l'holobionte du puceron du pois. Cet organisme semble être un modèle idéal pour le développement de nouvelles approches métagénomiques, car il abrite une communauté d'une complexité relativement simple, avec des génomes assez bien connu, ce qui permet de contrôler la fiabilité des méthodes développées. Nous ferons dans un premier temps un état de l'art des méthodologies adaptées à la métagénomique. Ensuite, nous appliquerons des méthodes bioinformatiques pour étudier finement la diversité symbiotique de notre modèle, à travers l'étude du polymorphisme observable à partir de données métagénomiques. Pour ce faire, nous tirerons profit d'un large projet de reséquencage, couvrant les différents biotypes du puceron du pois. Enfin, dans une dernière partie, nous proposerons une méthode originale qui permet d'étudier des variations structurales souvent négligées dans de telles données.

Chapitre 2

État de l'art : méthodes d'analyse de données métagénomiques

Le séquençage d'échantillons environnementaux (métagénomique) permet d'apporter un regard nouveau sur la diversité, la structure et le fonctionnement des interactions biologiques, en particulier au sein des communautés microbiennes. Mais l'étude des communautés microbiennes par l'intermédiaire de courtes lectures pose certains problèmes. En particulier, il est difficile d'identifier parmi l'ensemble des lectures celles qui proviennent d'un même organisme. Les tâches classiques de la génomique prennent ainsi une autre dimension quand il s'agit de traiter des données métagénomiques, et de nouvelles questions méthodologiques se posent. Traditionnellement, on résume les questions posées au sujet d'une communauté microbienne à "Qui sont les organismes présents dans la communauté ?" et "Que font-ils ?". Le développement de méthodes bioinformatiques pour la métagénomique s'est donc naturellement orienté vers ces deux questions, que sont la caractérisation taxonomique et fonctionnelle de communautés microbiennes, auxquelles s'ajoutent la métagénomique comparative, dont l'objet est d'étudier les similarités entre différentes communautés. Ce chapitre présente un aperçu de ces méthodes, construit en partie grâce à plusieurs revues bibliographiques : [Segata et al., 2013, Soueidan and Nikolski, 2015, Mande et al., 2012, Garza and Dutilh, 2015, Sangwan et al., 2016, Brown, 2015, Peabody et al., 2015].

2.1 Caractérisation taxonomique

La caractérisation taxonomique, ou *metagenomic profiling*, vise à répondre à une question qui ne se pose généralement pas lors de l'étude d'un seul organisme, en décrivant au niveau taxonomique les organismes présents dans l'échantillon. Il s'agit, à partir de lectures métagénomiques, d'identifier et éventuellement de quantifier les organismes qui sont présents au sein d'une communauté. Cette section vise à présenter les différentes familles de méthodes qui permettent de répondre à cette question. Ces méthodes diffèrent notamment par le recours ou non à des bases de données existantes, construites à partir des génomes déjà séquencés et identifiés. Généralement, les méthodes qui reposent lourdement sur ces bases de données se révèlent difficilement capables d'identifier des organismes non caractérisés précédemment, ce qui est un problème critique en métagénomique. Par ailleurs, la performance de l'assignation taxonomique varie selon les différentes techniques. Certaines fournissent au mieux un

inventaire des espèces présentes, tandis que d'autres permettent d'étudier des variations plus fines entre des souches microbiennes. C'est sous l'angle de ces deux aspects que cette section présente les différentes méthodes permettant la caractérisation taxonomique d'échantillons métagénomiques. Nous nous pencherons brièvement sur les méthodes dédiées aux données de *barcoding*, avant d'étudier plus en détail celles permettant l'assignation taxonomique de lectures issus de métagénomique plein-génome, recourant ou non à des bases de données de référence.

2.1.1 Séquençage d'amplicons

Bien qu'il ne s'agisse pas à proprement parler de métagénomique, le séquençage environnemental d'amplicons est une méthode répandue pour l'analyse de la composition taxonomique d'une communauté microbienne. Parce qu'il ne s'agit pas d'une méthode plein-génome, un effort de séquençage modéré permet de séquencer une grande majorité des organismes présents. Une première approche pour analyser de telles données consiste à aligner les séquences contre des bases de données de séquences de marqueurs phylogénétiques (principalement l'ADN ribosomique 16S), tels que Silva [Quast et al., 2013] ou GreenGenes [DeSantis et al., 2006]. Les lectures sont alignées sur un alignement de référence, généralement à l'aide d'un programme d'alignement multiple tel que SINA [Pruesse et al., 2012] pour Silva. Étant donnée la taille restreinte des séquences des marqueurs utilisés, cette tâche reste relativement rapide. En revanche, cette méthode ne permet pas de traiter les organismes absents de ces bases de référence, et elle présente un certain nombre de biais liés notamment aux étapes d'amplification. Une autre approche, suivie par exemple par les outils Qiime [Caporaso et al., 2010] et Mothur [Schloss et al., 2009] consiste à effectuer un clustering des lectures suivant un seuil de similarité, de manière à former des OTUs (Operational Taxonomic Units). Les bases de données de référence peuvent éventuellement être utilisées pour annoter les OTUs ainsi obtenus par un taxon.

La principale faiblesse du metabarcoding pour la caractérisation taxonomique de communautés provient du fait que bien souvent un unique marqueur est utilisé pour décrire l'échantillon. Ces marqueurs offrent rarement une résolution suffisante pour déterminer quelles sont les espèces présentes dans l'échantillon, et *a fortiori* ne permettent pas d'identifier avec fiabilité différentes souches. Par ailleurs, d'éventuels transferts horizontaux de gènes d'ARN ribosomique peuvent également tromper l'assignation taxonomique [Schouls et al., 2003]. Enfin, la quantification de l'abondance des taxons peut être perturbée par des biais d'amplification ou des variations de nombre de copies du marqueur considéré [Schouls et al., 2003].

2.1.2 Méthodes plein-génome avec référence

Alignement de séquences

Dans le cas du séquençage d'une communauté composée d'organismes pour lesquels un grand nombre de génomes de référence est disponible, une première approche est d'aligner les lectures métagénomiques contre cette collection de génomes. Le programme BlastX [Altschul et al., 1990] a été majoritairement utilisé pour l'analyse des premiers jeux de données métagénomiques, mais des aligneurs plus rapides tels que DIAMOND [Buchfink et al., 2014] ont également été conçus pour mieux passer à l'échelle sur des jeux de données importants. Le programme MEGAN [Huson et al., 2016] est une référence dans l'analyse d'alignements de jeux de données métagénomiques. À partir d'un résultat de type Blast, il permet d'assigner une lecture à un taxon, qui est le plus petit ancêtre commun des taxons avec lesquels cette lecture s'aligne. Il propose en complément de multiples options de visualisation. En corrigeant le nombre d'alignements par la longueur des génomes de référence, des outils tels que GAAS [Angly et al., 2009] ou GRAMMY [Xia et al., 2011] permettent de quantifier l'abondance des différents taxons. La résolution atteinte par ces outils est hautement dépendante de la densité de la base de données de référence utilisée. Par ailleurs, l'assignation de lectures à des génomes de référence partageant une forte similarité de séquence (de souches proches par exemple) nécessite l'emploi de méthodes dédiées, qualifiées de *strain tracking*. Les outils Pathoscope [Francis et al., 2013] et Sigma [Ahn et al., 2015] permettent d'effectuer cette tâche.

L'alignement des lectures métagénomiques sur des bases de données de référence présente deux inconvénients majeurs. Le premier est lié au temps nécessaire à l'alignement des séquences. En effet, le passage à l'échelle sur des jeux métagénomiques pouvant comporter des centaines de millions de lectures est difficile. Puisqu'il est nécessaire d'aligner chaque lecture, le temps dévolu à l'alignement augmente rapidement avec la taille des projets de séquençage. Ensuite, cette approche est très dépendante de la quantité et de la qualité des génomes de référence disponibles. La plupart des microorganismes constituant les échantillons métagénomiques ne sont pas cultivables, et le nombre de génomes de référence est limité. Ainsi, à l'exception des bactéries d'intérêt médical pour lesquelles de nombreuses souches sont connues, ces méthodes ne permettent pas d'atteindre une résolution supérieure à l'espèce. Finalement, en dehors de communautés modèles, de telles approches ne permettent pas d'étudier finement la diversité et le potentiel fonctionnel des microbiomes.

À partir de marqueurs phylogénétiques

Les méthodes basées sur l'alignement de lectures contre des génomes complets rencontrent des problèmes de passage à l'échelle dus à la taille des séquences de référence et au nombre de lectures à traiter. En conséquence, certaines des techniques d'assignation taxonomique utilisant des références se contentent d'un ensemble de séquences de gènes au lieu de génomes complets. Amphora [Wang and Wu, 2013] utilise 31 gènes présents en copie unique dans les génomes. La faible longueur de ces séquences de référence favorise le passage à l'échelle par rapport aux méthodes d'alignement contre des génomes entiers. Par

rapport à la métagénomique ciblée, l'emploi de plusieurs gènes distincts diminue les erreurs dues à d'éventuels transferts horizontaux. Le fait que ces gènes soient présents en copie unique permet également une meilleure quantification de l'abondance des espèces. Enfin, la résolution obtenue est supérieure à celle permise en métagénomique ciblée car ces gènes évoluent généralement plus rapidement que l'ARN ribosomique 16S. Dans le meilleur des cas, des souches bactériennes distinctes peuvent éventuellement être distinguées. Metaphlan [Truong et al., 2015] utilise quant à lui une base de données contenant près d'un million de gènes spécifiques de certains taxons. Des outils complémentaires permettent de détecter l'existence de différentes souches bactériennes à partir de profils de SNPs au sein des gènes marqueurs, ou par la présence/absence de gènes d'un pangénome identifié à partir de génome de référence. L'une des limites de cette approche est qu'il est difficile de complémenter la base de données par de nouveaux génomes en identifiant de nouveaux marqueurs spécifiques.

Les outils reposant sur des bases de données de séquences de marqueurs phylogénétiques offrent de bons résultats lorsque la communauté étudiée est bien représentée dans la base de données [Sankar et al., 2015]. Ils présentent en revanche des limites pour caractériser des souches bactériennes, et nécessitent pour ce faire des bases de données très fournies, ce qui restreint ces analyses à des communautés modèles.

Méthodes sans alignement

L'alignement de grands ensembles de lectures sur une large collection de génomes complets est une tâche couteuse qui rend difficile l'alignement de chaque lecture d'un métagénome sur un génome de référence dans un temps raisonnable. Pour palier ce problème, des méthodes d'assignation taxonomique sans alignement ont été développées. La solution retenue pour permettre le passage à l'échelle de l'assignation taxonomique métagénomique est d'utiliser de courtes séquences, nommées *kmers*, soit des séquences de taille k. Du fait de leur courte longueur, il est possible d'énumérer et d'indexer tous les *kmers* présents dans des génomes. Le premier outil utilisant cette technique est Kraken [Wood and Salzberg, 2014]. Kraken embarque une large base de données, contenant les kmers présents dans près de 25,000 génomes de bactéries, archées et virus. Il associe chaque lecture d'un échantillon au génome lui ressemblant le plus. Une place est laissée à l'incertitude : si aucun génome n'est suffisamment proche de la lecture, cette dernière est associée à un niveau taxonomique plus élevé. Afin de permettre des performances compatibles avec les données métagénomiques, ces outils nécessitent l'usage de structure de données particulières. Kraken repose par exemple sur une table de hash, construite à l'aide de minimiseurs de kmers, ce qui permet de requêter rapidement des kmers voisins, qui ont de grandes chances de partager un même minimiseur. Dans son mode le plus rapide, Kraken peut assigner près de 4 millions de lectures par minute, ce qui en fait un outil très rapide. En revanche, il est nécessaire de charger en mémoire l'intégralité de l'index, ce qui nécessite des quantités élevées de mémoire (près de 70 GB pour l'index complet) et limite le nombre de génomes pouvant être inclus dans la base de données. Cette approche est raffinée dans l'outil Clark [Ounit et al., 2015], qui construit un index plus léger à partir de *kmers* spécifiques de chaque génome. Plus récemment, les outils

Kaiju [Menzel et al., 2016] et Centrifuge [Kim et al., 2016] ont été développés en utilisant une autre structure d'indexation ne reposant pas sur une table de hash des kmer, mais sur un FM-index, et qui permet de réduire l'usage mémoire et d'améliorer l'assignation. Ainsi, Centrifuge est par exemple capable de stocker 4300 génomes procaryotes description dans un index occupant 4 Go de mémoire.

Limites des méthodes basées sur des références

Toutes ces méthodes reposent fortement sur des bases de données constituées à partir de génomes déjà séquencés. Cette approche se heurte à la faible représentativité de ces bases de données. Si un grand nombre d'espèces ont été séquencées, le nombre de souches disponibles pour une même espèce varie grandement. Ainsi, la base de données Refseq comporte à ce jour près de 144,000 génomes provenant d'environ 11,000 espèces, soit près de 13 génomes par espèce en moyenne. Cet effort de séquençage important ne représente pourtant qu'une infime partie des espèces bactériennes existantes (grossièrement estimé à plus d'un milliard dans [Dykhuizen, 2005]). De plus, 60 % de ces espèces ne sont représentées que par une seule souche, et 14% des espèces les mieux représentées représentent 90% des génomes de RefSeq. Parmi ces espèces abondamment séquencées et assemblées, la plupart sont d'intérêt biomédical. Ainsi, les 3 espèces les plus abondantes dans RefSeq sont *Escherichia coli*, *Salmonella enterica* et *Staphylococcus aureus*, avec près de 10 000 souches chacune. L'assignation taxonomique de lectures issues de communautés peu étudiées souffre ainsi d'une faible résolution taxonomique : la caractérisation de la communauté se limite le plus souvent à l'inventaire des espèces présentes. Pourtant, des différences fonctionnelles importantes peuvent s'expliquer par des variations génomiques à des échelles inférieures. Pour répondre à cette limite, quelques approches ont été développées pour rechercher et exploiter les variations par rapport aux séquences de référence. ConStrains [Luo et al., 2015] et StrainPhlan [Truong et al., 2017] utilisent des profils de SNPs, détectés sur les gènes marqueurs de la base de données Metaphlan. S'il est ainsi possible de reconstruire une phylogénie des différentes souches identifiées, cette analyse est restreinte à quelques gènes, et ne permet donc pas d'accéder à la séquence génomique complète des micro-organismes, ce qui empêche d'évaluer l'impact fonctionnel des différentes souches.

2.1.3 Méthodes sans référence

La diversité microbienne étant en grande partie inconnue, les méthodes requérant la comparaison à des bases de données de référence montrent rapidement leurs limites. Des méthodes qualifiées de *de novo* ont été développées pour identifier de nouveaux génomes à partir de données métagénomiques en recourant pas ou peu à des génomes de référence, et sont le principal moyen d'identifier les membres des microbiotes. Dans ces méthodes, le principal objectif est de regrouper les lectures provenant du même organisme.

Assemblage metagénomique

Principe de l'assemblage

Les limites actuelles des méthodes de séquençage rendent impossible de séquencer en une seule lecture des génomes complets. L'assemblage est la tâche permettant de transformer une multitude de courtes lectures en des portions plus longues du génome. Les programmes d'assemblage génèrent généralement un graphe représentant les chevauchements entre les lectures. On distingue deux principales familles d'assembleurs, en fonction du type de graphe utilisé par l'assemblage. Les assembleurs à *overlap graph* tels que Celera [Denisov et al., 2008] utilisent les lectures complets comme nœuds du graphe. Les assembleurs reposant sur des graphes de *De Bruijn*, tels que Abyss [Simpson et al., 2009] recourent quant à eux à un graphe de *kmers* (mots de taille k) plus petits que les lectures, qui rend plus facile la détection de chevauchements.

Dans ce graphe, l'assembleur recherche des chemins, dont la séquence va former des contigs, représentatifs du génome de l'organisme séquencé. Bien que ce soit possible, il est rare qu'un génome soit assemblé sous la forme d'un unique contig, car les graphes d'assemblage sont souvent complexes, ce qui contraint l'assembleur à interrompre les contigs. En particulier, c'est la présence de répétitions dans le génome qui fait qu'un assemblage est généralement découpé en contigs plus petits que le génome. Par exemple, lorsqu'un graphe de *De Bruijn* est utilisé pour l'assemblage, toute répétition d'un kmer au sein du génome se traduit par une structure en "X" dans le graphe. Cette structure ne peut être résolue par l'assembleur sans information extrinsèque, ce qui force le programme à interrompre les contigs au niveau de telles répétitions. L'emploi de kmers de plus grande taille permet de moins rencontrer ce problème.

Une autre source de complexité dans la structure de ces graphes est l'existence de chemins alternatifs, qui peuvent être dus soient à des erreurs de séquençage, soit à de véritables variations de la séquence génomique (par exemple les deux allèles d'un individu diploïde). Ces variants génèrent dans le graphe des structures en forme de bulle. Dans un graphe de *De Bruijn* constitué de kmers de taille k, une mutation ponctuelle ou erreur de séquençage se traduit par une bulle dont la longueur est de k kmers. La stratégie employée par la plupart des assembleurs est de retirer les kmers les moins abondants dans le graphe, qui correspondent généralement à des erreurs de séquençage, et à fusionner les bulles restantes dues au polymorphisme. Toutefois, l'assemblage des génomes très hétérozygotes reste problématique.

Un autre paramètre important dans l'assemblage d'un génome est la taille de celui-ci. Un grand génome nécessite d'importantes ressources informatiques, ce qui peut s'avérer limitant. À titre d'exemple, l'assemblage d'un jeu de données provenant d'un échantillon de sol peut occuper près de 350GB de RAM avec un assembleur moderne [Li et al., 2015]. Les plus grands génomes sont également plus sujets aux répétitions. Il y a par exemple près de 22,000 kmers de taille 101 répétés dans le génome d'*Escherichia Coli* (4.6Mb), et 92 millions dans un génome humain (3.4Gb).

Il est ensuite nécessaire d'évaluer la qualité d'un assemblage. Pour cela, un premier critère important est la longueur et le nombre des contigs. Un bon assemblage est constitué d'un

faible nombre de contigs de grande taille, dont la somme des longueurs approche la longueur du génome ciblé. Un indicateur fréquemment utilisé et qui synthétise ces critères est le N50, qui est la longueur minimale permettant de couvrir au moins la moitié du génome avec des contigs plus grands. Toutefois, bien qu'étant des méthodes *de novo*, la bonne évaluation des assemblages nécessite également de vérifier la véracité des contigs, par exemple en les alignant au génome de référence attendu quand cela est possible. C'est rendu possible par des outils tels que Quast [Gurevich et al., 2013], qui reportent le nombre d'erreurs commises lors de l'assemblage.

Méthodes d'assemblage métagénomique

Les méthodes d'assemblage citées précédemment ont été développées dans l'objectif d'assembler un unique génome provenant d'une unique espèce. Lorsqu'elles sont appliquées à des données métagénomiques, comme dans [Venter et al., 2004], les principales difficultés rencontrées en génomique classique sont exacerbées à cause de la diversité présente dans les communautés bactériennes.

Premièrement, une multitude d'espèces peuvent être représentées en quantités déséquilibrées dans les données de séquençage. Les assembleurs traditionnels utilisent l'information de la couverture du génome pour identifier des répétitions et les erreurs de séquençage. En contexte métagénomique, où l'abondance des différentes espèces varie, et où des régions génomiques peuvent être partagées par plusieurs espèces, cette stratégie n'est plus valable. Ainsi, dans [Venter et al., 2004], les génomes les plus couverts ont été considérés comme des répétitions par l'assembleur Celera, et une étape préalable a été nécessaire pour mieux les assembler. Comme indiqué précédemment, dans le cas d'un assemblage par graphe de *De Bruijn*, toute séquence répétée de longueur supérieure à k interrompt les contigs. Cette situation se produit fréquemment en métagénomique, où des espèces apparentées et pouvant partager certains gènes sont présentes. Les régions répétées entre les génomes à assembler s'ajoutent aux répétitions à l'intérieur d'un génome, et complexifient l'assemblage.

Deuxièmement, le polymorphisme existant au sein des communautés bactériennes complique l'assemblage de plusieurs manières. De nombreux variants d'une même espèce microbienne peuvent être séquencés au sein d'un échantillon métagénomique. Ce polymorphisme n'est pas équitablement réparti le long des génomes, et il est difficile d'assembler conjointement les régions conservées et caractéristiques de souches différentes. D'éventuelles variations structurales au sein de la population compliquent également la tâche.

Afin de résoudre ces problèmes, des algorithmes dédiés ont été développés pour l'assemblage *de novo* de données métagénomiques, tels que IDBA-UD [Peng et al., 2012], MetaVelvet[Namiki et al., 2012] ou MegaHit [Li et al., 2015]. Tous ces programmes présentent différentes particularités algorithmiques qui, en principe, permettent à la fois le passage à l'échelle sur de larges jeux de données métagénomiques, et l'assemblage de mélanges d'espèces. Par exemple, MegaHit emploie successivement des graphes de *De Bruijn* de tailles de kmers croissantes. Les plus petites valeurs de k permettent de mieux filtrer les erreurs de séquençage et d'assembler les régions les moins bien couvertes, tandis que les valeurs de k les

plus grandes permettent de résoudre certaines répétitions si la couverture le permet.

Limites de l'assemblage métagénomique

Malgré le développement de ces outils dédiés, le problème de l'assemblage métagénomique n'est pas résolu. Le challenge CAMI [Sczyrba et al., 2017] a permis de confronter différents outils sur des thématiques propres à la métagénomique, incluant l'assemblage. Les résultats illustrent la difficulté de cette tâche avec les méthodes actuelles. Dans ce concours, pour le jeu de données de "haute complexité", l'assemblage le plus long ne couvre que 70% de la communauté, au prix de près de 8000 erreurs dans l'alignement avec le génome de référence ciblé. La qualité des assemblages obtenus dépend tout d'abord de la couverture des génomes. Naturellement, les génomes les moins couverts sont difficilement assemblés. De manière moins intuitive, certaines assembleurs peinent à assembler des génomes très abondants. Une parade employée par certains assembleurs (tels que MegaHit [Li et al., 2015]) est d'utiliser différentes tailles de kmers, adaptées à des abondances différentes. En complément, la présence de génomes fortement apparentés dans la communauté rend difficile l'assemblage de ces espèces pour tous les outils considérés. Ce point semble particulièrement problématique, étant donné l'existence au sein de la plupart des communautés d'un continuum de diversité entre les individus.

Par ailleurs, la question de la pertinence de représenter l'assemblage d'un métagénome sous forme de séquences linéaires peut se poser. Des bulles créées par le polymorphisme ponctuel ou des branchement dus aux variations structurales sont présentes dans le graphe d'assemblage, mais absentes des contigs. Les assembleurs suppriment ces variations, en écrasant les bulles et en arrêtant les contigs lorsque des branchements se produisent. Le résultat est un ensemble de contigs, dont la séquence est un consensus de celles des organismes séquencés, et dans lequel les variations structurales ne sont pas représentées. Une tendance actuelle est d'aller au-delà de cette représentation linéaire d'un génome par des séquences au format FASTA. Les assemblages sont construits à partir de graphes, dont les régions linéaires sont extraites pour donner des contigs. Par rapport au graphe d'assemblage, les contigs généralement donnés en sortie du programme d'assemblage contiennent donc moins d'information. La figure 2 donne un exemple de la plus-value apportée par la représentation d'un assemblage sous forme de graphe. Ainsi, il est proposé de remplacer les génomes de référence linéaires par des graphes rendant compte des variations génomiques [Paten et al., 2017], et certains assembleurs tels que Spades [Bankevich et al., 2012] incluent dans leurs résultats un graphe au format GFA où la diversité structurale des génomes peut être observée.

Finalement, il est à ce jour impossible d'assembler complétement et fidèlement les organismes d'un métagénome, et de rendre compte de la diversité génomique dans ces communautés. Les assembleurs retournent des contigs les plus longs possible, tout en évitant de construire des chimères en assemblant des lectures issues de différents organismes. La diversité intra-spécifique est le plus souvent ignorée, de manière à retourner un consensus des génomes présents. L'assemblage n'est donc pas suffisant pour caractériser la diversité

métagénomique d'un échantillon, et des outils complémentaires sont nécessaires pour retrouver les contigs issus d'une même espèce.

Binning de séquences métagénomiques

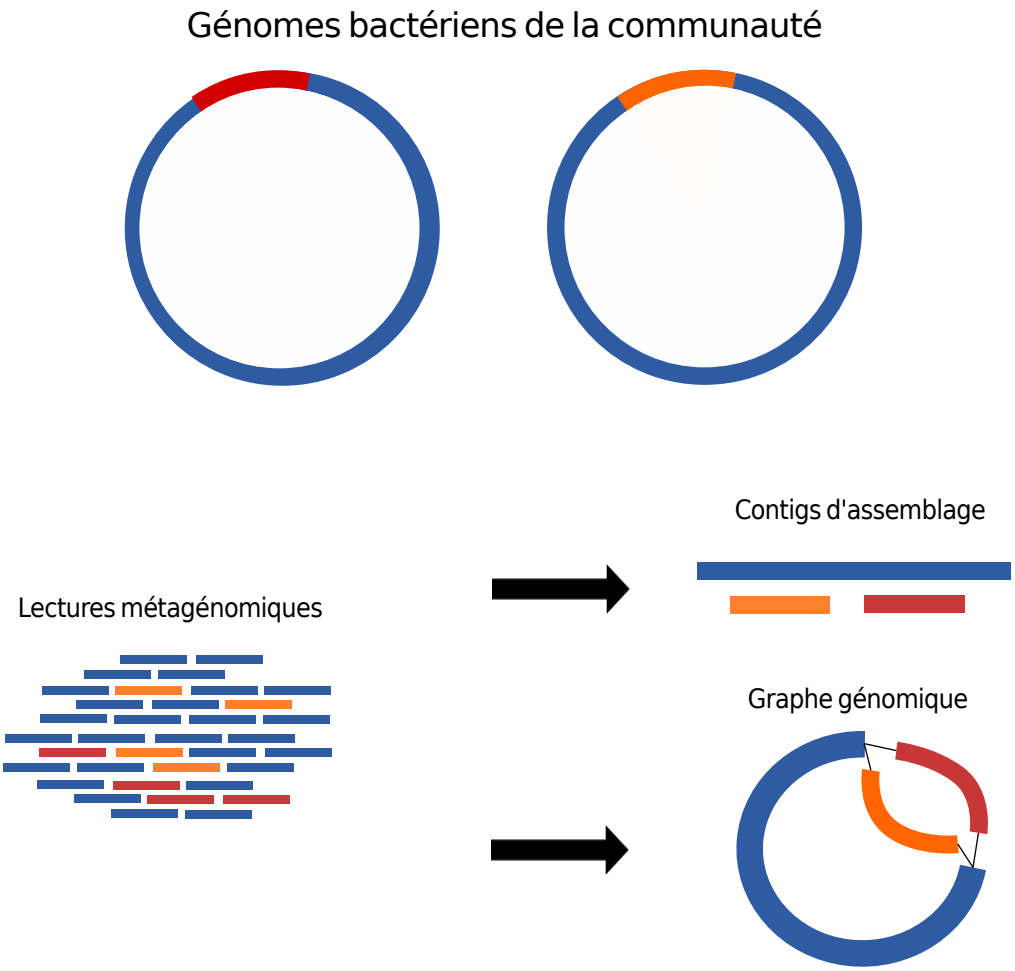
Les méthodes de binning ont pour but de regrouper des séquences de même origine taxonomique. Elles prennent en entrée des contigs préalablement assemblés, et les placent dans des clusters en fonction de leur origine taxonomique supposée. Étant donné qu'aucune information de référence n'est utilisée, ces bins ne sont pas associés à un taxon, mais ils peuvent ensuite être assemblés relativement facilement par des méthodes classiques, ce qui permet en théorie de reconstituer les génomes complets des membres de la communauté.

Une première famille de méthodes de binning utilise le contenu nucléotidique de la séquence. La première méthode de ce type est TETRA [Teeling et al., 2004], qui calcule pour chaque séquence des profils de tétranucléotides et la corrélation entre ces profils, ce qui permet de regrouper les lectures provenant d'organismes proches. La limite majeure des méthodes basées sur le contenu nucléotidique est qu'elles ne s'appliquent qu'à des fragments génomiques de grande taille (près de 10 kilobases), dans lesquels le contenu nucléotidique est représentatif de celui du génome entier. Elles ne peuvent donc s'appliquer qu'à des contigs préalablement assemblés et de grande taille, ce qui limite fortement leur intérêt. En plus de la composition nucléotidique, il est possible de classifier des contigs en fonction de leur couverture, selon l'hypothèse que des contigs avec des couvertures similaires proviennent des mêmes génomes. Certaines méthodes utilisent la couverture *différentielle* entre des jeux de données provenant par exemple d'endroits différents. L'idée sous-jacente est que des séquences provenant du même organisme, en plus d'avoir des niveaux de couverture similaires dans un échantillon, auront une couverture qui covarie de la même manière au sein de plusieurs échantillons. Les outils utilisant à la fois l'information de couverture et de composition obtiennent généralement les meilleures performances dans le binning de contigs. Ce sont ces outils qui se sont imposés comme les plus performants pour le binning de séquences. On peut citer parmi les principaux logiciels Concoct [Alneberg et al., 2014a], GroopM [Imelfort et al., 2014] ou MaxBin [Wu et al., 2014]. Ces multiples alternatives retournent parfois des résultats différents, ce qui a encouragé le développement d'outils pour la comparaison visuelle de leurs résultats (VizBin [Laczny et al., 2015]), ou de validation de la qualité des bins (CheckM [Parks et al., 2015]).

La principale limite rencontrée par ces techniques est qu'elles regroupent des contigs, issus d'un assemblage préalable. Dans ces contigs, la majorité de la diversité intra-spécifique a été éliminée par l'assembleur. Il est donc difficile, voire impossible pour ces méthodes de caractériser des variations génomiques fines, comme la présence de différentes souches bactériennes. Par ailleurs, le binning est sensible aux éventuelles erreurs lors de l'assemblage, comme la création de contigs chimériques. Enfin, certaines particularités génomiques rendent le binning difficile : une région génomique particulièrement riche en variants pour une souche particulière peut par exemple être affectée à un nouveau cluster différent du reste du génome.

Peu d'outils proposent un binning au niveau des lectures, car elles sont trop courtes

Figure 2 – *Illustration de l'intérêt des graphes d'assemblage.* *Dans cet exemple simple, la communauté à assembler contient deux bactéries contenant une large région variable. L'assemblage sous forme de contigs retourne trois contigs distincts, tandis que l'assemblage sous forme de graphe rend compte de la diversité structurale dans cette communauté.*



pour apporter une signature génomique fiable, et trop nombreuses pour proposer des outils passant à l'échelle sur de grands jeux de données. LSA (Latent Strain Analysis) [Cleary et al., 2015] est un outil original permettant le binning de lectures métagénomique, qui permet un assemblage indépendant de chaque bin. La méthode repose sur la comparaison de profils d'abondance de kmers au sein de plusieurs échantillons. À partir d'une matrice donnant l'abondance de chaque kmer dans chaque jeu de données, LSA procède à une décomposition en valeurs singulières (SVD) qui permet de sélectionner des kmers montrant le même profil de covariance, et provenant vraisemblablement du même génome. Une étape de clustering permet ensuite de construire des ensembles de lectures associées. Les bins ainsi reconstruits peuvent permettre d'isoler des souches distinctes, ce qui correspond à une résolution rarement atteinte par les méthodes de binning de contigs.

2.2 Métagénomique fonctionnelle

La métagénomique shotgun offre la possibilité d'accéder aux génomes entiers des membres de communautés microbiennes. Bien qu'il soit difficile de reconstruire des génomes complets avec les méthodes actuelles, les méthodes de binning et/ou d'assemblage présentées précédemment permettent de reconstruire des séquences suffisamment longues pour accéder au contenu génique des métagénomes. L'analyse de la diversité fonctionnelle se fait en plusieurs étapes. Tout d'abord, la prédiction de gènes permet d'identifier dans les contigs les séquences codant pour une protéine. L'annotation fonctionnelle permet ensuite d'associer chaque gène à une fonction. Enfin, l'ensemble des fonctions détectées dans un métagénome peuvent être mises en relation pour reconstituer des voies métaboliques.

2.2.1 Prédiction et annotation de gènes

La prédiction de séquences codantes peut se faire par comparaison à des bases de données protéiques, ou bien *ab initio*, ce qui permet de détecter de nouveaux gènes. Dans le premier cas, les séquences nucléotidiques sont converties en séquences protéiques suivant les 6 sens de lectures possibles, puis comparées à des bases de données de protéines connues. Ce type d'analyse est par exemple permis par l'outil BlastX [Altschul et al., 1990]. Étant donné l'incomplétude des bases de données de référence, les méthodes *ab initio* sont privilégiées en métagénomique. Différents outils ont été développés expressément pour cet usage, tels que MeteGeneMark [Zhu et al., 2010] ou Orphelia [Hoff et al., 2009]. Les difficultés liées à la métagénomique sont essentiellement dues à la longueur généralement plus courte des contigs par rapport aux assemblages d'une seule espèce. La plupart des techniques reposent sur des modèles (HMM, machine learning) entrainés à partir de génomes connus.

L'étape suivante consiste à assigner aux gènes détectés des fonctions métaboliques. Le processus est analogue à celui réalisé en génomique classique, avec un nombre de gènes généralement très supérieur. Il s'agit de comparer les séquences protéiques obtenues à l'étape précédente avec des bases de données protéiques. Celles-ci sont nombreuses, on peut citer

KEGG [Kanehisa, 2004], PFAM [Finn et al., 2014] ou Uniprot [Bateman et al., 2017]. Ces différentes bases de données ne couvrent pas toutes les fonctions connues, et des outils tels qu'InterPro [McDowall and Hunter, 2011] ou MG-RAST [Keegan et al., 2016] permettent d'en exploiter plusieurs.

2.2.2 Reconstruction de réseaux métaboliques

L'une des finalités de l'annotation des gènes est la reconstruction de réseaux métaboliques. L'outil MetaPath [Liu and Pop, 2010] aligne sur un grand réseau métabolique des lectures métagénomiques afin d'identifier quelles en sont les composantes présentes dans un jeu de données. L'analyse fonctionnelle de séquençages métagénomiques peut par exemple permettre l'identification de nouvelles voies métaboliques, qui peuvent être utiles à l'hôte de communautés symbiotiques [Cecchini et al., 2013]. Dans le cas d'échantillons environnementaux, il est possible d'identifier les composés pour lesquels un organisme dépend de son environnement [Borenstein et al., 2008]. En comparant les topologies des réseaux métaboliques de plusieurs espèces vivant dans le même environnement, il est possible d'identifier des coopérations [Levy et al., 2015] ou des compétitions [Kreimer et al., 2012] au sein d'une communauté.

Les données métagénomiques offrent ainsi la possibilité d'accéder au fonctionnement de communautés complexes. Cette étape intervient cependant après de nombreuses autres analyses, et peut souffrir d'erreurs survenues au cours de l'assemblage, de l'assignation taxonomique ou de l'annotation, *a fortiori* dans le cas d'organismes peu connus.

2.3 Métagénomique comparative

La métagénomique comparative consiste à comparer des ensembles d'échantillons métagénomiques, tels que des séries temporelles ou des échantillons collectés à des endroits différents.

Un premier ensemble des méthodes de métagénomique comparative s'inscrit dans la continuité des méthodes d'assignation taxonomique. Ces méthodes calculent une distance entre des ensembles de taxons ou OTUs associés à des abondances. La distance calculée dépend d'un indice de similarité, qui peut prendre en compte ou non l'abondance de chaque groupe. On peut citer parmi les indices quantitatifs la distance de Bray-Curtis et parmi les indices qualitatifs la distance de Jaccard. Ces distances sont implémentées dans certains logiciels d'analyse de données métagénomiques tels que MEGAN [Huson et al., 2016], qui proposent également des visualisations de données multidimensionnelles (par exemple par PCoA).

Étant donné les difficultés à établir une bon inventaire taxonomique de métagénomes, une alternative est de comparer les échantillons par leurs séquences, sans tenter de les assigner dans des OTUs. On parle alors de métagénomique comparative *de novo*. Ces outils utilisent les kmers présents dans les différents échantillons pour calculer divers indices de similarité. Par

les algorithmes et les structures de données utilisés, ils réussissent à comparer très rapidement de grands jeux de données. On peut citer parmi ces méthodes Simka [Benoit et al., 2016] et MASH [Ondov et al., 2016].

2.4 Conclusions

Cette revue des méthodes existantes pour l'analyse des données métagénomiques permet d'identifier celles qui seraient à même de répondre aux problématiques posées par cette thèse. Un des enseignements qui ressort de ce travail est que l'essentiel des méthodes développées sur la métagénomique vise à caractériser la diversité microbienne avec une résolution relativement grossière, typiquement en identifiant ou caractérisant des espèces bactériennes. Cela s'explique en partie par le fait que les communautés microbiennes sont souvent à la fois mal connues et très complexes. L'assemblage métagénomique, qui est une démarche fréquemment employée, est limité dans la mesure où il est encore très difficile d'assembler les génomes complets des membres d'une communauté microbienne.

Le modèle de l'holobionte du puceron du pois est particulier dans le sens où la plupart des associés symbiotiques ont été caractérisés de manière plus ou moins complète. Un génome de référence est connu pour la majorité d'entre eux, et le symbiote obligatoire *Buchnera aphidicola* est particulièrement bien connu pour son rôle fonctionnel. Partant de ce constat, les méthodes *de novo* semblent avoir peu d'informations supplémentaires à apporter, et ne tirent pas partie des génomes déjà assemblés. En incorporant une information extrinsèque, sous la forme par exemple de bases de données de gènes ou de génomes de référence, il est possible d'atteindre une toute autre résolution, et de caractériser plus finement le contenu voire le potentiel fonctionnel de ces systèmes. La thématique de la métagénomique de souches bactériennes est très active actuellement, mais reste restreinte à des communautés modèles comme le microbiome humain, et nécessite de s'intéresser à des microorganismes déjà connus [Segata, 2018].

Au cours de cette thèse, nous avons donc cherché à employer et développer des méthodes tirant profit des génomes de référence pour décrire plus finement la diversité génomique et l'histoire évolutive des symbiotes du puceron du pois. Dans le chapitre 3, nous utiliserons à la fois l'alignement de lectures sur ces références et la recherche de variants pour caractériser précisément le génotype des symbiotes. Puis dans le chapitre 4, nous présenterons une méthode nouvellement développée pour corriger des génomes de référence existants mais éloignés, en prenant en compte d'éventuelles variations structurales.

Chapitre 3

Caractérisation multi-échelle de la diversité symbiotique dans le complexe du puceron du pois par une approche métagénomique

3.1 Présentation de la publication

Le puceron du pois est un modèle d'étude des relations symbiotiques, mais la diversité génomique et l'histoire évolutive des symbiotes qui lui sont associés restent mal connus. En particulier, peu d'études ont été menées à des échelles taxonomiques fines, que ce soit pour expliquer l'histoire évolutive à l'intérieur de l'espèce du puceron du pois, ou pour mieux caractériser des variants génomiques de symbiotes. Dans ce chapitre, qui est constitué d'une publication acceptée dans *Microbiome* [Guyomar et al., 2018], nous montrons comment l'information plein-génome apportée par la métagénomique peut répondre à de telles questions.

À partir d'un large jeu de données de reséquençage de pucerons du pois à travers différents biotypes, nous avons mis en place une approche qui repose sur l'alignement sur des génomes de référence, qui ont été assemblés si nécessaire, et la recherche de variants. Cette approche permet d'exploiter pleinement les données de métagénomique, et se veut généralisables à d'autres modèles. L'étape d'assemblage et d'alignement a permis de confirmer les connaissances préalables sur les espèces impliquées dans le complexe du puceron du pois, en montrant que ce système est dominé par quelques symbiotes déjà identifiés. La comparaison de profils de variants génomiques sur tout le génome et pour de multiples échantillons a ensuite révélé une hétérogénéité de la diversité intra-spécifique. Des études phylogénétiques ont mené à différents scénarios pour l'association entre l'hôte et chacun de ses symbiotes. Dans l'ensemble, les résultats indiquent de fréquents transferts horizontaux au sein du complexe. Les dynamiques évolutives diffèrent également entre les symbiotes, ce qui est révélateur d'histoires ou de contraintes évolutives différentes. Enfin, l'analyse attentive des données métagénomiques a montré que deux souches différentes d'un même symbiote pouvaient coexister dans un hôte. mené à des scénarios évolutifs pour les différents symbiotes du complexe.

Dans ces travaux, nous présentons une démarche qui permet d'interroger la diversité

microbienne d'holobiontes à partir de données métagénomiques, et qui a généré en des résultats importants sur le modèle du puceron du pois. Finalement, cet holobionte qui semble simple du point de vue des espèces qui y sont présentes, recèle une diversité importante et révélatrice de son histoire. Ces travaux ouvrent la voie à une meilleure compréhension des relations entre l'hôte et son symbiote, et à la description d'autres holobiontes par des techniques similaires.

RESEARCH

Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches

Cervin Guyomar[1,2], Fabrice Legeai[1,2], Emmanuelle Jousselin[3], Christophe Mougel[1], Claire Lemaitre[2] and Jean-Christophe Simon[1]

Abstract

Background: Most metazoans are involved in durable relationships with microbes which can take several forms, from mutualism to parasitism. The advances of NGS technologies and bioinformatics tools have opened opportunities to shed light on this hidden but very influential diversity. The pea aphid is a model insect system for symbiont studies, and is organized in a complex of biotypes, each adapted to a specific host plant. It harbours both an obligatory symbiont supplying key nutrients and facultative symbionts bringing some novel functions to the host, such as protection against biotic and abiotic stresses. Little is known on how the symbiotic genomic diversity is structured at different scales: across host biotypes, amongst individuals of the same biotype, or within individual aphids; which limits our understanding on how these multi-partner symbioses evolve and function.

Results: We present a framework well adapted to the study of genomic diversity and evolutionary dynamics of the pea aphid holobiont from metagenomic read sets, based on mapping to reference genomes and whole genome variant calling. Our results revealed that the pea aphid microbiota is dominated by a few heritable bacterial symbionts reported in earlier works, with no discovery of new microbial associates. However, we detected a large and heterogeneous genotypic diversity associated with the different symbionts of the pea aphid. Partitioning analysis showed that this fine resolution diversity is distributed across the three considered scales. Thorough phylogenetic analyses highlighted frequent horizontal transfers of facultative symbionts between host lineages, indicative of flexible associations between the pea aphid and its microbiota. However, the evolutionary dynamics of symbiotic associations strongly varied depending on the symbiont, reflecting different histories and possible constraints. In addition, at the finest intra-host scale, we showed that different symbiont strains may coexist inside the same aphid host.

Conclusions: We present a methodological framework for the detailed analysis of NGS data from microbial communities of moderate complexity, and gave major insights into the extent of diversity in pea aphid-symbiont associations and the range of evolutionary trajectories they could take.

Keywords: host-microbiota interactions; aphids; metagenomics; symbiosis; phylogeny

1 Background

Symbioses have been studied for long in the case of simple binary interactions between a host and a single symbiont. Many studies have unveiled the functional impacts and the evolutionary consequences of these symbioses including acquisition of novel functions, transmission patterns ([1, 2]), genomic changes ([3]), reproductive manipulations (reviewed in ([4]), or cost/benefit balance of symbiotic relationships ([5, 6]). Yet, the advances of molecular techniques in the last decades have revolutionized the description and our understanding of host-microbe interactions, and revealed that every plant or animal is interacting in some way with multiple microbes ([7]). Biology is undergoing a paradigm shift where individual phenotypes should be considered as resulting from the combined expression of the host and associated microbe genomes (metagenomes) ([8]). As a reflection of this conceptual shift, the term "holobiont" is now used to name the complex ecosystem of a host and its community of associated organisms ([9, 10]). Similarly, the

Full list of author information is available at the end of the article
*Equal contributor

term "hologenome" is used to describe the collection of genomes of a host and its microbiota ([11]). A prerequisite to understand the functional, ecological and evolutionary implications of host-microbiota associations for holobionts is to evaluate the extent and partitioning of diversity at different scales involving individuals and populations of holobionts. This can be obtained from i) a full inventory of the microbial entities associated with the host, including transient low abundant symbionts, and ii) a fine characterization of the genomic diversity of microbial partners both within and between individual hosts from different populations. Inter-individual host diversity is often ignored when pooling together several individuals, or underestimated by insufficient sampling in the population, and intra-host variability is rarely considered, but these two levels are essential to infer the evolutionary dynamics of host microbiota interactions ([12]) and to better link microbiota diversity with associated phenotypic changes in the host ([12]).

Next generation sequencing techniques can provide whole genome sequencing data of communities of organisms. Some host sequencing projects contain microbe-related reads that are often considered as "contaminant" in the analysis of the host genome. These datasets can actually be analyzed and provide meaningful insights about organisms seen as holobionts. Shotgun metagenomic sequencing has several features which enables high resolution analysis of taxonomic and genetic diversity associated with holobionts. First, because it is a without a priori technique, it can capture all of the microbial diversity in environmental or host samples, including unknown bacteria, viruses or eukaryotic symbionts. Secondly, it provides whole genome information, which enables to detect genetic variation at a fine scale, and therefore offers the potential to track the evolutionary history of the holobiont partners, including acquisition source, and gain-loss dynamics of microbial diversity. One criticism on metagenomic studies investigating the genetic diversity associated with holobionts, is that most of the current phylogenetic analyses using the bacterial 16S ribosomal RNA gene are led at a coarse scale. They cannot assess accurately the specificity of the association between a host and its symbionts because bacteria with similar 16S rRNA (usually above 97% sequence identity) can have substantial differences on the rest of the genome and therefore have different impact on their host phenotypes [13]. Whole genome metagenomic sequencing allows investigating fine-scale diversity and yields robust phylogenetic information. Moreover, the whole genome information can be used to explore the phenotypic effects of symbiotic communities by using gene annotations and reconstructing holobiont metabolic networks ([14]).

Over the last decades, numerous computational methods have been developed to improve the analysis of metagenomic reads. These bioinformatics developments can be grouped into two main approaches: de novo genome assembly and metagenomic sequence profiling, that is the grouping of sequences from one or several metagenomes into groups of the same taxonomical origin. Both of these approaches have been mainly applied to examine diversity at the species-level. If tremendous progress has been achieved in de novo metagenomics assembly ([15]), the inherent goal remains to build a set of consensus sequences representing the actual species in the metagenomics sample, and polymorphism information is usually discarded, preventing the recovery of strain-level genomic variations ([16, 17]). On the other hand, metagenomic profiling when based on reference databases is either restricted to few marker genes ([18, 19]) or can perform strain level assignation only for model systems or very well studied organisms for which many strains are already characterized (for instance for biomedically important pathogens [20, 21]. Finally, reference-free metagenomic profiling approaches, also called binning approaches, are often based on previous assemblies that have already discarded polymorphism information [22, 23], or, when using co-abundance signals, may lead to incorrect binning when conserved and variable regions of a same species are sorted in different bins [24].

Overall, one of the main pitfalls of current holobiont analyses is the characterization of microbes at strain/genotype level. Apart from model communities for which comprehensive strain databases are available, fine variations in symbiont genomes are not accurately addressed by the current metagenomics-dedicated methods. Then, a basic but efficient strategy consists in converting the problem into several non-metagenomic ones, namely analyzing each symbiont and its corresponding read subsets independently using classical genomic variation methods. The major difficulty remains to be able to partition unambiguously the read datasets, and this is definitely easier when disposing of good reference genomes for all the symbionts. In the present paper, we present a framework designed to recover strain-level genomic variations from metagenomic reads preliminary mapped on reference genomes. When a given symbiont lacks a good reference genome, it is then built de novo from the metagenomic datasets.

To assess the potential value offered by this framework, we applied it to a biological system of moderate complexity regarding microbial communities, and with good prior knowledge of the expected symbiotic diversity. The pea aphid Acyrthosiphon pisum is a model species for insect symbioses, and shows several features which make it relevant for studying the forces structuring microbial diversity in holobionts. Pea aphids shelter an obligate bacterial symbiont, *Buchnera aphidicola* which provides

the host with essential amino acids absent or scarce in the insect diet (i.e. phloem sap [25]. In addition, several secondary symbionts are commonly found in pea aphid populations at different frequencies. Some of these secondary symbionts have been shown to provide ecological advantages to their hosts, for example by increasing protection against natural enemies or by conferring thermal tolerance [26]. While the primary symbiont is strictly maternally inherited [27], secondary symbionts are vertically transmitted with a lower fidelity, and can be horizontally transmitted [28], but neither the mechanisms nor the magnitude of these events of horizontal transfers are fully understood [29]. The pea aphid actually forms a complex of at least 15 biotypes, each biotype being adapted to a specific set of host plants [30]. Estimates of divergence time between biotypes suggest that this complex may have diversified 5,000-10,000 years ago, which coincides with the onset of plant domestication for agriculture [31]. Population genetic analyses revealed that these biotypes form a continuum of divergence, with partially isolated host races and reproductively isolated cryptic species ([32]. Several studies revealed that pea aphid biotypes also differ in their composition and frequency of secondary symbionts, but secondary symbionts seem to contribute very little to plant specialization of their hosts ([33, 34, 35, 30]. In addition, strain variation has been characterized in some secondary symbionts infecting the pea aphid complex [34], and found in some cases associated with large phenotypic differences in their hosts [36, 29]. Overall, the available literature on the pea aphid symbionts indicates large variation across host populations, both in bacterial species and strains, with important functional, ecological and evolutionary impacts on pea aphid holobionts. Although there have been recent attempts to uncover the bacterial communities associated with the pea aphid complex with deep sequencing of 16S ribosomal RNA [33, 37], no study has been yet conducted to fully characterize the diversity of pea aphid microbiota notably at different scales of organization, and at a whole genome scale. The pea aphid appears to be a relevant system to develop a metagenomic framework applied to the analysis of microbial diversity and structure in holobionts. It is located at a sweet spot of complexity, with a symbiotic community of moderate size and with various modes of transmission of symbionts between hosts. It offers an interesting case of diversity partitioning between host populations through genetically and ecologically differentiated biotypes, and it is a species for which ample genomic resources are available for both the host and its associated symbionts.

In this paper, we analysed metagenomic data from a large dataset of pea aphid resequenced genomes to explore the extent and partitioning of microbial diversity at the different scales presented above. By mapping the reads on a set of reference genomes, we assigned the majority of the reads to microbial taxa associated with the pea aphid complex. This enabled a high resolution inventory of the genomic diversity of bacterial symbionts found in the pea aphid complex. Variant calling and phylogenetic approaches on the whole set of symbiotic bacteria revealed contrasted levels of genomic variability and various transmission patterns between symbionts, presumably resulting from different evolutionary histories and ecologies of host-symbiont associations.

2 Methods

2.1 Biological samples

Pea aphids were collected on different plants of the *Fabaceae* family mainly in eastern France where host plant diversity is high, but also in southern and western France (Additional file 1). Individuals were sampled as parthenogenetic (clonal) females and brought to the laboratory to initiate individual clonal lineages. After at least two generations of culture on broad bean *Vicia faba* (a plant on which all pea aphid biotypes can feed [38]), DNA was extracted from each clone in order to i) genotype them with several polymorphic microsatellite markers, ii) detect repeated genotypes (i.e. individuals having the same multilocus genotypes and thus presumably belonging to the same clone) and remove them from further analyses to keep a single copy per genotype, and iii) check biotype membership of each lineage through assignment tests (see [39] for further details). Briefly, individuals with a membership equal or larger than 90% in the genetic cluster corresponding to their assigned biotype were selected for further sequencing scheme. In this study, 14 biotypes out of the 15 described for the pea aphid complex were each represented either by single or pooled individuals. Thirty two individual resequenced genotypes encompassing 11 biotypes were those already used in [40]. This study also includes 18 new samples corresponding to pools of 14 to 35 individuals, each with a distinct multilocus genotype but belonging to the same biotype following assignment tests, representing overall 12 biotypes. Overall, the 50 samples used in this study are described in Additional file 1. Since these samples were composed of clones reared in the laboratory for at least two generations prior to DNA extraction for sequencing, their microbiota was largely composed of the heritable fraction, which was the focus of our study.

The DNA of the aphids and their microbiota was extracted using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) following the manufacturer instructions, and sequenced in paired end using Illumina HiSeq 2000 instruments, resulting in 2*100 bp reads with a mean insert size of 250 bp. The average read depth for the pea aphid genome was 15X for individual sequencing (42.5 million reads on average) and ranged from 20

to 50X for pool sequencing (197.5 million reads on average). FastQC files were generated for each sample, and no anomaly in the sequencing data was observed. The FastQ files of the paired reads from the 50 samples are stored and publicly available at the Sequence Read Archive of the National Center for Biotechnology Information database, under the BioProject IDs PRJNA255937, PRJNA385905 and PRJNA454786.

2.2 Bioinformatics analyses

Full details on the analysis presented in the following parts are available on the website https://aphid-microbiome.netlify.com. This includes the source code of every custom script used during the analyses. Mapping based disentanglement of holobiont genomes Sequencing of both host and microbial DNA produces metagenomic datasets, containing reads originating from different organisms. This metagenomic context was dealt with by mapping read sets using BWA-MEM [41] with default parameters against a set of reference genomes, including the pea aphid nuclear and mitochondrial genomes, the primary symbiont genome (*Buchnera aphidicola*), and genomes of known pea aphid secondary symbionts, when available. This was the case for *Hamiltonella defensa 5A*, *Serratia symbiotica Tucson*, *Rickettsiella viridis* and *Regiella insecticola 5.15*. For the *Rickettsia* symbiont, no closely related reference genome was available and we produced our own reference genome by de novo assembly, as explained in the paragraph below. For *Spiroplasma*, we used a draft genome previously assembled from unmapped reads of a particular pea aphid sample, as described in [40]. For *Fukatsuia symbiotica*, we used the draft genome sequenced from the conifer aphid *Cinara confinis* [42, 43]. In addition, we included in the reference set the variant genomes of the phage APSE of *H. defensa* [44] and several plasmid sequences associated to symbionts detected in the pea aphid. In particular, we added three *Rickettsia* plasmid sequences from other insects in order to map *Rickettsia* plasmidic reads in the absence of a reference sequence for *A. pisum*. After the mapping step, several statistics were computed, including the mapping rate, the average coverage for each genome, the fraction of the reference genome covered by at least five reads, and the mean edit distance for the reads mapping on each reference genome. Reads associated to each symbiont were extracted using Samtools [45], and all downstream analyses were conducted independently and with the same settings for each symbiont. The reference genomes used for this step are summarised in Table 1. Additional statistics on the genomes used are available in Additional file 2.

2.3 Assembly of *Rickettsia sp.* genome

Using the results of a previous mapping of pea aphid reads on the genome of *Rickettsia bellii*, we identified two samples from the *Pisum sativum* biotype with high *Rickettsia* coverage (Ps_ind1 and Ps_ind2). These two samples were pooled together, resulting in a 100X coverage on the genome of *R. bellii*. Reads that mapped on the pea aphid genome were filtered out, and the remaining ones were assembled using SPAdes version 3.11.1 [55], with default parameters. Contigs with blast matches on *Rickettsia bellii* and *Rickettsia sp MEAM1* were extracted. To increase contiguity and genome completeness, some pairs of contigs were bridged together using the gapfiller MindTheGap [56] that performs local assembly using the whole read set.

The resulting assembly was 1,070,000 bp long (for comparison, *R. bellii* is 1.5 Mb long and *Rickettsia sp.* strain MEAM1 is 1.24 Mb), in 327 contigs, and had a N50 of 4,483 bp. 82.4% of complete genes were found using Busco v3.0.1 and the bacteria_odb9 gene set, which is very close to the 83.7% obtained for the reference genome of *Rickettsia bellii*. Compared to *Rickettsia bellii*, we observed a major improvement of the genome coverage as 84% more reads mapped on the newly assembled genome across the whole dataset.

2.4 Analysis and taxonomic assignation of unmapped reads

Unmapped reads were extracted using Samtools [45], and low quality reads were removed using Trimmomatic [57] with the following parameters: LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. Remaining unmapped reads were taxonomically assigned using Centrifuge [58]. Only assignation hits larger than 40 base pairs were kept. Results were visualized using the Pavian R package [59].

2.5 Genome wide variant calling

Variant calling was performed for the whole set of symbionts identified in the pea aphid samples (*B. aphidicola*, *H. defensa*, *R. insecticola*, *S. symbiotica*, *Rickettsia sp.*, *Spiroplasma sp.*, *F. symbiotica* and *R. viridis*). It was also performed on the pea aphid mitochondrial genome, in order to capture the host matriline diversity. By essence, secondary symbionts were not present and equally abundant in all the samples, and a minimal coverage was required to run variant calling. Only symbionts with more than 10X sequencing depth and a homogeneous coverage along the genome were kept for this analysis. For instance, two symbionts in five samples were discarded because more than 90% of genomic positions were covered by less than two reads. This metric was smaller than 30% in the remaining samples.

Samtools mpileup [45] was used with options "-t DP,DPR" on the alignments to detect both SNPs and indels, and the coverage of the different alleles was reported. The generated bcf file was processed using

Table 1 Summary of reference genomes used for mapping

Organism name	Sequence ID	Accession	Reference
Acyrthosiphon pisum	Genome	SAMN00000061	[46]
Buchnera aphidicola	Genome APS	BA000003.2	[47]
	Plasmid pLeu	AP001071.1	[47]
	Plasmid pTrp	AP001070.1	[47]
Hamiltonella defensa 5AT	Genome 5AT	CP001277.1	[48]
	Plasmid pHD5AT	CP001278.1	[48]
	Phage APSE 1	AF157835.1	[49]
	Phage APSE3	EU794053.1	[44]
	Phage APSE4	EU794051.1	[44]
	Phage APSE5	EU794050.1	[44]
	Phage APSE6	EU794054.1	[44]
	Phage APSE7	EU794052.1	[44]
Regiella insecticola 5.15	Genome	AGCA01000000	[50]
	Plasmid pRILSR1	CM000957.1	[51]
Serratia symbiotica strain Tucson	Genome	GCA_000186485.2	[52]
Spiroplasma sp.	Genome	Upon request	[40]
Candidatus Fukatsuia symbiotica	Genome	GCA_900128755.1	[42]
Rickettsiella viridis	Genome		[53]
Rickettsia sp.	Genome	Upon request	This paper
	plasmid pREIS3	CM000771.1	
	plasmid pRF	GQ329881.1	
	plasmid pRAF	CP001613.1	
Wolbachia sp. wRi	Genome	GCA_000022285.1	[54]

bcftools [60] with options "-mv -Ov". Abundance tables of reference and alternative alleles for each polymorphic site and for each sample were extracted for further filtering using vcftools [61] and processed using a custom R script (available on the https://aphid-microbiome.netlify.com). In order to remove false positive variants due to sequencing errors, rare variants were removed by applying two coverage filters: for each sample, variants covered by less than four reads, or with less than 10% frequency were removed. Regions with exceptionally high or low coverage were excluded from the analysis. Genomic positions were considered of low coverage when at least 75% of samples had a coverage inferior to the median coverage of all variants along the genome. Similarly, high coverage genomic positions were discarded when the coverage was at least five times superior to the median coverage for at least 75% of the samples. In addition, for closely related reference genomes, such as *R. insecticola*, *H. defensa* and *F. symbiotica*, homologous genomic regions were detected by performing a pairwise blast search, and regions with a homology greater than 80% were excluded.

2.6 Phylogenetic inference

Variant frequencies were used to compute the variant profile of each sample by selecting the most abundant allele at each site. In the case of equally covered alleles, the reference allele was kept. This situation made it difficult to determine the most abundant genotype in the sample, but was rare in our dataset. We therefore decided to remove from the analysis samples in which more than 5% of variable sites yielded alleles with equal abundances. It was the case for three pool sequencing samples with low symbiotic coverage.

To investigate the evolutionary relationships between the genomes of the different samples, a phylogenomic analysis on a set of genes encoding membrane proteins was performed when an annotated reference genome was available. We first selected a list of genes, in order to compute the putative sequences for these genes in all samples. The Uniprot database was queried to retrieve DNA sequences of membrane protein transcripts (under the "Cell membrane" keyword) for the different studied symbionts (the complete list of genes used can be found in Additional file 3). Membrane proteins were selected as they are assumed to show a higher mutation rate than usual phylogenetic markers [62], and therefore are more appropriate to capture recent phylogenetic events. This query resulted in sets of 96, 118, 141 and 96 genes for *B. aphidicola*, *H. defensa*, *S. symbiotica* and *R. insecticola*, respectively. For each sample, the putative sequences of the selected proteins were inferred by replacing the reference alleles by

the alternative alleles associated to the different variant profiles.

The gene sequences of each selected protein were aligned using MAFFT [63] (v7.310, linsi mode) and the resulting multiple alignments were concatenated. The lengths of the alignments for the analyzed symbionts were 92,293 bp for *B. aphidicola*, 118,344 bp for *H. defensa*, 100,027 bp for *R. insecticola* and 144,360 bp for *S. symbiotica*. To validate that our alignments were not subject to substitution saturation, a Xia's test was run, as implemented in DAMBE6 [64]. Because most software of phylogenetic inference struggle to estimate branch lengths for identical sequences, we pre-processed our concatenated alignments by keeping only one sequence for each set of identical sequences. We used RaxML [65] (version 8.2.10, options -f a - 1000 -m GTRGAMMA), a phylogenetic inference program based on maximum likelihood method, to infer the phylogeny of the samples of the considered genes. The GTRGAMMA model was used with no partitioning of the data matrix, with 1,000 bootstrap iterations. Phylogenetic trees were edited and compared using functions of Ape [66], and Dendextend [67] R packages.

To cross-validate the phylogenetic relationships inferred on gene sets and also use the information contained in whole genome data, we used a clustering approach of whole genome variant profiles. Pairwise comparisons of variant profiles were performed; the numbers of differences between all pairs of profiles were then computed, and divided by the total number of variants detected on the genome, as implemented in the AW-clust algorithm proposed in [68]. The distance matrix was then used to perform Neighbor Joining clustering and build a phylogenetic tree based on whole genome variant profile information. Tree topologies were visually compared between the gene set and whole genome approaches. For *F. symbiotica*, *Rickettsia sp.*, *Rickettsiella viridis* and *Spiroplasma sp.*, we did not perform a gene-based phylogeny since their reference genomes are not well assembled nor annotated. In that case, neighbor joining was performed on whole genome variant profiles to infer phylogenetic relationships between samples.

Outgroups were used to root the phylogenetic trees. For *B. aphidicola*, we used sequencing data of two Japanese *A. pisum* lineages, known to be highly divergent from European lineages [31]. For other symbionts, we used close related symbiont species: *H. defensa* from the whitefly *Bemisia tabaci* (GenBank 2777848), *S. symbiotica* SCt-Vlc from the conifer aphid *Cinara tujafilina* (FR904230), *Spiroplasma melliferum* KC3 from *Apis mellifera* (GCA_000236085.3), *Rickettsia sp.* MEAM1 from *Bemisia tabaci* (GCA_002285905.1) and *Rickettsiella grylli* from crickets (GCA_000168295.1). For *R. insecticola*, the closest known symbiont was *F. symbiotica*, and reciprocally, the outgroup for *F. symbiotica* was *R. insecticola*.

2.7 Phylogenetic reconciliations

We used reconciliation analyses as implemented in Jane 3 [69] to infer cospeciation and host shifts events along the evolutionary history of each symbiont. The history of symbiotic relationships is commonly disclosed by comparing host mitochondrial phylogeny and symbiotic phylogeny. Many studies use phylogenetic congruence between these two types of genomes to elucidate patterns of symbiotic inheritance [70, 71]. However, achieving a high resolution in reconstructing host phylogenetic information for closely related lineages from mitochondrial DNA is challenging [27]. Since the primary endosymbiont *B. aphidicola* is known to be strictly maternally inherited [25], our strategy to overcome this limitation was to use its phylogeny as a proxy for the host mitochondrial phylogeny. *B. aphidicola* is known to have a high mutation rate [72] as highlighted in [32], and therefore appears to be a good indicator of the recent host history [70]. In reconciliation analyses, the parasite phylogeny (in our case, the secondary symbiont) is "mapped" onto the host phylogeny (i.e. each node in the parasite tree is assigned to a node in the host phylogeny). In such a map, the diversification events of the parasites are linked to their host phylogenetic history, so that four types of events are considered: cospeciation events, host switches, sorting events and duplication events. For the host phylogeny we used the matriline phylogeny inferred for *B. aphidicola* gene set data which showed a better resolution than the aphid mitochondrial phylogeny, and tested for each secondary symbiont whether primary and secondary symbiont phylogenies showed significant cospeciation (indicative of vertical transmission), using gene-based phylogeny for *S. symbiotica*, *H. defensa* and *R. insecticola*, and Neighbor Joining analysis of whole genome variants for *F. symbiotica*, *Spiroplasma sp.*, *Rickettsia sp.* and *Rickettsiella*. For each cospeciation analysis, we first pruned aphid samples for which the focal symbiont was detected but had insufficient read coverage to obtain reliable data for phylogenetic inferences (i.e. we did not consider the symbionts in a sample when their coverage was comprised between 1X and 10X), in order to avoid overestimating losses in the reconciliation process (i.e. considering that a symbiont was absent in an aphid sample while it was actually present but with insufficient data to perform a reliable variant calling). The focal symbiont was considered as absent when the coverage was inferior to 1X. We ultrametrised the host and symbiont trees using Grafen's method using Ape package in R. We then ran Jane 3 [69] with the number of "generations" (iterations of the algorithm) set to 100 and the "population" (number of samples per generation) set to 100 and used

default cost setting (cospeciation=0 and all other events =1). The cost of the best solution was compared to the distribution of the costs found in 500 randomizations in which the tip mappings were permuted at random. When the cost of the observed reconciliation is lower than expected by chance, the cospeciation signal is significant.

3 Results

3.1 Most of the microbiome diversity is captured by the mapping approach

On average, 90% of the reads were assigned by mapping to the pea aphid nuclear or mitochondrial genome. The nuclear genome average coverage was 13X for individual sequencing and 66X for pool sequencing. 5.62% of the reads mapped on the genome of *B. aphidicola* and its plasmids, with an average coverage of 628X for individual sequencing and 3,694X for pooled sequencing. The coverages for the different secondary symbionts were very diverse, and ranged from 0 (secondary symbiont was absent) to 1,300X (see Additional file 1). Presence and absence of symbionts as inferred from read depth was in agreement with the results of PCR diagnostic tests conducted for individual samples [40], and the few mismatches observed in the previous study were corrected by the choice of more appropriate reference sequences for *Rickettsia sp.*, *R. viridis* and *Spiroplasma sp.*

To further ensure that the used reference genomes were appropriate, we looked at the proportion of the genome covered by metagenomics reads, and the average edit distance of reads mapping on each symbiont genome (minimum number of editing operations between the read and the corresponding part of the reference genome). Overall, more than 97% of the genomic positions of our reference genomes were covered by at least five reads. For *F. symbiotica*, we also checked that the mean edit distance of mapped reads was not larger than that of other symbionts for which we had reference genomes or did a de novo assembly. Mean edit distance was 1.43 for *F. symbiotica* and ranged between 0.71 and 4.0 for other symbionts (average value was 1.67). Apparently, the use of a *F. symbiotica* genome assembled from another aphid host does not hamper the quality of the mapping.

Sequencing depth data are summarized in a presence/absence matrix, as seen in Figure 1 and are fully detailed in Additional file 1. Since only a few infected individual aphids were enough to enable the detection of a symbiont in a pooled sample, pooled data generally contained a higher richness in secondary symbionts (on average 4.28 secondary symbionts per sample for pooled samples compared to 1 for individual sequencing).

3.2 A low number of unmapped reads validates the mapping approach

A few reads did not map onto any reference genome. The average rate of reads that did not map after quality control was 0.82% (median: 0.62%, min: 0.25%, max: 4.76%). It confirms that mapping metagenomic reads on this set of reference genomes is able to capture most of the genomic diversity of the pea aphid complex. The unmapped rate was heterogeneous between samples, and appeared linked to the symbiotic composition of the samples. Samples infected by symbionts for which a draft reference genome was used for mapping (*Spiroplasma* and *Rickettsia*) contained more unmapped reads. These reads probably originate from genomic regions absent or too divergent from these draft reference genomes. When considering samples containing only symbionts with good quality and closely related genomes, the average unmapped rate lowered to 0.69%.

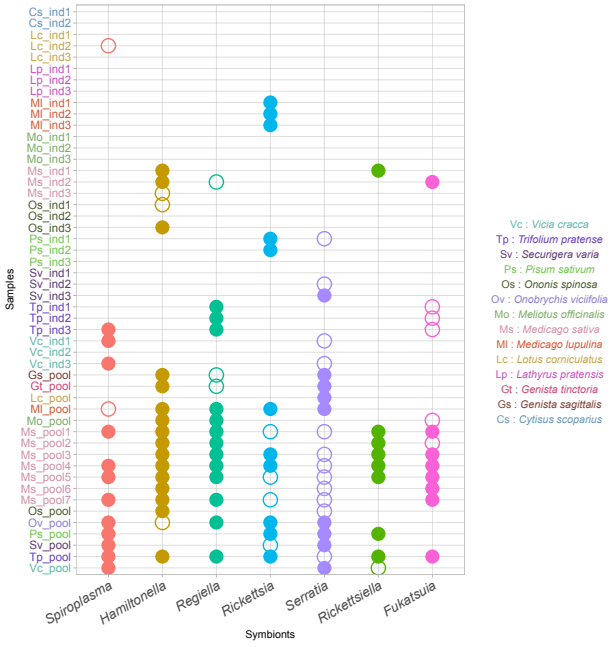
The nature of those unmapped reads was further explored by conducting a taxonomic assignation of such reads with Centrifuge (version 1.0.3) [58] and its default database. Overall, only 4.9% of the unmapped reads were assigned to a taxon. The taxonomic assignation of unmapped reads is summarized in Additional file 4, and can be explored for all samples on the website https://aphid-microbiome.netlify.com/. It is in accordance with mapping results. Some reads of host or symbiotic origins that were not mapped to the appropriate reference genome were however accurately assigned by Centrifuge. Other taxa were also found by Centrifuge assignation, either because of over-assignation by the program, or because some environmental organisms were sequenced along with the pea aphid and its symbionts. These reads represented a small fraction of the unmapped reads. Most unmapped reads were not taxonomically assigned by Centrifuge, probably because they contained sequencing errors, or were too distant to any reference sequence in the Centrifuge database. Overall, these results indicate that the microbiota of the pea aphid complex is dominated by a few heritable symbionts, and that we achieved a close to exhaustive inventory of the microbiome of our pea aphid samples.

3.3 Different levels of intra-specific diversity for the pea aphid symbionts

The overall genomic diversity of the selected samples was estimated for each symbiont by measuring the density of variable sites between the two most different symbiont genomes in the dataset. Only pooled samples were considered in this analysis, in order to have a more comparable sample size for each symbiont.

Variant calling results are summarized in Table 2. They show strong contrasts in genomic diversity between the different symbiont taxa associated with the pea aphid complex. *H. defensa* and *Regiella insecticola* showed the highest diversity, with 12.6 and 16.8 variants per kilobase (kb), respectively. Conversely, genomic diversity was extremely low for *Rickettsiella viridis*, with on average 0.027

Figure 1 Presence/absence pattern for bacterial symbionts as detected in the metagenomic dataset. Pea aphid individuals (ind) and populations (pool) were analyzed. Empty circles indicate a coverage greater than 1X. Filled circles indicate a coverage greater than 10X, enabling phylogenetic analysis. *A. pisum* and *Buchnera aphidicola* genomes were detected in every sample.



variants per kb. The other symbionts (*B. aphidicola*, *F. symbiotica*, *Spiroplasma sp.*, *Rickettsia sp.* and *S. symbiotica*) showed intermediate levels of genomic diversity (with respectively 3.0, 1.59, 1.28, 1.19 and 1.0 variants per kb). Consequently, the lengths of the branches of the phylogenetic trees built for these various symbionts were highly variable.

3.4 Phylogenomic analysis of *Buchnera aphidicola* from the pea aphid complex

By analysing genomic variation over the whole genome of *B. aphidicola*, we built a well-supported phylogeny of the pea aphid obligatory symbiont. No substitution saturation was detected using the Xia's test [64] (see Additional file 6). Fig. 2 shows the results of the phylogenomic analysis for *B. aphidicola* across all datasets, using maximum likelihood based inference on a 96 gene set alignment. The tree topology obtained from the gene set was compared with a whole genome variant profile clustering. Overall, the two phylogenetic methods gave similar results, as shown in Additional file 5. The few mismatches observed between the two topologies mainly involved nodes with low support in both trees.

As previously observed using partial sequences of pseudogenes data [32], *B. aphidicola* genomes associated with the pea aphid complex are separated into two distinct clades.

Matrilines from the same biotype were generally clustered together, but some were scattered across the phylogeny (e.g. *Vicia cracca* and *Ononis spinosa* biotypes did not form single clusters). The fact that some samples from the same biotype did not cluster together likely results from incomplete lineage sorting or ongoing gene flow between biotypes (Peccoud, Ollivier, et al. 2009). When comparing *B. aphidicola* and mitochondrial phylogenies (see Additional file 7), the well-supported branches of the latter were identically retrieved on the endosymbiont phylogeny, but *B. aphidicola* phylogeny was better resolved. This confirms the suitability of using B. aphidicola phylogeny as a framework for examining evolutionary dynamics of secondary symbiont infections. Overall, we built a solid phylogenetic framework for *B. aphidicola* with good branch supports, that we further used to contrast primary and secondary symbiont histories.

3.5 Phylogenetic insights on the evolutionary histories of host-secondary symbiont associations

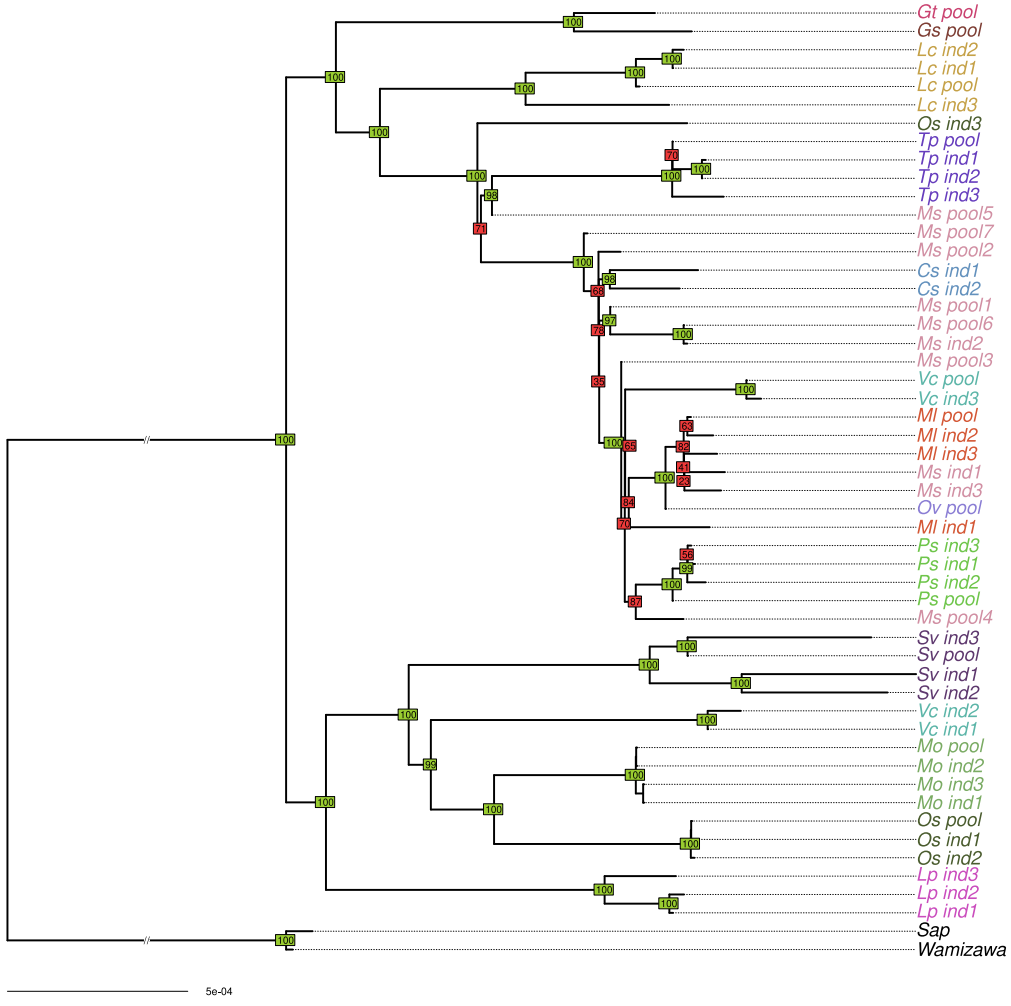
We then examined the evolutionary histories of the associations between secondary symbionts and their pea aphid hosts by comparing one by one the matriline phylogeny reconstructed from *B. aphidicola* with the phylogeny of each of the seven secondary symbionts detected with sufficient coverage in our metagenomics dataset.

Visual comparison of the matriline phylogeny with *Hamiltonella defensa* phylogeny revealed some congruent

Table 2 Summary of variant calling results. Outgroup samples were excluded to report the diversity within the dataset

Symbiont	Number of samples	Number of SNPs/kb	Number of Indels/kb	Maximum distance between two samples (variants/kb)
Serratia symbiotica	9	1.46	0.13	1
Buchnera aphidicola	50	12.61	0.56	3.03
Hamiltonella defensa	16	22.16	0.91	12.61
Regiella insecticola	12	18.2	0.56	16.75
Rickettsia sp.	9	1.19	0.12	1.19
Rickettsiella viridis	8	0.03	0	0.03
Fukatsuia symbiotica	8	2.21	0.04	1.6
Spiroplasma sp.	12	1.95	0	1.28

Figure 2 Phylogeny of Buchnera aphidicola. Phylogeny was inferred by Maximum likelihood based on a concatenate of 96 membrane protein-coding genes. Bootstrap values above or below 90 appear in green and red, respectively.



nodes but also several differences in tree topologies indicating frequent horizontal transfers inside the pea aphid complex (Figure 3). Reconciliation analyses detected nine possible events of host shifts and six cospeciation events, which yielded a co-diversification scenario that is less costly than expected by chance. In addition, three events of loss were detected. This reflects mixed patterns of transmission with overall vertical transmission of this secondary symbiont along the evolutionary history of the pea aphid complex, combined with multiple events of horizontal transfers and some losses (see Additional file 8). *Spiroplasma sp.* phylogeny also showed many incongruences with the matriline phylogeny, presumably reflecting frequent horizontal transfers (Figure 4). Reconciliation analysis inferred eight potential host-switch events and only three cospeciation events (see Additional file 8). In that case, the cospeciation hypothesis was rejected, indicative of a shorter association of *Spiroplasma* with the pea aphid complex.

Regiella insecticola phylogeny retrieved two well-differentiated clades (Figure 5). Whole genome variant calling indicated that more than 30,000 variants distinguish these two clades, while intra-clade variation was much lower, with at best 8,000 variants called. These two clades may have infected the pea aphid complex separately, and seem to be preferentially associated with different biotypes (*Medicago sativa* for clade 1 and *Trifolium pratense* for clade 2). Given the low variation within each lineage relative to the large divergence between the two lineages, we can confidently assume that the acquisition of these symbionts by the different aphid hosts occurred after their divergence. The matriline phylogeny and the *R. insecticola* phylogeny showed several incongruences within and between the two clades, suggesting frequent horizontal transfers, as suggested above for *H. defensa* and *Spiroplasma*. Accordingly, the reconciliation analysis detected 10 events of host switch and a single cospeciation event. The signal of cospeciation between *Regiella* and *Buchnera* was not significant, supporting horizontal transmission and frequent losses events of this symbiont in the pea aphid complex (see Additional file 8).

Despite of the low genomic diversity found for *Rickettsiella viridis*, most nodes of the phylogeny are well supported (Figure 6). Reconciliation analysis revealed only one cospeciation event along with six host-switch events. Accordingly, no significant cospeciation signal was found. This result, combined with the fact that this symbiont is found in only three biotypes of our sample and is poorly diverse, suggests a very recent history of this association in the pea aphid complex. In our sample, *Fukatsuia symbiotica* was associated preferentially with the *Medicago sativa* biotype, either because of its recent acquisition, low rate of horizontal transfers or strong incompatibilities/counter-selection in other biotypes. Phylogenetic analysis revealed a few incongruences between tree topologies of host matrilines and *F. symbiotica* (Figure 7). This pattern presumably reflects cases of horizontal transfer, in agreement with the reconciliation analysis that detected three host switch events. However, we found a significant signal of cospeciation (four putative events), indicative of overall vertical transmission within the Medicago sativa biotype.

Several incongruences were observed between the phylogenies of *Rickettsia sp.* and *B. aphidicola* (Figure 8). The reconciliation analyses recovered four host switch events and four cospeciation events; the cospeciation signal was not significant.

The *Serratia* symbiotica phylogeny delineated several clades for this symbiont (Figure 9). Nine samples are infected by this symbiont in eight different biotypes, indicating that *S. symbiotica* is represented in most of the biotypes but at a moderate prevalence across the complex. Some incongruences were observed between *S. symbiotica* and primary symbiont phylogenies but all involved nodes with low support on the *S. symbiotica* phylogeny. Reconciliation analyses revealed a few number of host switch events along with significant cospeciation (Additional file 8). However, the fact that *S. symbiotica* is found at a moderate prevalence suggests some failures in vertical transmission, leading to loss events in pea aphid lineages (three losses were indeed detected by the reconciliation test).

3.6 Intra-host coexistence of two *Regiella insecticola* strains

Investigation of inter-sample phylogenetic relationships was led by considering the most abundant alleles for each sample. However, some samples may be polymorphic at some sites, with both the reference and alternative alleles covered in metagenomic dataset. While intra-sample genomic variability is expected for pooled samples, which originate each from a diverse host population, it would be more surprising for individual sequencing samples. However, we observed that genome sequences from two individual clones of the Trifolium biotype (Tp_ind1 and Tp_ind2) showed a high number (32,000) of intra-sample polymorphic sites along the *R. insecticola* genome. These two samples showed no sign of polymorphism for the primary symbiont and mitochondrial genomes, excluding the hypothesis of contamination during the sequence data production.

Figure 10 shows the coverage distribution for major and minor alleles of *R. insecticola* in the Tp_ind1 sample. A similar distribution was obtained for the Tp_ind2 sample. These bimodal distributions suggest that two genotypes of *R. insecticola* coexist in these two samples. We estimated the read depth of the two genotypes with the most abundant genotype in Tp_ind1 covered at around 40X, and the other genotype at around 10X (25X and 10X for

Figure 3 a) Phylogeny of *Hamiltonella defensa*. Phylogeny was inferred by Maximum likelihood based on a concatenate of 118 membrane protein-coding genes. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Hamiltonella defensa* phylogeny with host matriline phylogeny

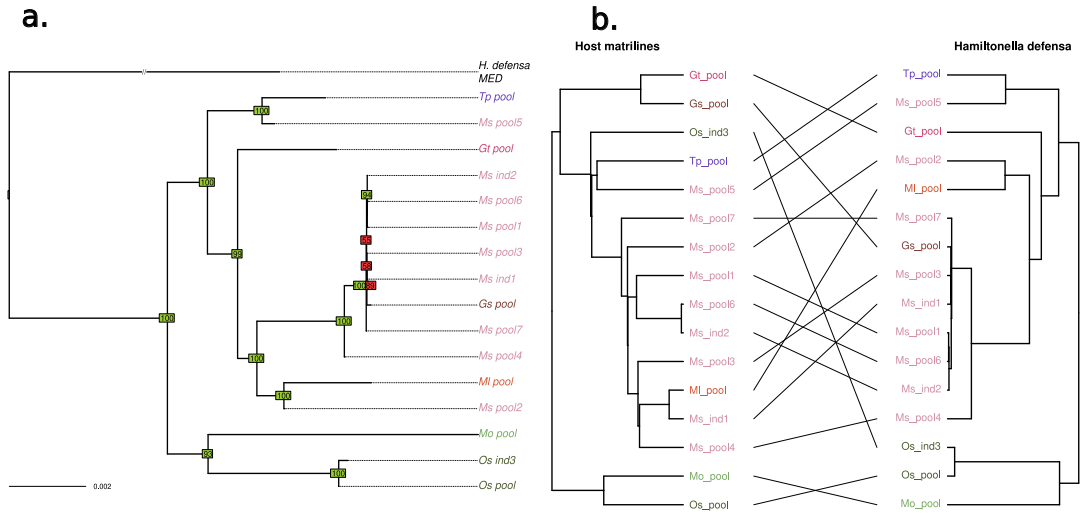


Figure 4 a) Phylogeny of *Spiroplasma sp.*. Phylogeny was inferred by Neighbor Joining based on a whole genome variant calling. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Spiroplasma sp.* phylogeny with host matriline phylogeny

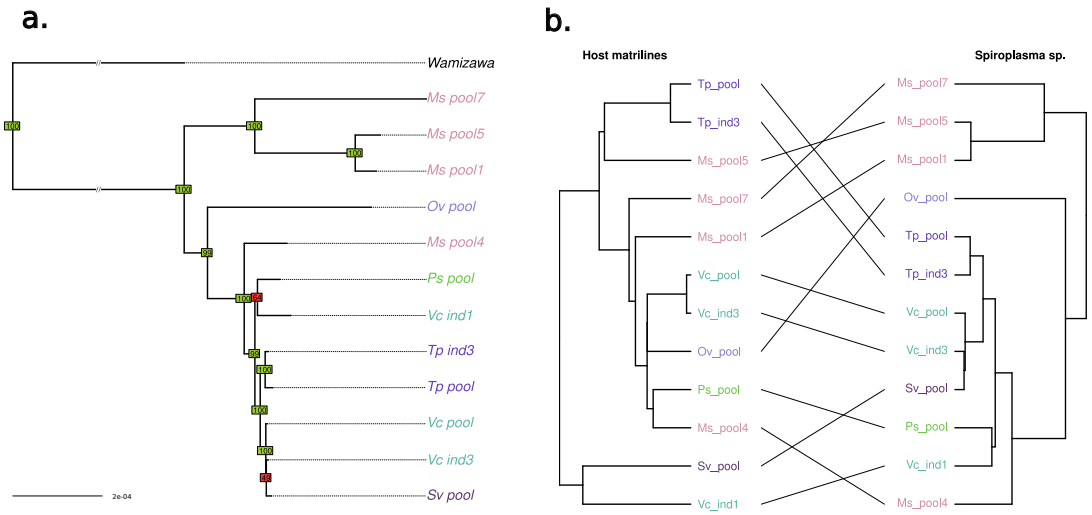


Figure 5 a) Phylogeny of *Regiella insecticola*. Phylogeny was inferred by Maximum likelihood based on a concatenate of 96 membrane protein-coding genes. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Regiella insecticola* phylogeny with host matriline phylogeny

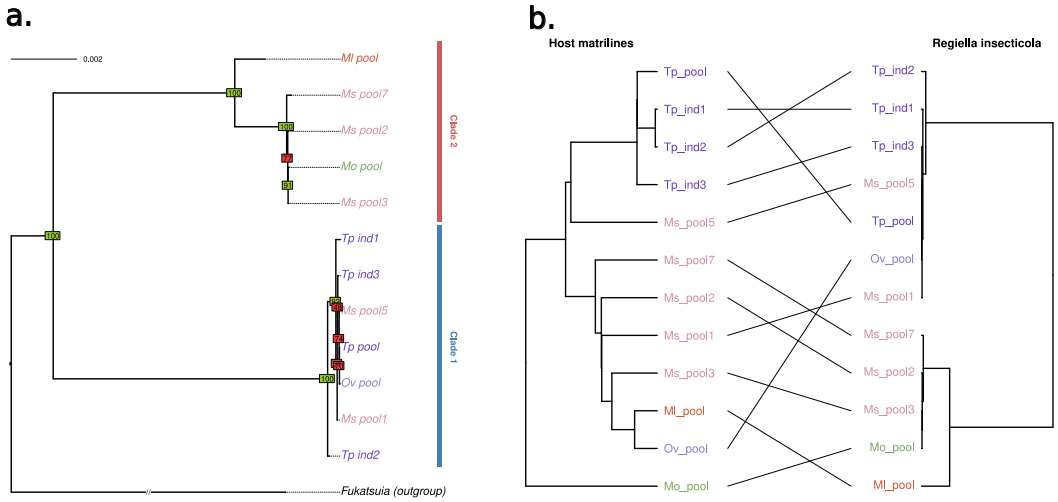


Figure 6 a) Phylogeny of *Ricketsiella viridis*. Phylogeny was inferred by Neighbor Joining based on a whole genome variant calling. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Rickettsiella viridis* phylogeny with host matriline phylogeny

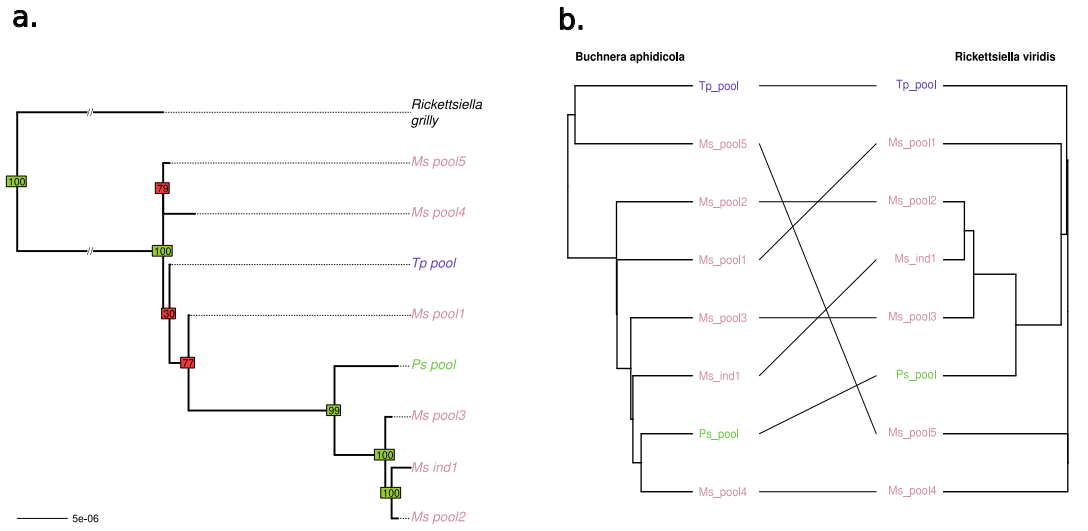
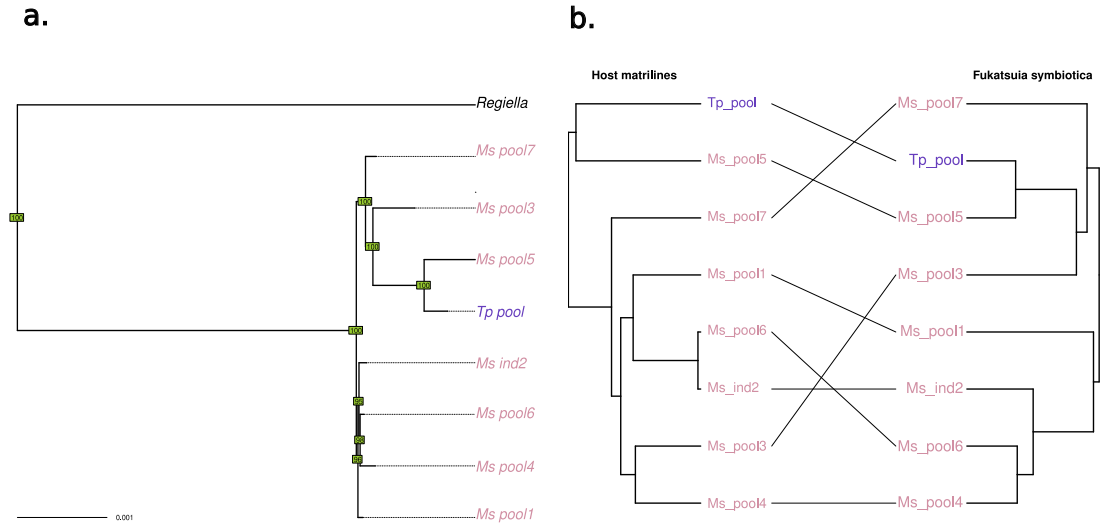


Figure 7 a) Phylogeny of *Fukatsuia symbiotica*. Phylogeny was inferred by Neighbor Joining based on a whole genome variant calling. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Fukatsuia symbiotica* phylogeny with host matriline phylogeny



Tp_ind2). The variant profiles for these two genotypes were close to the ones observed for the two clades of *R. insecticola* described in Figure 5.

Sequencing data thus indicate the coexistence of two *R. insecticola* lineages inside particular samples, but it does not prove this coexistence inside individual aphids, because samples denominated as "individual sequencing" actually resulted from the sequencing of a pool of individual aphids from the same clone. Therefore, it is possible that aphids from the same clone host different symbiont genotypes. To challenge this hypothesis, we performed experimental validation on individual aphids picked in the clonal lineage maintained in culture in our laboratory. A deletion of 32 bp differentiating the two clades was identified on the contig of accession AGCA01000518 (see Additional file 9). We designed primers to amplify the region corresponding to this deletion. Electrophoresis confirmed the presence of the two haplotypes in individual aphids from the Tp_ind1 and Tp_ind2 clonal lineages, while a single haplotype was detected in aphids from the Tp_ind3 clone. This validation confirmed the coexistence inside single individual aphid hosts of two distinct genotypes of *R. insecticola*.

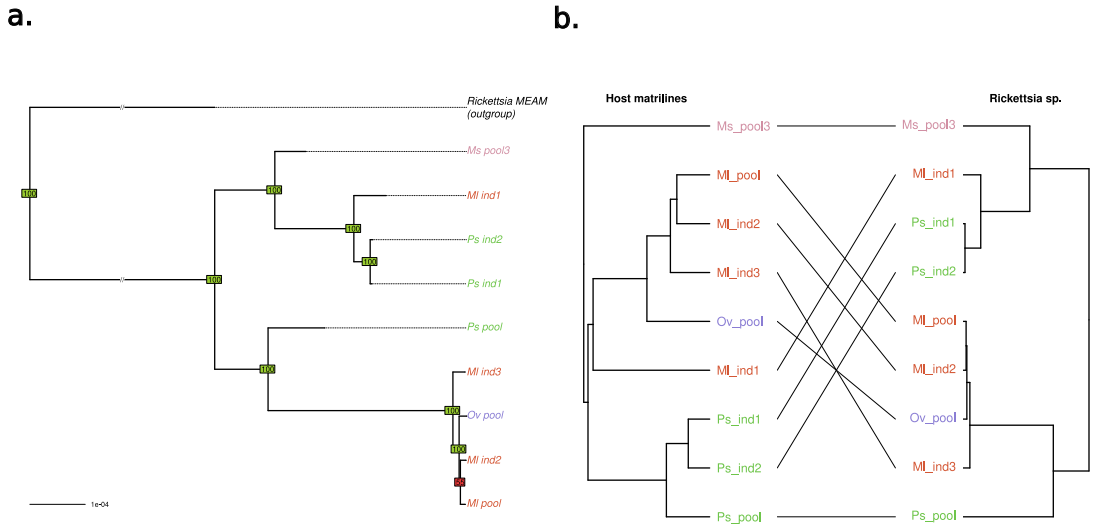
4 Discussion

We present here a framework to explore multi-scale genomic diversity in holobiont systems of low complexity, which is generally the case of insect holobionts. We applied this approach to metagenomic datasets of the pea aphid complex by considering microbial variation across host biotypes, amongst individuals of the same biotype, and within individual aphids. This work allowed to extract more than 99% of the metagenomic information and to draw a complete inventory of microbes associated to the pea aphid complex, revealing a microbiota dominated by a few bacterial symbionts. Our approach also revealed for the first time a large genomic diversity among A. pisum symbionts, with different diversity patterns between symbiont taxa presumably reflecting distinct evolutionary histories, genomic features, transmission patterns and ecological influences across pea aphid biotype-symbiont associations. Finally, phylogenomic analyses highlighted that frequent horizontal transfers and losses of facultative symbionts have probably been common events during the diversification of the *A. pisum* complex.

4.1 Guidelines for analysing multi-scale holobiont metagenomic diversity

The method proposed to finely analyse holobiont metagenomic diversity was based on the mapping of metagenomic reads on a set of reference genomes. By doing so, the entangled metagenomic read set was transformed into symbiont specific read subsets, which enabled finer analyses such as intra-sample variability detection or strain level diversity analyses. The method is reliable for the pea aphid holobiont, which has a restricted number of symbiotic partners, and for which reference genomes are partly available. The rate of unmapped reads was below 1% for most samples, and variations depending on the composition of symbiotic communities were observed,

Figure 8 a) Phylogeny of *Rickettsia sp.*. Phylogeny was inferred by Neighbor Joining based on a whole genome variant calling. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Rickettsia sp.* phylogeny with host matriline phylogeny



indicating that the availability and quality of reference genomes are important to achieve a good assignation of the metagenomic reads. When distant reference genomes are used for mapping, highly divergent regions and large insertions or deletions obviously limit the assignation success rate. Overall, mapping of metagenomics reads on a set of reference genomes (when available) or de novo assembled genomes (when coverage is sufficient), followed by a strain-level analysis of genomic variation appears to be an appropriate characterization method in the case of the pea aphid holobiont.

A large number of aphid samples were sequenced in order to investigate microbial diversity across the pea aphid complex of biotypes. However, sequencing data from host aphids did not allow accessing directly to individual bacterial genotypes, and we had to build genotypes based on the most abundant alleles in each bacterial population. In our dataset, individual sequencing samples had either a low intra-sample polymorphism or a mix of genotypes we could easily disentangle (as for example *R. insecticola* in the *Trifolium* biotype). However, pooled samples were analysed so that only the most abundant alleles were kept to reconstruct the genotype of each symbiotic lineage. Overall, this assumption leads to underestimate the actual diversity in the pooled samples. Compared to individual host genotype sequencing, pooled sequencing allows to capture a greater diversity of symbiotic lineages, but suffers limitations in reconstructing individual bacterial genotypes, due to methodological problems in handling large intra-sample polymorphism. Despite this limitation and the fact that we applied stringent filters to discard ambiguous variants, in most cases we could retrieve a sufficient number of reliable variants from metagenomics reads to compare symbiont diversity and to build well-resolved phylogenies.

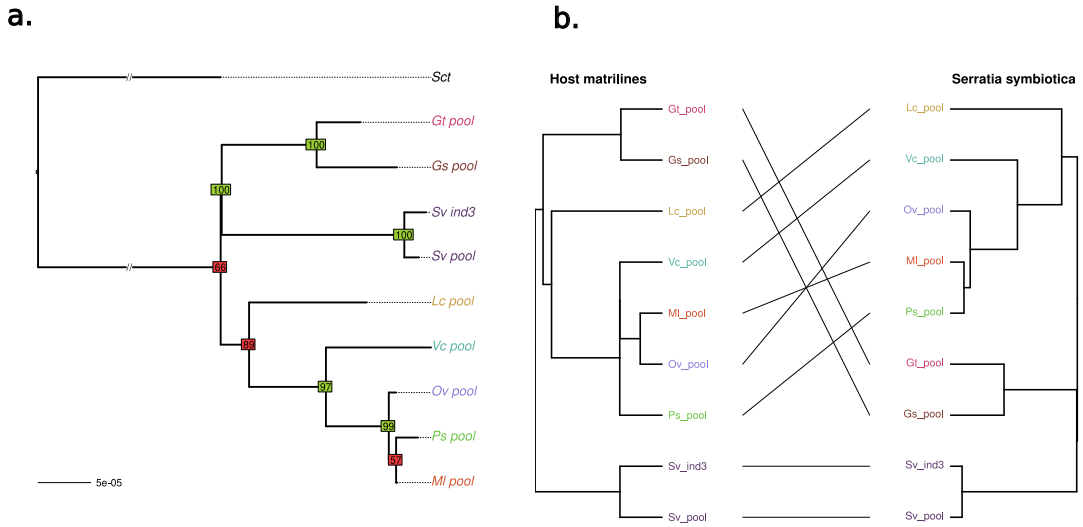
Search for genomic variants was restricted to SNPs and short insertions and deletions. The analysis of large genomic rearrangements may bring additional information on the symbiotic genomic diversity [73]. Short variant information seems to be sufficient to reconstruct symbiotic phylogenetic trees, since most phylogenetic studies rely on gene sequences analyses, and generally do not integrate rearrangements, but this structural variation should not be neglected in order to reconstruct full genomes for the main microbial genotypes existing in pea aphid holobionts.

4.2 Multi-scale diversity inventory of an holobiont

Previous studies on the pea aphid's microbiota focused on the detection of symbionts using 16S rRNA PCR based detection or 16S amplicon sequencing [33, 74]. The drawbacks of these methods are that they are restricted to bacteria, have generally low taxonomic resolution, suffer from several biases due to DNA amplification and may be unable to identify new microbial partners [75].

To overcome these limitations, we used shotgun metagenomic sequencing, which captures whole genomic information about the host and its associated microbial community. We successfully assigned most of the reads to host and symbiont reference genomes forming the pea

Figure 9 a) Phylogeny of *Serratia symbiotica*. Phylogeny was inferred by Maximum likelihood based on a concatenate of 141 membrane protein-coding genes. Bootstrap values above or below 90 appear in green and red, respectively.
b) Tanglegram of *Serratia symbiotica* phylogeny with host matriline phylogeny

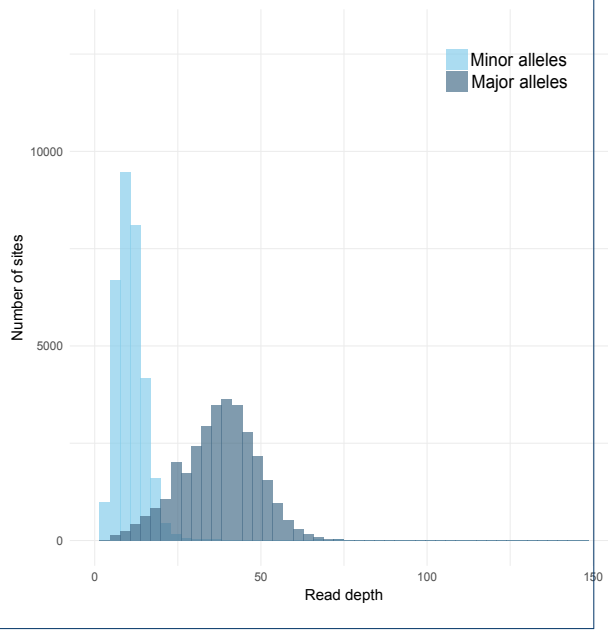


aphid holobiont, and checked that no new bacterial symbiont was abundant in unmapped reads. Also, we found no evidence for the occurrence of *Wolbachia* in our large metagenomic dataset though this symbiont has been reported in *A. pisum* in three previous studies [76, 74, 33]. One explanation could be that none of the pea aphids used for individual or pooled resequencing projects was infected by this symbiont. Alternatively, detection of *Wolbachia* in previous studies could result from artefacts or DNA amplification from aphid endoparasitoids. Because DNA was extracted from aphid clonal lineages cultured in laboratory conditions for two generations (to avoid contamination from aphid parasite microbiota and limit environmental microbes), only the inherited part of the microbiota was sequenced. In contrast with a previous study based on 16S rRNA sequencing [33], no gut associate microbe was found in our metagenomic dataset, suggesting either a low prevalence of such microbes in pea aphid populations, their loss in culture because of poor vertical transmission, or an artefact of 16S rRNA data. Finally, apart from the bacteriophage APSE, no fungal or viral associates were found. However, because of their small genome sizes, unreferenced viruses could have been missed in the unmapped reads analysis. In addition, RNA viruses are common in arthropods and need specific detection methods [77] Therefore, further analyses are required to in depth examination of the pea aphid virome with dedicated approaches [77]. These results altogether indicate an apparent low complexity of the pea aphid microbiota when considered at a species-level scale, and are in accordance with previous works on aphids and other sap-feeders insects showing low richness of host associated microbial communities and mostly composed of a few heritable bacterial symbionts [78, 79].

4.3 Contrasting evolutionary dynamics of pea aphid-secondary symbiont associations

b) Tanglegram of Hamiltonella defensa phylogeny with host matriline phylogeny The history of the symbiosis between aphids and their primary symbiont *B. aphidicola* is well known, with a 160-280 million years old association [80]. Although *B. aphidicola* can be experimentally transferred between aphid matrilines and has been lost in a few aphid taxa [81], it is considered as a strictly maternally inherited symbiont and no horizontal transfer has been observed so far at different phylogenetic scales [70, 82, 83]. For *A. pisum*, we observed in the present work a close congruence between mitochondrial and *B. aphidicola* phylogenies, indicating a persisting association between the host and its primary symbionts, and a codiversification of both partners in recent evolutionary time. Genome-wide analysis of *B. aphidicola* diversity in the pea aphid complex showed a diversification of pea aphid matrilines which corresponds well to the adaptive radiation that led to the complex of biotypes, and confirmed previous results obtained from pseudo-gene sequences of *Buchnera* [31, 72]. Using our well-resolved *B. aphidicola* phylogeny, we were able to contrast the evolutionary trajectories of pea aphid matrilines with that of every *A. pisum* secondary symbiont and to propose different history scenarios of pea aphid-secondary symbiont associations.

Figure 10 Read depth for polymorphic alleles of R. insecticola genome in Tp_ind1 sample. Bimodal distribution indicates the occurrence of two distinct genotypes within the same aphid host sample.



Several secondary symbionts are known in *A. pisum* and other aphid species, but the nature of their association with aphid hosts is variable, from free association to co-obligatory symbiont with intermediate stages of dependency [42]. Recent data provide evidence for a higher rate of mother to offspring transmission for most of the secondary symbionts presented here [84], but some indirect proofs of horizontal transfers have also been reported [28, 34]. Their underlying mechanisms are still unclear, with host plant, natural enemies, or paternal transmission as candidate paths for horizontal transfers[85]. In this study we showed a contrasting genomic diversity for the different symbionts, from poorly diverse symbionts such as *Rickettsiella viridis* to highly heterogeneous ones such as *Regiella insecticola*. This heterogeneity in genomic diversity could result from the combination of several factors, such as differences in evolutionary rates, population size, transmission modes and host-symbiont association histories [86, 87, 84]. It is also very likely that these symbiotic associations are constrained by different factors including host compatibility to new infection [88] and selection [34]. For example, some symbionts like *Serratia* seems to have a wide host range [89] while others like *Fukatsuia* tend to be more restricted in terms of biotypes. In the specific case of *R. viridis*, although we cannot totally discard this hypothesis, the very low genomic variation is unlikely to result from a low mutation rate considering the level of diversity of *R. viridis* which is two orders of magnitude less than for the other symbionts associated to pea aphids and that there is no particular mention of this pattern in the literature. Instead, this low population genomic diversity in *R. viridis* might rather result from its relatively recent acquisition by a few A.pisum lineages, likely from a small number of sources. Evolutionary dynamics of symbiotic associations in the pea aphid complex were studied here by comparing phylogenetic trees of secondary symbionts with that of the obligatory symbiont B. aphidicola, as a proxy of pea aphid matriline phylogeny. While symbiotic species showing phylogenetic congruence with *B. aphidicola* probably reflect co-speciation with their aphid host lineages, incongruent symbiont phylogenies are expected to result from different events such as horizontal transfers or symbiont loss/gain events. Accordingly, incongruences between matriline and secondary symbiont phylogenies were observed for all secondary symbionts considered in this study. Host switches were detected for every secondary symbiont by reconciliation analyses, supporting the hypothesis of frequent horizontal transfers proposed in previous studies on that system [34] Reconciliation analyses also detected a few events of loss for most symbionts and those could result from failures in vertical transmission as sometimes observed in laboratory conditions [85]. With reconciliation analyses, we also found several cases of significant signals of co-speciation between secondary symbionts and their host matrilines. Since secondary symbionts of the pea aphid are maternally inherited with a generally good fidelity [84], this is not a surprising result. However, these results need to be interpreted with care as for some samples (pooling several individuals), we only reconstructed the most abundant genotype for each symbiont and might have therefore underestimated the phylogenetic diversity of the biotype-symbiont associations and the complexity of co-diversification scenarios. In any case, our approach suggests that cospeciation signals as well as the numbers of gain and loss estimated from reconciliation tests greatly differ between secondary symbionts, reflecting mixed patterns of transmission and different dynamics and durations of these symbiotic associations among the pea aphid complex. In the case of *Regiella insecticola*, we revealed an even more complex situation. We indeed found that *R. insecticola* populations in pea aphid biotypes encompass two highly differentiated genotypes, likely representing two distinct events of infection by symbiont strains that diverged much before the diversification of pea aphid biotypes. Horizontal transfers of these two genotypes were also detected within the pea aphid complex, indicating more recent host switch.

Overall, these evolutionary scenarios of symbiotic associations in the pea aphid complex suggest that the rate and source of horizontal transfers are very variable across symbionts, in accordance with previous studies at lower

resolutions [28]. Yet, these results may be extended by larger phylogenetic studies in the pea aphid complex but also in other aphid and arthropod taxa, and by investigations of the amount and mechanisms of gain (horizontal transfers) and loss of secondary symbionts in natural populations of pea aphids.

4.4 Intra-host coexistence of different *Regiella insecticola* strains

Our metagenomic approach on the pea aphid microbiota also revealed an unexpected level of diversity. Indeed, this study showed evidence for the coexistence of two divergent R. insecticola genotypes within the same individual aphid. While the within host coexistence of symbiotic strains from the same lineage has already been reported in some arthropods [90], it has been rarely found in aphids (but see [42]). This bi-infection of *R. insecticola* strains inside individual aphids has been observed for two clones, where the two existing strains were both very different and equally abundant, facilitating detection and characterization of their infection status. However, some less obvious cases of multi-infection in other samples or by other symbionts might have been undetected. The development of dedicated techniques to analyse intra-sample polymorphism may help to better understand these events of coinfection and their evolutionary implications. The discovery of this symbiotic coinfection raises new questions concerning the effects of these strains, individually or in conjunction, on host fitness and phenotype, their localization and interaction in the aphid host, and the stability of this coinfection.

An important aspect which requires dedicated studies is how this genomic diversity in pea aphid microbiota translates into functional differences and influences the holobiont phenotype. It is known that strain level genomic variation can have considerable consequences on the expression of the host extended phenotype. For instance, previous works demonstrated that the level of natural enemy protection provided by H. defensa is highly different between two Genista biotypes infected by genetically distinct strains of the protective symbiont [36]. Here, the reconstructed H. defensa phylogeny confirmed that these two *Genista* biotypes host highly different symbiotic populations, while sharing close matriline history. Genome-wide variant discovery may help to infer metabolic differences between *H. defensa* genotypes and their associated APSE phages that could cause the variation in protection levels of the hosts [91]. Similarly, a functional annotation of the genomic differences between the two highly divergent genotypes of *R. insecticola* found singly or in co-infection within the same host, may reveal different impacts on the host phenotype.

5 Conclusions

We conducted a multi-scale analysis of genomic diversity associated with the pea aphid microbiota, ranging from the common species- and biotype-levels analysis, to a more innovative intra-specific analysis, and we were able to uncover the genomic diversity at each considered scale.

Improved understanding of host-microbiota relationships may benefit from large holobiont sequencing projects, and we believe the framework we developed here is applicable to other holobiont systems of low complexity. By analysing whole genome variation in the pea aphid holobiont, we confirmed that its microbiome diversity is limited to a few inherited symbionts, but we revealed a generally large genomic diversity observed at different levels of the holobiont organization. This genomic diversity in populations of secondary symbionts seems to be mainly shaped by the dynamics of symbiotic associations, which could take multiple routes and lead to different evolutionary trajectories.

This work paves the way for new studies relying on metabolic and functional approaches and aiming to examine how genomic variation in microbiota affects host fitness and phenotypic traits. Moreover, a full understanding of the evolutionary history and ecology of symbiotic associations requires a larger investigation of the sources of genomic diversity at different geographical, temporal and trophic scales.

Although the metagenomic framework we developed here for the pea aphid system yielded significant knowledge improvements in patterns of genomic diversity and evolution in host-symbiont associations, we pinpointed some limitations in our approach such as the availability of reference genomes and the difficulty to handle metagenomic data of high complexity. Methods to analyse fine-scale diversity from metagenomic dataset are still rare, and require either well annotated reference genomes, or simple communities where organisms are easy to disentangle. More advanced methods have to be developed to assess metagenomic diversity in either complex or non-model holobionts.

Acknowledgements
Authors warmly thank Jean Peccoud for advice on approaches to explore the evolutionary dynamics of pea aphid symbiont associations and the GenOuest BioInformatics Platform that provided the computing resources necessary for bioinformatic analyses.

Competing interests
The authors declare that they have no competing interests.

Funding
CG was supported by Université of Rennes 1 through a PhD grant. This work was supported by the Plant health and Environment division of INRA and the ANR Speciaphid (ANR-11-BSV7-005-01) to JCS

Availability of data and materials
Individual aphid sequencing datasets used for the current study are available under BioProject ID PRJNA255937. Pool sequencing datasets will be registered shortly.

Author's contributions
CG performed the analyses and wrote the paper, guided by FL, EJ, CM, CL and JCS. EJ supervised phylogenetic approaches and performed the reconciliation analyses. All authors read, revised and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Author details
[1]INRA, UMR 1349 INRA/Agrocampus Ouest/Université Rennes 1, Institut de Génétique, Environnement et Protection des Plantes (IGEPP), Le Rheu, France. [2]Univ Rennes, Inria, CNRS, IRISA, , F-35000 Rennes, France. [3]INRA, UMR CBGP (INRA/IRD/Cirad/Montpellier SupAgro), Campus International de Baillarguet, Montpelier, France.

References

1. Munson, M.A., Baumann, P., Clark, M.A., Baumann, L., Moran, N.A., Voegtlin, D.J., Campbell, B.C.: Evidence for the establishment of aphid-eubacterium endosymbiosis in an ancestor of four aphid families. Journal of Bacteriology **173**(20), 6321–6324 (1991). doi:10.1126/science.aaf3951
2. Thao, M.L., Moran, N.A., Abbot, P., Brennan, E.B., Burckhardt, D.H., Baumann, P.: Cospeciation of psyllids and their primary prokaryotic endosymbionts. Applied and Environmental Microbiology **66**(7), 2898–2905 (2000). doi:10.1128/AEM.66.7.2898-2905.2000
3. Gil, R., Sabater-Muñoz, B., Latorre, A., Silva, F.J., Moya, A.: Extreme genome reduction in Buchnera spp.: toward the minimal genome needed for symbiotic life. Proceedings of the National Academy of Sciences of the United States of America **99**(7), 4454–8 (2002). doi:10.1073/pnas.062067299
4. Charlat, S., Hurst, G.D.D., Merçot, H.: Evolutionary consequences of Wolbachia infections. Trends in Genetics **19**(4), 217–223 (2003). doi:10.1016/S0168-9525(03)00024-6
5. Oliver, K.M., Russell, J.A., Moran, N.A., Hunter, M.S.: Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proceedings of the National Academy of Sciences **100**(4), 1803–1807 (2003). doi:10.1073/pnas.0335320100. 1011.1669v3
6. Russell, J.A., Moran, N.A.: Costs and benefits of symbiont infection in aphids: variation among symbionts and across temperatures. Proc. Biol. Sci. **273**(1586), 603–610 (2006). doi:10.1098/rspb.2005.3348
7. Christian, N., Whitaker, B.K., Clay, K.: Microbiomes: Unifying animal and plant systems through the lens of community ecology theory. Frontiers in Microbiology **6**(SEP), 869 (2015). doi:10.3389/fmicb.2015.00869
8. McFall-Ngai, M., Hadfield, M.G., Bosch, T.C.G., Carey, H.V., Domazet-Lošo, T., Douglas, A.E., Dubilier, N., Eberl, G., Fukami, T., Gilbert, S.F., Hentschel, U., King, N., Kjelleberg, S., Knoll, A.H., Kremer, N., Mazmanian, S.K., Metcalf, J.L., Nealson, K., Pierce, N.E., Rawls, J.F., Reid, A., Ruby, E.G., Rumpho, M., Sanders, J.G., Tautz, D., Wernegreen, J.J.: Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences **110**(9), 3229–3236 (2013). doi:10.1073/pnas.1218525110. NIHMS150003
9. Rosenberg, E., Koren, O., Reshef, L., Efrony, R., Zilber-Rosenberg, I.: The role of microorganisms in coral health, disease and evolution. Nature Reviews Microbiology **5**(5), 355–362 (2007). doi:10.1038/nrmicro1635
10. Rohwer, F., Seguritan, V., Azam, F., Knowlton, N.: Diversity and distribution of coral-associated bacteria. Marine Ecology Progress Series **243**, 1–10 (2002). doi:10.3354/meps243001. arXiv:1011.1669v3
11. Bordenstein, S.R., Theis, K.R.: Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. PLoS Biology **13**(8), 1002226 (2015). doi:10.1371/journal.pbio.1002226
12. Douglas, A.E., Werren, J.H.: Holes in the hologenome: Why host-microbe symbioses are not holobionts. mBio **7**(2), 02099 (2016). doi:10.1128/mBio.02099-15
13. Jaspers, E., Overmann, J.: Ecological significance of microdiversity: Identical 16S rRNA gene sequences can be found in bacteria with highly divergent genomes and ecophysiologies. Applied and Environmental Microbiology **70**(8), 4831–4839 (2004). doi:10.1128/AEM.70.8.4831-4839.2004
14. Thomas, G.H., Zucker, J., Macdonald, S.J., Sorokin, A., Goryanin, I., Douglas, A.E.: A fragile metabolic network adapted for cooperation in the symbiotic bacterium Buchnera aphidicola. BMC Systems Biology **3**(1), 24 (2009). doi:10.1186/1752-0509-3-24
15. Albertsen, M., Hugenholtz, P., . . . , A.S.-N., undefined 2013: Albertsen et al. 2013. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. researchgate.net
16. Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T.S., Shapiro, N., Blood, P.D., Gurevich, A., Bai, Y., Turaev, D., Demaere, M.Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiutė, M., Hansen, L.H., Sørensen, S.J., Chia, B.K.H., Denis, B., Froula, J.L., Wang, Z., Egan, R., Don Kang, D., Cook, J.J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.W., Singer, S.W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M.D., Lingner, T., Lin, H.H., Liao, Y.C., Silva, G.G.Z., Cuevas, D.A., Edwards, R.A., Saha, S., Piro, V.C., Renard, B.Y., Pop, M., Klenk, H.P., Göker, M., Kyrpides, N.C., Woyke, T., Vorholt, J.A., Schulze-Lefert, P., Rubin, E.M., Darling, A.E., Rattei, T., McHardy, A.C.: Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nature Methods **14**(11), 1063–1071 (2017). doi:10.1038/nmeth.4458
17. Awad, S., Irber, L., Brown, C.T.: Evaluating Metagenome Assembly on a Simple Defined Community with Many Strain Variants. Doi.Org, 155358 (2017). doi:10.1101/155358
18. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C.: Metagenomic microbial community profiling using unique clade-specific marker genes. Nature methods **9**(8), 811–4 (2012). doi:10.1038/nmeth.2066
19. Truong, D.T., Tett, A., Pasolli, E., Huttenhower, C., Segata, N.: Microbial strain-level population structure & genetic diversity from metagenomes. Genome Research **27**(4), 626–638 (2017). doi:10.1101/gr.216242.116
20. Francis, O.E., Bendall, M., Manimaran, S., Hong, C., Clement, N.L., Castro-Nallar, E., Snell, Q., Schaalje, G.B., Clement, M.J., Crandall, K.A., Johnson, W.E.: Pathoscope: Species identification and strain attribution with unassembled sequencing data. Genome Research **23**(10), 1721–1729 (2013). doi:10.1101/gr.150151.112
21. Ahn, T.H., Chai, J., Pan, C.: Sigma: Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics **31**(2), 170–177 (2015). doi:10.1093/bioinformatics/btu641
22. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti, L., Loman, N.J., Andersson, A.F., Quince, C.: Binning metagenomic contigs by coverage and composition. Nature Methods **11**(11), 1144–1146 (2014). doi:10.1038/nmeth.3103
23. Imelfort, M., Parks, D., Woodcroft, B.J., Dennis, P., Hugenholtz, P., Tyson, G.W.: GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ **2**, 603 (2014). doi:10.7717/peerj.603
24. Cleary, B., Brito, I.L., Huang, K., Gevers, D., Shea, T., Young, S., Alm, E.J.: Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nature Biotechnology **33**(10), 1053–1060 (2015). doi:10.1038/nbt.3329
25. Baumann, P.: Biology of Bacteriocyte-Associated Endosymbionts of Plant Sap-Sucking Insects. Annual Review of Microbiology **59**(1), 155–189 (2005). doi:10.1146/annurev.micro.59.030804.121041
26. Tsuchida, T., Koga, R., Shibao, H., Matsumoto, T., Fukatsu, T.: Diversity and geographic distribution of secondary endosymbiotic bacteria in natural populations of the pea aphid, Acyrthosiphon pisum. Molecular Ecology **11**(10), 2123–2135 (2002). doi:10.1046/j.1365-294X.2002.01606.x
27. Funk, D.J., Wernegreen, J.J., Moran, N.A.: Intraspecific variation in symbiont genomes: Bottlenecks and the aphid-Buchnera association. Genetics **157**(2), 477–489 (2001). doi:10.1126/science.1129844
28. Sandström, J.P., Russell, J.A., White, J.P., Moran, N.A.: Independent origins and horizontal transfer of bacterial symbionts of aphids. Molecular Ecology **10**(1), 217–228 (2001). doi:10.1046/j.1365-294X.2001.01189.x

29. Oliver, K.M., Moran, N.a., Hunter, M.S.: Variation in resistance to parasitism in aphids is due to symbionts not host genotype. Proceedings of the National Academy of Sciences of the United States of America **102**(36), 12795–800 (2005). doi:10.1073/pnas.0506131102
30. Simon, J.-C., Carre, S., Boutin, M., Prunier-Leterme, N., Sabater-Munoz, B., Latorre, A., Bournoville, R.: Host-based divergence in populations of the pea aphid: insights from nuclear markers and the prevalence of facultative symbionts. Proceedings of the Royal Society B: Biological Sciences **270**(1525), 1703–1712 (2003). doi:10.1098/rspb.2003.2430
31. Peccoud, J., Simon, J.-c., Mclaughlin, H.J., Moran, N.A.: Post-Pleistocene radiation of the pea aphid complex revealed by rapidly evolving endosymbionts. Proceedings of the National Academy of Sciences of the United States of America **106**(38), 16315–16320 (2009). doi:10.1073/pnas.0905129106
32. Peccoud, J., Ollivier, A., Plantegenest, M., Simon, J.-C.: A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proceedings of the National Academy of Sciences of the United States of America **106**(18), 7495–500 (2009). doi:10.1073/pnas.0811117106
33. Gauthier, J.P., Outreman, Y., Mieuzet, L., Simon, J.C.: Bacterial communities associated with host-adapted populations of pea aphids revealed by deep sequencing of 16S ribosomal DNA. PLoS ONE **10**(3), 0120664 (2015). doi:10.1371/journal.pone.0120664
34. Henry, L.M., Peccoud, J., Simon, J.C., Hadfield, J.D., Maiden, M.J.C., Ferrari, J., Godfray, H.C.J.: Horizontally transmitted symbionts and host colonization of ecological niches. Current Biology **23**(17), 1713–1717 (2013). doi:10.1016/j.cub.2013.07.029
35. Ferrari, J., West, J.A., Via, S., Godfray, H.C.J.: Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. Evolution **66**(2), 375–390 (2012). doi:10.1111/j.1558-5646.2011.01436.x
36. Leclair, M., Pons, I., Mahéo, F., Morlière, S., Simon, J.C., Outreman, Y.: Diversity in symbiont consortia in the pea aphid complex is associated with large phenotypic variation in the insect host. Evolutionary Ecology **30**(5), 925–941 (2016). doi:10.1007/s10682-016-9856-1
37. Russell, J.A., Weldon, S., Smith, A.H., Kim, K.L., Hu, Y., Łukasik, P., Doll, S., Anastopoulos, I., Novin, M., Oliver, K.M.: Uncovering symbiont-driven genetic diversity across North American pea aphids. Molecular Ecology **22**(7), 2045–2059 (2013). doi:10.1111/mec.12211
38. Peccoud, J., Simon, J.C., Von Dohlen, C., Coeur d'acier, A., Plantegenest, M., Vanlerberghe-Masutti, F., Jousselin, E.: Evolutionary history of aphid-plant associations and their role in aphid diversification. Comptes Rendus - Biologies **333**(6-7), 474–487 (2010). doi:10.1016/j.crvi.2010.03.004
39. Jaquiéry, J., Stoeckel, S., Nouhaud, P., Mieuzet, L., Mahéo, F., Legeai, F., Bernard, N., Bonvoisin, A., Vitalis, R., Simon, J.C.: Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex. Molecular Ecology **21**(21), 5251–5264 (2012). doi:10.1111/mec.12048
40. Gouin, A., Legeai, F., Nouhaud, P., Whibley, A., Simon, J.C., Lemaitre, C.: Whole-genome re-sequencing of non-model organisms: Lessons from unmapped reads. Heredity **114**(5), 494–501 (2015). doi:10.1038/hdy.2014.85
41. Li, H., Durbin, R.: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**(14), 1754–1760 (2009). doi:10.1093/bioinformatics/btp324. 1303.3997
42. Meseguer, A.S., Manzano-Marín, A., Coeur d'Acier, A., Clamens, A.L., Godefroid, M., Jousselin, E.: Buchnera has changed flatmate but the repeated replacement of co-obligate symbionts is not associated with the ecological expansions of their aphid hosts. Molecular Ecology **26**(8), 2363–2378 (2017). doi:10.1111/mec.13910. 086223
43. Manzano-Marin, A., Szabo, G., Simon, J.C., Horn, M., Latorre, A.: Happens in the best of subfamilies: establishment and repeated replacements of co-obligate secondary endosymbionts within Lachninae aphids. Environmental Microbiology **19**(1), 393–408 (2017). doi:10.1111/1462-2920.13633
44. Degnan, P.H., Moran, N.A.: Diverse phage-encoded toxins in a protective insect endosymbiont. Applied and Environmental Microbiology **74**(21), 6782–6791 (2008). doi:10.1128/AEM.01285-08
45. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R.: The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**(16), 2078–2079 (2009). doi:10.1093/bioinformatics/btp352. 1006.1266v2
46. Richards, S., Gibbs, R.A., Gerardo, N.M., Moran, N., Nakabachi, A., Stern, D., Tagu, D., Wilson, A.C.C., Muzny, D., Kovar, C., Cree, A., Chacko, J., Chandrabose, M.N., Dao, M.D., Dinh, H.H., Gabisi, R.A., Hines, S., Hume, J., Jhangian, S.N., Joshi, V., Lewis, L.R., Liu, Y.S., Lopez, J., Morgan, M.B., Nguyen, N.B., Okwuonu, G.O., Ruiz, S.J., Santibanez, J., Wright, R.A., Fowler, G.R., Hitchens, M.E., Lozado, R.J., Moen, C., Steffen, D., Warren, J.T., Zhang, J., Nazareth, L.V., Chavez, D., Davis, C., Lee, S.L., Patel, B.M., Pu, L.L., Bell, S.N., Johnson, A.J., Vattathil, S., Williams, R.L., Shigenobu, S., Dang, P.M., Morioka, M., Fukatsu, T., Kudo, T., Miyagishima, S.Y., Jiang, H., Worley, K.C., Legeai, F., Gauthier, J.P., Collin, O., Zhang, L., Chen, H.C., Ermolaeva, O., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Maglott, D., Murphy, T., Pruitt, K., Sapojnikov, V., Souvorov, A., Thibaud-Nissen, F., Câmara, F., Guigó, R., Stanke, M., Solovyev, V., Kosarev, P., Gilbert, D., Gabaldón, T., Huerta-Cepas, J., Marcet-Houben, M., Pignatelli, M., Moya, A., Rispe, C., Ollivier, M., Quesneville, H., Permal, E., Llorens, C., Futami, R., Hedges, D., Robertson, H.M., Alioto, T., Mariotti, M., Nikoh, N., McCutcheon, J.P., Burke, G., Kamins, A., Latorre, A., Ashton, P., Calevro, F., Charles, H., Colella, S., Douglas, A.E., Jander, G., Jones, D.H., Febvay, G., Kamphuis, L.G., Kushlan, P.F., Macdonald, S., Ramsey, J., Schwartz, J., Seah, S., Thomas, G., Vellozo, A., Cass, B., Degnan, P., Hurwitz, B., Leonardo, T., Koga, R., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., De Vos, M., Duncan, E.J., Evans, J., Ghanim, M., Heddi, A., Kaloshian, I., Vincent-Monegat, C., Parker, B.J., Pérez-Brocal, V., Rahbé, Y., Spragg, C.J., Tamames, J., Tamarit, D., Tamborindeguy, C., Vilcinskas, A., Bickel, R.D., Brisson, J.A., Butts, T., Chang, C.C., Christiaens, O., Davis, G.K., Duncan, E., Ferrier, D., Iga, M., Janssen, R., Lu, H.L., McGregor, A., Miura, T., Smagghe, G., Smith, J., Van Der Zee, M., Velarde, R., Wilson, M., Dearden, P., Edwards, O.R., Gordon, K., Hilgarth, R.S., Rider, S.D., Srinivasan, D., Walsh, T.K., Ishikawa, A., Jaubert-Possamai, S., Fenton, B., Huang, W., Rizk, G., Lavenier, D., Nicolas, J., Smadja, C., Zhou, J.J., Vieira, F.G., He, X.L., Liu, R., Rozas, J., Field, L.M., Campbell, P., Carolan, J.C., Fitzroy, C.I.J., Reardon, K.T., Reeck, G.R., Singh, K., Wilkinson, T.L., Huybrechts, J., Abdel-Latief, M., Robichon, A., Veenstra, J.A., Hauser, F., Cazzamali, G., Schneider, M., Williamson, M., Stafflinger, E., Hansen, K.K., Grimmelikhuijzen, C.J.P., Price, D.R.G., Caillaud, M., Van Fleet, E., Ren, Q., Gatehouse, J.A., Brault, V., Monsion, B., Diaz, J., Hunnicutt, L., Ju, H.J., Pechuan, X., Aguilar, J., Cortés, T., Ortiz-Rivas, B., Martínez-Torres, D., Dombrovsky, A., Dale, R.P., Davies, T.G.E., Williamson, M.S., Jones, A., Sattelle, D., Williamson, S., Wolstenholme, A., Cottret, L., Sagot, M.F., Heckel, D.G., Hunter, W.: Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biology **8**(2), 1000313 (2010). doi:10.1371/journal.pbio.1000313
47. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., Ishikawa, H.: Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature **407**(6800), 81–86 (2000). doi:10.1038/35024074
48. Degnan, P.H., Yu, Y., Sisneros, N., Wing, R.a., Moran, N.a.: Hamiltonella defensa, genome evolution of protective bacterial endosymbiont from pathogenic ancestors. Proceedings of the National Academy of Sciences of the United States of America **106**(22), 9063–8 (2009). doi:10.1073/pnas.0900194106
49. van der Wilk, F., Dullemans, A.M., Verbeek, M., van den Heuvel, J.F.J.M.: Isolation and Characterization of APSE-1, a Bacteriophage Infecting the Secondary Endosymbiont of Acyrthosiphon pisum. Virology **262**(1), 104–113 (1999). doi:10.1006/viro.1999.9902
50. Hansen, A.K., Vorburger, C., Moran, N.A.: Genomic basis of endosymbiont-conferred protection against an insect parasitoid. Genome Research **22**(1), 106–114 (2012). doi:10.1101/gr.125351.111
51. Degnan, P.H., Leonardo, T.E., Cass, B.N., Hurwitz, B., Stern, D., Gibbs, R.A., Richards, S., Moran, N.A.: Dynamics of genome evolution in facultative symbionts of aphids. Environmental Microbiology **12**(8), 2060–2069 (2010). doi:10.1111/j.1462-2920.2009.02085.x
52. Burke, G.R., Moran, N.A.: Massive genomic decay in Serratia

symbiotica, a recently evolved symbiont of aphids. Genome Biology and Evolution **3**(1), 195–208 (2011). doi:10.1093/gbe/evr002
53. NikohN., Tsutomu T., Maeda T., Yamaguchi K., Shigenobu S., Koga R., T., F.: Genomic Insight into Symbiosis-Induced Insect Color Change by a Facultative Endosymbiont "Candidatus Rickettsiella viridis" Allied to Arthropod and Human Pathogens. submitted to PLoS Pathogens (2017)
54. Klasson, L., Westberg, J., Sapountzis, P., Naslund, K., Lutnaes, Y., Darby, A.C., Veneti, Z., Chen, L., Braig, H.R., Garrett, R., Bourtzis, K., Andersson, S.G.E.: The mosaic genome structure of the Wolbachia wRi strain infecting Drosophila simulans. Proceedings of the National Academy of Sciences **106**(14), 5725–5730 (2009). doi:10.1073/pnas.0810753106. arXiv:1408.1149
55. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A.: SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology **19**(5), 455–477 (2012). doi:10.1089/cmb.2012.0021
56. Rizk, G., Gouin, A., Chikhi, R., Lemaitre, C.: MindTheGap: integrated detection and assembly of short and long insertions. Bioinformatics **30**(24), 3451–3457 (2014). doi:10.1093/bioinformatics/btu545
57. Bolger, A.M., Lohse, M., Usadel, B.: Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics **30**(15), 2114–2120 (2014). doi:10.1093/bioinformatics/btu170
58. Kim, D., Song, L., Breitwieser, F.P., Salzberg, S.L.: Centrifuge: Rapid and sensitive classification of metagenomic sequences. Genome Research **26**(12), 1721–1729 (2016). doi:10.1101/gr.210641.116
59. Breitwieser, F.P., Salzberg, S.L.: Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. bioRxiv (2014), 2014–2017 (2016). doi:10.1101/084715
60. Li, H.: A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics **27**(21), 2987–2993 (2011). doi:10.1093/bioinformatics/btr509. 1203.6372
61. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R.: The variant call format and VCFtools. Bioinformatics **27**(15), 2156–2158 (2011). doi:10.1093/bioinformatics/btr330. NIHMS150003
62. Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., Nielsen, R.: Genes under positive selection in Escherichia coli. Genome Research **17**(9), 1336–1343 (2007). doi:10.1101/gr.6254707
63. Katoh, K., Kuma, K.I., Toh, H., Miyata, T.: MAFFT version 5: Improvement in accuracy of multiple sequence alignment. Nucleic Acids Research **33**(2), 511–518 (2005). doi:10.1093/nar/gki198
64. Xia, X., Xie, Z.: DAMBE: Software package for data analysis in molecular biology and evolution. Journal of Heredity **92**(4), 371–373 (2001). doi:10.1093/jhered/92.4.371
65. Stamatakis, A., Ott, M.: Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Royal Society (2008). doi:10.1098/rstb.2008.0163. http://rstb.royalsocietypublishing.org/cgi/doi/10.1098/rstb.2008.0163
66. Paradis, E., Claude, J., Strimmer, K.: APE: Analyses of phylogenetics and evolution in R language. Bioinformatics **20**(2), 289–290 (2004). doi:10.1093/bioinformatics/btg412
67. Galili, T.: dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics **31**(22), 3718–3720 (2015). doi:10.1093/bioinformatics/btv428. NIHMS150003
68. Gao, X., Starmer, J.: Human population structure detection via multilocus genotype clustering. BMC Genetics **8**(1), 34 (2007). doi:10.1186/1471-2156-8-34
69. Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R.: Jane: a new tool for the cophylogeny reconstruction problem. Algorithms for molecular biology : AMB **5**(1), 16 (2010). doi:10.1186/1748-7188-5-16
70. Jousselin, E., Desdevises, Y., Coeur d'acier, A.: Fine-scale cospeciation between ¡i¿Brachycaudus¡/i¿ and ¡i¿Buchnera aphidicola¡/i¿: bacterial genome helps define species and evolutionary relationships in aphids. Proceedings. Biological sciences / The Royal Society **276**(1654), 187–196 (2009). doi:10.1098/rspb.2008.0679
71. Baldo, L., Ayoub, N.A., Hayashi, C.Y., Russell, J.A., Stahlhut, J.K., Werren, J.H.: Insight into the routes of Wolbachia invasion: High levels of horizontal transfer in the spider genus Agelenopsis revealed by Wolbachia strain and mitochondrial DNA diversity. Molecular Ecology **17**(2), 557–569 (2008). doi:10.1111/j.1365-294X.2007.03608.x
72. Moran, N.A., McLaughlin, H.J., Sorek, R.: The Dynamics and Time Scale of Ongoing Genomic Erosion in Symbiotic Bacteria. Science **323**(5912), 379–382 (2009). doi:10.1126/science.1167140
73. Chevignon, G., Boyd, B.M., Brandt, J.W., Oliver, K.M., Strand, M.R.: Culture-Facilitated Comparative Genomics of the Facultative Symbiont Hamiltonella defensa. Genome Biology and Evolution **10**(3), 786–802 (2018). doi:10.1093/gbe/evy036
74. Russell, J.A., Weldon, S., Smith, A.H., Kim, K.L., Hu, Y., Łukasik, P., Doll, S., Anastopoulos, I., Novin, M., Oliver, K.M.: Uncovering symbiont-driven genetic diversity across North American pea aphids. Molecular Ecology **22**(7), 2045–2059 (2013). doi:10.1111/mec.12211
75. Escobar-Zepeda, A., De Le??n, A.V.P., Sanchez-Flores, A.: The road to metagenomics: From microbiology to DNA sequencing technologies and bioinformatics. Frontiers in Genetics **6**(DEC), 348 (2015). doi:10.3389/fgene.2015.00348
76. Wang, Z., Wu, M.: A phylum-level bacterial phylogenetic marker database. Molecular Biology and Evolution **30**(6), 1258–1262 (2013). doi:10.1093/molbev/mst059
77. François, S., Filloux, D., Fernandez, E., Ogliastro, M., Roumagnac, P.: Viral metagenomics approaches for high-resolution screening of multiplexed arthropod and plant viral communities. In: Viral Metagenomics, pp. 77–95. Springer, ??? (2018)
78. Colman, D.R., Toolson, E.C., Takacs-Vesbach, C.D.: Do diet and taxonomy influence insect gut bacterial communities? Molecular Ecology **21**(20), 5124–5137 (2012). doi:10.1111/j.1365-294X.2012.05752.x
79. Jing, X., Wong, A.C.N., Chaston, J.M., Colvin, J., McKenzie, C.L., Douglas, A.E.: The bacterial communities in plant phloem-sap-feeding insects. Molecular Ecology **23**(6), 1433–1444 (2014). doi:10.1111/mec.12637
80. Moran, N.A., Munson, M.A., Baumann, P., Ishikawa, H.: A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proceedings of the Royal Society B: Biological Sciences **253**(1337), 167–171 (1993). doi:10.1098/rspb.1993.0098
81. Moran, N.A., Yun, Y.: Experimental replacement of an obligate insect symbiont. Proceedings of the National Academy of Sciences **112**(7), 2093–2096 (2015). doi:10.1073/pnas.1420037112
82. Moran, N.A., von Dohlen, C.D., Baumann, P.: Faster evolutionary rates in endosymbiotic bacteria than in cospeciating insect hosts. Journal of Molecular Evolution **41**(6), 727–731 (1995). doi:10.1007/BF00173152
83. Chong, R.A., Moran, N.A.: Evolutionary loss and replacement of Buchnera, the obligate endosymbiont of aphids. ISME Journal **12**(3), 898–908 (2018). doi:10.1038/s41396-017-0024-6
84. Rock, D.I., Smith, A.H., Joffe, J., Albertus, A., Wong, N., O'Connor, M., Oliver, K.M., Russell, J.A.: Context-dependent vertical transmission shapes strong endosymbiont community structure in the pea aphid, Acyrthosiphon pisum (2017). doi:10.1111/mec.14449
85. Oliver, K.M., Degnan, P.H., Burke, G.R., Moran, N.A.: Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. Annual Review of Entomology **55**(1), 247–266 (2010). doi:10.1146/annurev-ento-112408-085305
86. Moran, N.A., McCutcheon, J.P., Nakabachi, A.: Genomics and Evolution of Heritable Bacterial Symbionts. Annual Review of Genetics **42**(1), 165–190 (2008). doi:10.1146/annurev.genet.41.110306.130119
87. McCutcheon, J.P., Moran, N.A.: Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology **10**(1), 2670 (2011). doi:10.1038/nrmicro2670
88. Łukasik, P., Guo, H., van Asch, M., Henry, L.M., Godfray, H.C.J., Ferrari, J.: Horizontal transfer of facultative endosymbionts is limited by host relatedness. Evolution **69**(10), 2757–2766 (2015). doi:10.1111/evo.12767
89. Henry, L.M., Maiden, M.C.J., Ferrari, J., Godfray, H.C.J.: Insect life history and the evolution of bacterial mutualism. Ecology Letters **18**(6), 516–525 (2015). doi:10.1111/ele.12425

90. Valette, V., Bitome Essono, P.Y., Le Clec'h, W., Johnson, M., Bech, N., Grandjean, F.: Multi-infections of feminizing Wolbachia strains in natural populations of the terrestrial isopod Armadillidium Vulgare. PLoS ONE **8**(12), 82633 (2013). doi:10.1371/journal.pone.0082633
91. Brandt, J.W., Chevignon, G., Oliver, K.M., Strand, M.R.: Culture of an aphid heritable symbiont demonstrates its direct role in defence against parasitoids. Proceedings of the Royal Society B: Biological Sciences **284**(1866), 20171925 (2017). doi:10.1098/rspb.2017.1925

Additional Files

Additional file 1 — Read depth of reference genomes for each sample
Additional file 2 — Sets of membrane protein genes selected for phylogenetic inference
Additional file 3 — Summary of unmapped reads taxonomic assignation by Centrifuge
Additional file 4 — Comparison of *Buchnera aphidicola* phylogeny inferred by gene-set phylogeny and whole genome clustering
Additional file 5 — Comparison of *Buchnera aphidicola* and mitochondrial phylogenies
Additional file 6 — Results of phylogenetic reconciliation by Jane

Chapitre 4

Développement d'une méthode d'assemblage guidé par référence en contexte métagénomique

4.1 Contexte dans le cadre de la thèse

Le chapitre précédent est un exemple du potentiel des données métagénomiques pour l'étude fine du fonctionnement et de l'évolution des holobiontes. Il combine de manière innovante des méthodes bioinformatiques disponibles en génomique et en métagénomique, mais souffre, à travers deux points particuliers, des limites des méthodes actuellement disponibles présentées au cours du chapitre 2. Le premier de ces points est la nécessité de disposer d'un génome de référence de qualité pour rendre possible la détection de variants génomiques. Si dans le cas du puceron du pois nous avons pu accéder ou assembler de tels génomes, il n'en est pas le cas de tous les systèmes, car les génomes bactériens sont globalement mal décrits, et l'assemblage métagénomique est une tâche difficile. Le second point qui n'a pas été abordé dans l'étude précédente est l'existence de polymorphismes structuraux dans les données métagénomiques, qui pourraient pourtant avoir d'importants impacts fonctionnels. En raison de la faible longueur des lectures Illumina, il s'agit également d'une tâche difficile pour laquelle peu d'outils sont disponibles, et aucun dans le cadre de la métagénomique.

Dans ce chapitre, qui est rédigé comme une ébauche d'un article scientifique, nous proposons une méthode visant à résoudre ces deux problèmes. Il s'agit d'une méthode d'assemblage guidé par un génome de référence, qui permet l'identification de différents variants structuraux au sein d'un métagénome, et facilite l'assemblage de génomes à partir de lectures métagénomiques. Cette méthode repose sur deux étapes : l'assemblage de lectures recrutées par l'alignement sur un génome de référence, puis l'utilisation d'une version améliorée d'une version améliorée du programme *MindTheGap*, précédemment développé dans l'équipe [Rizk et al., 2014], qui permet de reconstituer les régions non assemblées, et d'identifier la structure du génome recherché. Nous avons étendu les fonctionnalités de *MindTheGap*, initialement prévu pour assembler des insertions, pour permettre de reconstruire les séquences manquantes entre plusieurs contigs, sans faire d'hypothèse sur leur ordonnancement dans le génome.

Nous proposons deux applications de cette méthode, centrées sur le complexe du puceron

du pois. Dans la première, nous appliquons notre méthode à l'assemblage du génome de *Buchnera aphidicola*, le symbiote obligatoire des pucerons, en utilisant un génome éloigné comme guide. Nous évaluons ainsi la capacité de notre approche à assembler un génome inconnu à partir d'une référence distante, tout en disposant d'un génome de référence pour validation. Notre approche permet un assemblage complet des 50 échantillons du jeu de donnés utilisé précédemment, et montre des performances supérieures à l'assembleur métagénomique Megahit [Li et al., 2015], tout en étant en moyenne 6 fois plus rapide. Dans un second temps, nous présentons les résultats de l'assemblage du bactériophage APSE, qui est associé au symbiote facultatif *Hamiltonelle defensa*, et est connu pour présenter une cassette de virulence très variable et très certainement impliquée dans la protection des pucerons à l'égard de parasitoïdes [Degnan and Moran, 2008]. Dans ce cadre, *MindTheGap* permet de reconstruire ces génomes de phage, et est capable d'identifier différents variants de cette cassette au sein d'un même échantillon. En particulier, un nouveau variant de ce phage est identifié, et nous proposons un premier résultat sur le contenu génique de sa cassette de virulence.

Ainsi, bien qu'il ne s'agisse pas d'une version finale de l'article, ce chapitre présente le potentiel de *MindTheGap* pour reconstruire des génomes et leurs variations structurales à partir de métagénomes. Il ouvre des perspectives à la fois sur la communauté symbiotique du puceron du pois, dans lequel les cortèges symbiotiques ont des effets phénotypiques importants pour leurs hôtes, et sur l'étude de communautés plus complexes, où *MindTheGap* pourrait être un outil prometteur pour reconstruire des génomes et leur variabilité structurale.

4.2 Introduction

The advances of molecular techniques revealed the importance of microorganisms in every ecosystem. A wide range of microbial communities are now studied to reveal how microbes shape functional and evolutionary components of ecosystems such as soils, seas and holobionts. In particular, whole-genome metagenomic sequencing makes it possible to understand the full functional potential of microbial communities by accessing the whole genomic sequence of both culturable and unculturable microbes. However, extracting relevant information from complex metagenomic datasets is a challenging task. Current metagenomic datasets are a mixture of short reads originating from different species. Thus, reconstructing genomes from metagenomic data requires two steps : the assembly of reads into longer sequences, and the partitioning of sequences based on their taxonomic origin.

Metagenomic assembly consists in forming contigs prior to the taxonomic binning of sequences. Many recent software are devoted to this task [Nurk et al., 2016, Peng et al., 2012, Li et al., 2015]. However, because of the high complexity of such data, assembling contigs from metagenomic reads is challenging. Metagenomic assembly is subject to the same limitations as regular genomic assembly : highly repetitive regions of the genome cannot be assembled using short reads and break the assembly contiguity. In addition, metagenomic samples show two levels of genomic diversity which hamper the assembly task. First, microbial communities are made of a mixture of species with uneven abundances. Second, some genomic diversity and polymorphism exist within each species. As a result, metagenomic assemblies are very fragmented because of homologous regions between microbial species and highly variable regions within the species. Large benchmarking of several metagenomic assemblers highlights the difficulty of this task, especially when closely related species are present in the microbial mixture [Sczyrba et al., 2017]. After assembly, metagenomic contigs can be grouped together based on their nucleotidic composition, coverage, and/or by using existing reference databases, resulting in bins of contigs at a given operational taxonomic level [Alneberg et al., 2014b, Wu et al., 2014, Albertsen et al., 2013]. However, this step is error-prone due to both the intra-specific variability and the regions shared between metagenomic species, which can result in incorrectly binned contigs.

Because of the complexity of microbial communities, metagenomic assemblies are fragmented, and obtained at the cost of a high computational resource usage. An alternative to this problem would be to operate on a subset of metagenomic reads assigned in a first step to a given species. Binning methods relying on the nucleotidic composition of reads cannot be applied to the current Illumina reads because of their short length [Teeling et al., 2004]. Alternatively, it is possible to select reads by reference-based approaches, which struggle to classify reads from badly known species, and hardly scale up to large datasets when based on alignment methods [Huson et al., 2016]. A relevant strategy to assemble a given genome from metagenomic data is to map reads against the closest available reference genome to assemble new contigs. The quality of the assembly is therefore highly dependent on the distance with the reference genome. In particular, any genomic region absent from the reference genome

will be missed. Alternatively, reads originating from other organisms can be removed from the read set by mapping, enabling the genomic assembly of the reads belonging to the genome of interest. However this approach can only be applied to simple communities with reference genomes available, such as holobionts of moderate complexity [Burke and Moran, 2011].

One of the main pitfalls of metagenomic assembly approach is the coassembly of different genomes. However, only some specific microbes among the microbial community may be of interest for researchers. This is the case for the investigation of most host-symbiont interactions in holobionts, in which a few microbes may be responsible for important host phenotypic changes, or for the study of new pathogenic strains of known microbes. In all these use cases, functional, structural or phylogenetic genomic analyses require the assembly of a new genome of interest for metagenomic data. In that context, neither *de novo* metagenomic assembly nor assembly from reads selected by reference alignment are able to return assemblies of good quality. Nonetheless, it seems possible to use the best of these two strategies, by selecting reads from regions of homology with reference genomes, and using *de novo* assembly to reconstruct the missing regions.

Several tools, such as MITObim [Hahn et al., 2013], LOCAS [Klein et al., 2011], Pilon [Walker et al., 2014] or IMR/DENOM [Gan et al., 2011], were designed following this idea, combining reference alignment and *de novo* assembly. However, these tools show some limitations because of which they are not adapted to metagenomic data. MITObim [Hahn et al., 2013] is devoted to the assembly of mitochondrial genomes, significantly shorter than bacterial genomes. It works in a iterative manner, extending contigs at best of a read length at each iteration. The time required to reconstruct large insertions is therefore prohibitive. In IMR/DENOM [Gan et al., 2011], a similar iterative strategy is used to detect Short Nucleotide Polymorphism (SNP) compared to a reference sequence. In addition, a whole-genome assembly is performed in a second time to detect large structural variations. While this is a very efficient approach for the genome of *Arabidopsis thaliana* studied in this case, several features of the pipeline are not appropriate for metagenomic data. The alignment of *de novo* assembled contigs would require in that case a costly and error prone metagenomic alignment. In addition, the method is reported to perform badly with heterogeneous samples [Landman et al., 2014], and is openly available. Similarly, Pilon is designed to correct reference genomes by calling short nucleotide variants and locally assembling large insertions. The local assembly step recruits pairs of reads with one read of the pair properly mapping on the flanks of the insertion. Therefore, the tool requires paired end reads dataset, and, more importantly, is unable to assemble regions larger than the range of paired-end reads. It also uses the structure of the reference genome, and is therefore unable to detect structural rearrangements. LOCAS [Klein et al., 2011] uses a comparable approach with its SUPERLOCAS module. Reads mapping on reference genome regions and orphan reads are assembled in distinct overlap graphs. Compared to *De Bruijn* graphs, overlap graphs require time consuming overlap computation. While this is not a problem for the low coverage resequencing datasets targeted by LOCAS, this tool hardly scales up to recent high volume metagenomic datasets. In addition, LOCAS does non take into account the specificities of metagenomic samples, and is unable to return more than one assembled sequences for a given genomic region. Finally, other tools,

such as SHEAR [Landman et al., 2014] or RECORD [Buza et al., 2015], rely heavily on the reference genome and give better results when applied to a highly similar genome. RECORD adds to the readset pseudoreads generated from a reference genome very similar to the target, assembled the extended readset into contigs. Contigs are then aligned to the reference genome to produce a corrected version.

Existing reference-guided assembly tools has limitations, especially when applied to metagenomic data. All of these tools work under the assumption of genomic homogeneity, which is very rarely satisfied in metagenomic context. Moreover, the architecture of the reference genome is used as a starting-point for the assembly in most of the case, and some tools are designed either for short genomes of small resequencing datasets.

In this work, we present a solution for the assembly of a genome of interest from metagenomic data, in a reference-based manner. This method is able to return several different solutions reflecting the metagenomic diversity, can assemble very large insertions, and makes no assumption on the ordering or direction of regions homologous with the reference. The method is based on two main steps. First, a portion of the reads homologous to the reference genome are recruited by mapping and assembled into contigs. Second, a targeted assembly is performed using the software *MindTheGap* [Rizk et al., 2014] with all the metagenomic reads, resulting in a *de novo* assembly of the missing regions. The first step uses a reference genome to facilitate the assembly of backbone contigs. Because no assumption is made on the ordering of contigs, putative structural variations are detected.

Using a closely related reference genome will result in a long assembly even at the first step. In the opposite case, when using a remote reference genome, a smaller part of the target genome is assembled, but the second step allows to enhance the genome by assembling regions absent from the reference genome. Our implementation of the gapfilling is able to return multiple sequences between a contig and one or several other contigs. As a result, it is possible to return several alternative genomes, accounting for the structural genomic diversity inside the sample, making the pipeline especially appropriate to metagenomic data. The task of *de novo* metagenomic assembly is simplified by using contigs built in the first step as starting points. This greatly reduces the computational cost and the number of assembly errors. This method fits well the case of targeted genome assembly from metagenomic samples.

We applied this method to reconstruct genomes from several metagenomic samples of the pea aphid *Acyrthosiphon pisum*. We first demonstrated the ability of *MindTheGap* to assemble the genome of the obligatory symbiont *Buchnera aphidicola* in a single contig, using a remote genome as a primer. The method was then applied to the bacteriophage APSE, which is associated to the facultative symbiont of the pea aphid *Hamiltonella defensa*, and is known to show structural variations of a virulence cassette, which confers a protection to the aphid against its main natural ennemies.

4.3 Material and Methods

4.3.1 Targeted assembly for metagenomic data

Strategy overview

The method described in this work relies on a two-step pipeline, described in Figure 3.

The first step uses a remote reference genome to build an incomplete but trustworthy assembly, matching with the conserved regions of the genome. The second step uses the whole set of metagenomic reads to extend the previously assembled contigs and form a complete assembly, without any *a priori* on the order and orientation of contigs. The result of the pipeline is a genome graph encompassing the stuctural diversity detected on the assembled genome. This graph can be exploited by extracting contigs, or paths of the graph that represent different strains.

Assembly of backbone contigs

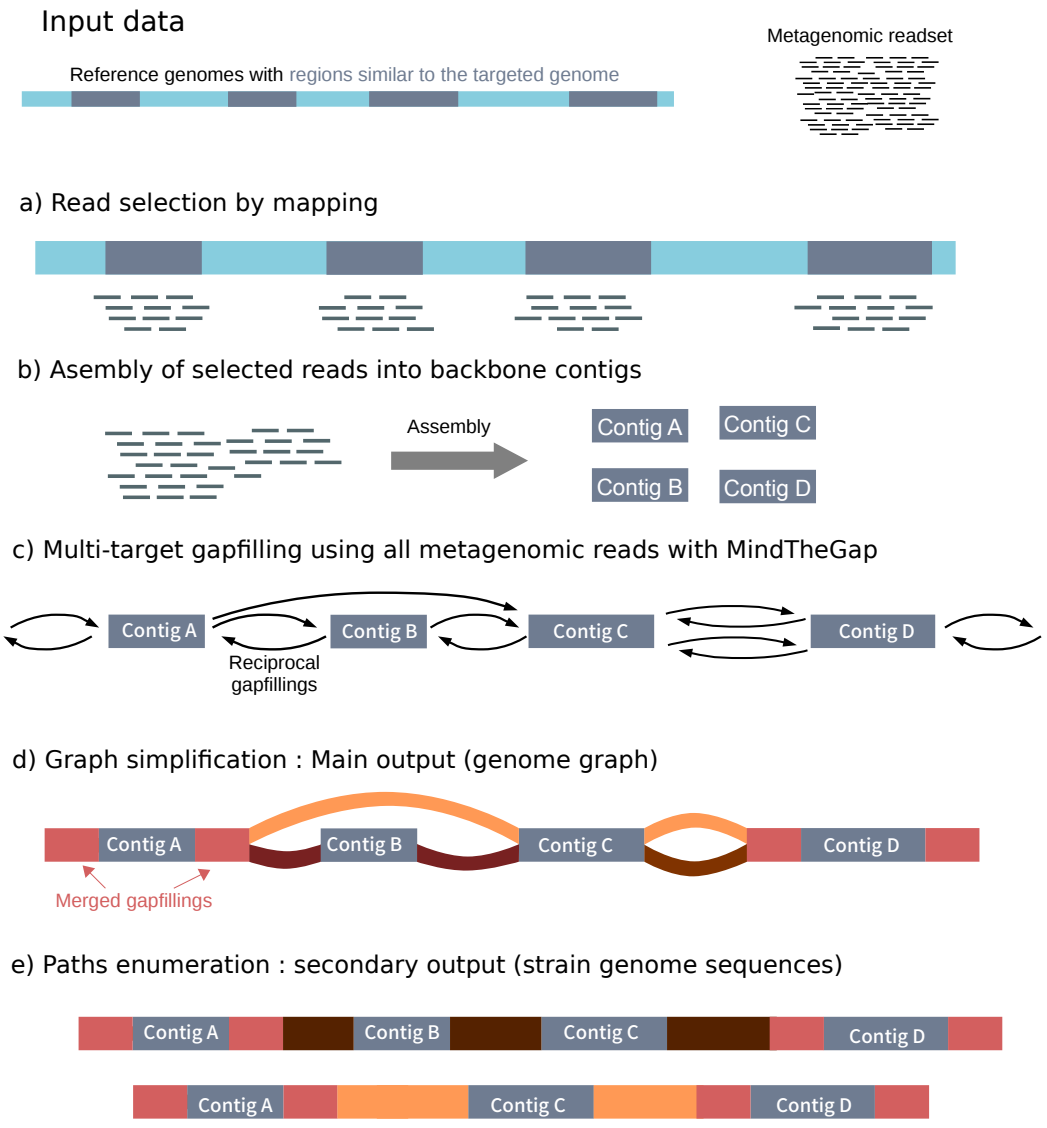
The first step requires a metagenomic readset and a reference genome, and returns contigs that are assembled using reads mapped on the reference. All metagenomic reads are mapped against the reference genome using BWA MEM [Li and Durbin, 2009], and the mapped reads are kept and *de novo* assembled using the Minia [Chikhi and Rizk, 2012] assembler. Although any assembler can be used in this step, we use Minia [Chikhi and Rizk, 2012] for its low memory footprint, and its assembly algorithm similar to the one used in the second step of the method. The number of mapped reads and the length of the assembly depend on the sequence similarity between the reference genome and the targeted genome. The goal of this step is to generate high quality contigs, that can reliably be used for the upcoming gapfilling. To ensure this, we set up Minia with more stringent parameters than for an usual assembly task, with higher kmer sizes and abundance thresholds. This helps to remove erroneous kmers from the de Bruijn graph, and build reliable contigs, at the cost of genome coverage and contiguity. In the same goal, only contigs longer than a user-defined threshold (500 bp by default) are kept, to ensure only high quality contigs are used for the upcoming step.

Parallel gapfilling with MindTheGap

The essential step of the pipeline is the gapfilling between backbone contigs, which enables the assembly of regions absent from the reference genome. The first step returns a partial assembly from reads mapping on the reference genome. In this second step, the whole readset is used to increase the contiguity and the length of the assembly, taking into account structural variants. This is made possible by a targeted assembly of the whole readset using the previously assembled contigs as primers. This step does not require the ordering of contigs, since all possible combinations are tested during gapfilling. As a result, structural variants can be detected, either compared to the reference genome or within the sample.

This step is based on the software *MindTheGap*. It was originally developed for the

Figure 3 – ***Overview of the MindTheGap reference-guided assembly pipeline***
a) Reads of the metagenomic sample with similarity with the reference genome are recruited by mapping.
b) Recruited reads are assembled into backbone contigs.
c) Gapfilling is performed between all the assembled contigs, with no assumption on the contigs order. (See figure 4).
d) The assembly graph is simplified, resulting in a genome graph that can be used for further analyses.
e) Optionally, paths of the graph can be enumerated to output linear assemblies.



detection of insertion events [Rizk et al., 2014]. *MindTheGap* is made of two main modules. The first one (*find module*) detects breakpoints by analyzing kmer coverage along a reference genome, and returns two kmers from either side of each insertion. The second module (*fill module*) performs a local assembly for each pair of breakpoint event kmers, resulting in one or several insertion sequences.

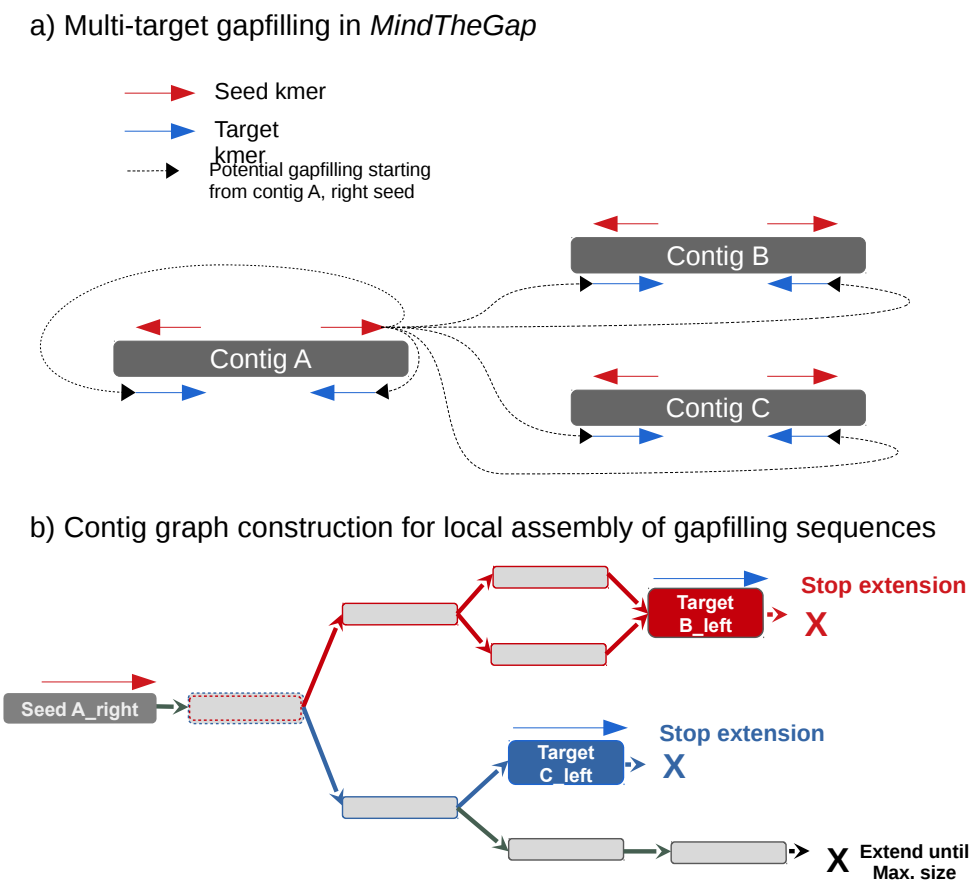
In this work, we took advantage of the second module of *MindTheGap* and adapted to the problem of reference-guided assembly. This module has been modified to make possible the gapfilling between a seed kmer and **multiple** target kmers, enabling the "all versus all" gapfilling within a set of contigs with only a linear increase of the runtime (compared to a quadratic increase for a naive "all versus all" gapfilling). The resulting algorithm is presented in Figure 4. A seed kmer is extracted at the end of each contig and its reverse-complement, resulting in a set of $2n$ kmers for n contigs. Similarly, a set of $2n$ target kmers is extracted at the beginning of each contig and its reverse-complement. For each seed kmer, a contig graph is created by starting from the seed kmer and performing a breadth first traversal of the *De Bruijn* graph representation of the whole readset. Contigs are consensus sequences returned by removing graph motifs such as bubbles (SNPs) and tip-ends (errors). In the contig graph, contigs are nodes, and edges represent the existence of a $k-1$ nucleotide overlap between two contigs. The creation of the contig graph is similar to the algorithm used in *Minia* [Chikhi and Rizk, 2012]. The traversal is stopped when the graph becomes too large (total assembled nucleotides) or too complex (number of contigs), following user-defined parameters. Importantly, if one of the target kmers is found during the contig graph construction, that contig is not extended further, avoiding redundant contig assembly, and saving time and memory. After the contig graph has been built, target kmers are searched within this contig graph, and insertion sequences joining the two breakpoint ends are built, by retraversing the contig graph from the seed kmer to contigs containing a given target kmer. For each seed-target couple, if several solutions are returned, redundant solutions above a 95% identity threshold are removed. Thanks to this multi-target version of the algorithm, only $2n$ contig graph constructions are necessary to search possible sequences between all pairs of contigs, instead of n^2 with the naive approach.

The whole process is parallelized by dispatching the $2n$ starting kmers to different threads. The main output is a genome graph in the GFA format (Graphical Fragment Assembly, `https://github.com/GFA-spec/GFA-spec`), giving the overlap relationships between contigs an their gapfillings. GFA is a file format designed for a unified representation of different forms of genomic graphs. It is supported by several assemblers [Koren et al., 2017, Simpson et al., 2009], and can be manipulated by the Bandage viewer [Wick et al., 2015] or programming libraries [Gonnella and Kurtz, 2017]. In that GFA output, *MindTheGap* indicates contigs and gapfillings as nodes, and overlaps as edges.

Graph simplification and visualisation

In order to return a standard fasta assembly, the genome assembly has to be processed. The complexity of the graph is reduced on several steps using a post-treatment program,

Figure 4 – ***Gapfilling a set of contigs using MindTheGap fill module.***
a) Seed and target kmers are extracted from the n input contigs, resulting in 2 sets of 2n kmer, seed and target ones. For each seed kmer, an instance of the gapfilling presented in b) is run.
b) A graph of contigs is built starting from a seed kmer. Extension is stopped when a target kmer of another contig if encountered, or a maximum assembly size is reached.
c) Gapfilling sequences assembled starting from contig A right seed. In that instance, 2 gapfilling sequences are assembled toward contig B, and one toward contig C



summarized in Figure 5.

First, it is likely that two contigs are linked in the graph by two gapfillings with reverse-complement sequences, one starting from the left contig and the other one starting from the right contig. Such reciprocal links are removed, when their sequence similarity is over a 95% threshold.

Secondly, several gapfilling sequences may be identical on extremities of their sequence, resulting in redundant sequences in the graph. A node merging algorithm is applied, in order to return contigs that do not share large identical subsequences. Sets of sequences sharing the same 100 first nucleotides are built. Within each set, the sequences are then compared to find the first divergence between all sequences. A new node is added to the graph, containing the repeated portion of the sequences, and repeated nodes are shortened accordingly. This process is applied iteratively to every node, including the newly created nodes, for which a subset of neighbors may still show identical sequences.

Finally, simple linear paths in the graph are merged, nodes whose length is lower than 500 bp are removed, and highly branching nodes (connected to more than 5 contigs) are cut. The resulting graph is a good representation of the *MindTheGap* assembly, and nodes can be extracted to be used as regular contigs.

Path enumeration for strain identification

Depending on the complexity of the dataset, several paths may be possible in each connected component of the cleaned graph, reflecting several potential strains. For each connected component of the graph, these different paths are enumerated. Currently, all maximal paths are enumerated. When possible, circular paths were preferred to linear paths, which implies that this process is for now optimized for circular genome assembly.

Implementation and availability

MindTheGap has been officially released in version 2.1, enabling the so-called "contig mode" for reference-guided assembly (https ://github.com/GATB/MindTheGap). *MindTheGap* is written in C++ using the GATB library [Drezen et al., 2014] (https ://github.com/GATB/gatb-core). The GATB library provides algorithms for the analysis of NGS datasets with high performances and a low memory footprint.

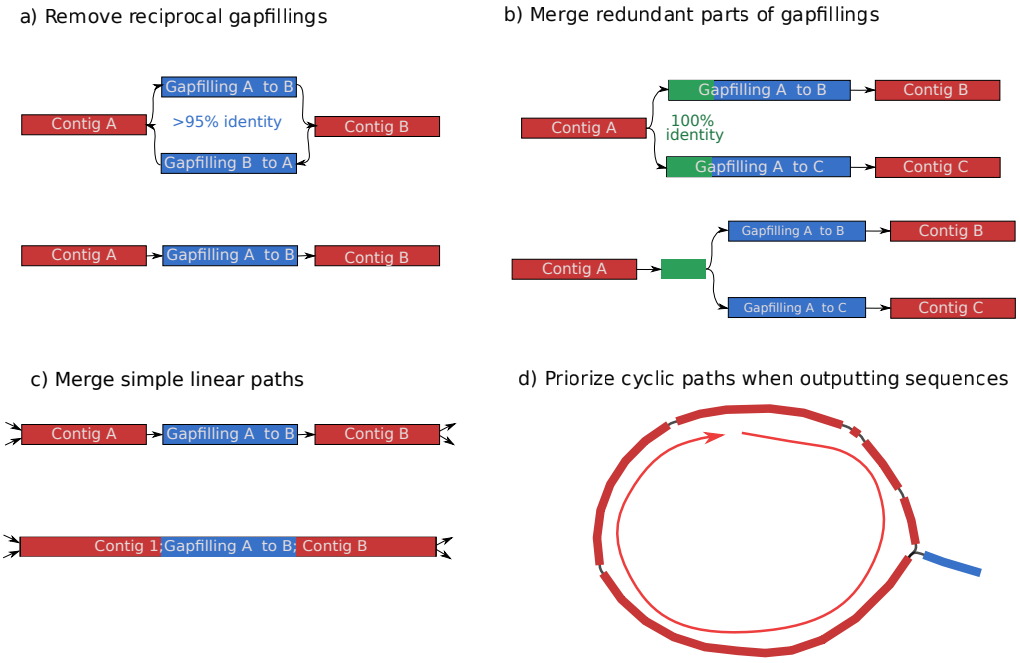
The graph simplification is performed using Julia written scripts, available upon request. They rely in the *Lighgraphs.jl* library (https ://github.com/JuliaGraphs/LightGraphs.jl/), that offers a lightweight and customizable graph implementation.

A complete pipeline including mapping, assembly and gapfilling is also available as a Python script distributed along with *MindTheGap*.

4.3.2 Application to metagenomic datasets

In this study, we applied *MindTheGap* to two different assembly scenarios, in the context of the pea aphid holobiont. We considered 50 metagenomic samples of paired end 100bp *Illumina*

Figure 5 – ***Graph simplifications applied to MindTheGap output***
a) : Reciprocal gapfillings with more than 95% identity are merged.
b) : Shared sequences between gapfillings originating from or leading to a same contig are merged to reduce sequence redundancy
c) : Simple linear paths, with no branching nodes, are merged into a single node.
d) : Graph traversal enumerates all possible paths through the graph. When detected, only circular paths are outputted.



	Assembly			Gapfilling	
Samples	**Kmer size**	**Min. abundance**	**Contig length**	**Kmer size**	**Min. abundance**
Individual	61	10	400	51	5
Pool	81	20	400	71	10

Table 3 – *Sets of parameters used for the guided assembly of Buchnera*

reads. Among them, 32 were individual sequencing of a single aphid clone, and 18 were pool sequencing of individual aphids originating from the same population. Pool sequencing samples are expected to show more genomic diversity than individual clones, and therefore to be more challenging for genome assembly. These datasets have already been studied in a previous work, in which the microbiota of each aphid sample was detailed [Guyomar et al., 2018]. The number of reads per dataset is ranging from 65 to 700 million, with pool sequencing datasets generally showing a higher coverage than individual datasets. However, more than 90% of the reads originate from the host, and are not relevant when focusing on symbiont genomes. This particular fact motivates the choice of a targeted assembly technique, which does not require to assemble pea aphid reads.

Guided assembly of *Buchnera aphidicola*

MindTheGap was first applied to the assembly of the bacterial genome of *Buchnera aphidicola*, obligatory symbiont of the pea aphid. As an obligate symbiont, it is present in all of the 50 samples.

As a remote reference genome for the guided assembly process, we used *Buchnera aphidicola* G002 from *Myzus persicae*. It shares 80.74% of ANI (Average Nucleotide Identity) with the genome of *B. aphidicola LSR1* from the pea aphid, that was used as a reference to assess the quality of our assemblies.

In order to take into account the differences of coverage between individual and pool sequencing samples, we used two sets of parameters, summarized in Table 3. In all cases, we chose more stringent parameters for assembly than for gapfilling, in order to first assemble a reduced set of trustworthy contigs, that can be extended during the gapfilling step, able to assemble low coverage variants and to increase contiguity.

Selecting representative strains from assembly graphs

Bacteria sequenced in a metagenomic samples do not necessarily share the exact same genotype. This is especially true for pool sequencing, where different aphid hosts and their microbiota are sequenced together. As a consequence, the assembly graph may exhibit polymorphism in the form of bubbles in the graph. To compare with traditional assembly methods, a selection of a representative paths among all the possible sequences was necessary.

An *adhoc* script was developed to convert the gapfilling graph into a genome assembly of *Buchnera*. Only the largest connected components (size greater than 1 kb.) were analyzed. For each connected components, all possible maximum paths were computed. Since several paths with very similar sequences may be generated, a clustering step was used to detect

potential outliers in the path set. As suggested and implemented is the Python module *pyani*, a metric based on the Hadamard product of average nucleotide identity (ANI) and alignment length, was used as a similarity measure between paths and a distance matrix between every possible paths was built. A standard clustering procedure was applied to this matrix using hierarchical clustering, and the clustering tree was cut under a threshold of 99.9% of similarity score, resulting in one cluster if all sequences are highly similar, or more if different sequences were computed from the graph.

For each sample, a path was selected as the assembled *Buchnera* genome for this sample. Except for the *Lus* sample for which a path not contaminated by *Hamiltonella* contigs was chosen, paths were selected randomly, and each assembly only represents a possible genotype within the sample.

Comparison with other approaches

The results were compared to those of a usual approach to assemble a particular genome from metagenomic data. A complete *de novo* assembly was performed for each sample using MegaHit [Li et al., 2015] and *Buchnera* contigs were selected by a Blast alignment againt the genome of *Buchnera aphidicola APS*. We used the *blastn* algorithm, and kept contigs with at least 50% of the length covered by Blast hits. The quality of each assembly was assessed using Quast [Gurevich et al., 2013] and the reference genome of *Buchnera aphidicola APS* from *A. pisum*. Similarly to what was done with *MindTheGap*, we did not include contigs smaller than 1 kb, mainly associated with plasmid sequences. We compared different assembly metrics such as reference genome coverage, number of contigs, assembly length and NGA50. NGA50 is a metric similar to N50 taking into account the reference genome length, and eventual misassemblies detected by a reference genome alignment.

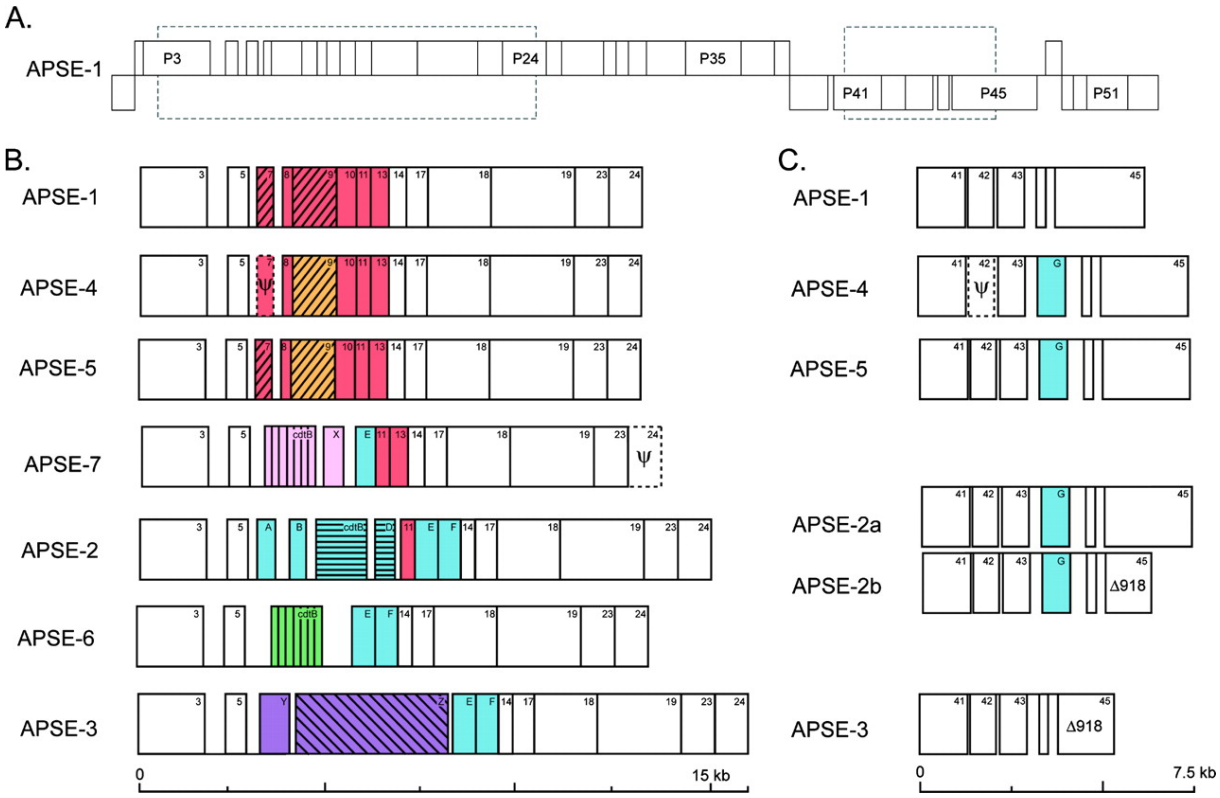
We also computed these metrics for the Minia assembly performed prior to the gapfilling step, mainly as a way to measure the relative contributions of reference-based assembly and gapfilling to the final assembly.

4.3.3 Guided assembly of the bacteriophage APSE

The bacteriophage APSE is strictly associated with *Hamiltonella defensa*, a secondary symbiont of *Acyrthosiphon pisum* and other aphids and sap-feeder insects [Degnan and Moran, 2008]. Several variants of this phage have been described in the pea aphid and other insects. These variants differ by a highly variable region of 10 to 15 kb, called a virulence cassette, and for which seven variants have been described to date [Degnan and Moran, 2008]. The gene content of these cassettes has been described by *Degnan et al*, and is shown in figure 6. APSE has been shown to be involved in the natural enemy resistance conferred by *Hamiltonella* to its host, and brings variable levels of protection depending on the considered phage variant [Brandt et al., 2017].

During the previous work presented in Chapter 3, the presence of APSE was investigated by aligning metagenomic reads on the APSE-1 reference genome. Because a large genome portion is shared between APSE variants, the coverage of APSE is a good approximation

Figure 6 – ***Gene content of the virulence cassette for the APSE variants known to date.*** *From Degnan et al [Degnan and Moran, 2008]*



of the abundance of the different variants. 18 samples out of 50 show an average genomic coverage greater than 30X, which is a reasonable limit to make possible the assembly. Most samples are highly covered, with a maximum coverage of 8,000X. The *MindTheGap* pipeline was applied using different parameters, depending mainly on the sample coverage. We used the genome of APSE1 as a reference for the mapping step of the pipeline.

The standard graph simplification procedure was performed on the genome graph produced by the pipeline. In order to output genomic graphs only representing the cassette structural variation, an extra-step of local alignment of alternative paths was performed for each polymorphic position of the graph, and only variants with less than 95% of identity. Phage genome(s) were then generated from the cleaned graph.

Genome comparison and annotation

The genomes were compared using the multiple alignment software Mauve [Darling et al., 2004] and its *progressiveAlignment* algorithm. The guide tree built by Mauve was used as a indication of the distances between phage genomes.

For each phage variant, a specific gene repertoire was created using the variant annotation proposed by [Degnan and Moran, 2008], and the available APSE genomes. In addition, a core gene set was built from the genes in conserved regions (from genes P1 to P5, P14 to P40 and P46 to P54.) Sequences of the core gene set were extracted from the APSE-1 genome. A Blast

alignment was performed between each genome and the APSE pangenome, and the blast hits coordinates were used to annotate the putative positions of genes on the newly assembled APSE genomes.

Comparison with alternative approaches

Two tools were compared to *MindTheGap* to reconstruct APSE genomes using the sample Gt. *MITObim* [Hahn et al., 2013] was used, using the APSE1 genome as a bait for the assembly. The software was run using the `-quick` option and 31 iterations. Similarly, Pilon [Walker et al., 2014] was used to attempt to correct the APSE1 reference genome using reads from the *Gt* sample.

4.4 Results

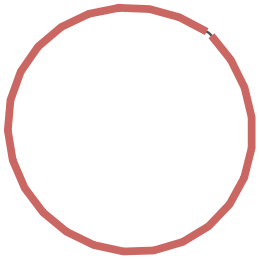
4.4.1 Single chromosome assembly of *Buchnera aphidicola* from metagenomic data

MindTheGap was first used a reference-guided assembly tool for the *Buchnera* obligatory symbiont of the pea aphid. It was applied to 50 metagenomic samples. A remote reference genome (80% ANI) was used a guide for the assembly, and the resulting genomes were compared to the closest reference available as a validation of our approach.

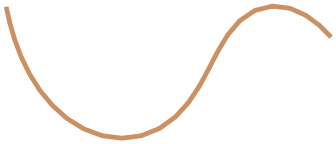
Graph representation encompasses *Buchera*'s genome and its variability in a single component

The genome graph was first analyzed by observing the number of connected components. Between two and seven connected components were generated, with 84% of samples assembled in two connected components. In 49 of the 50 samples, the accumulated length of the first connected component was over 640,000 bp, which was the targeted genome size. For the L3Ml04 sample, two connected components were obtained for a total length of 636 kb. The size of the other connected components was comprised between 400 and 9,000 bp. These sequences were either sequences of the pLeu plasmid of *Buchnera* (length 7,786bp) or unconnected contigs. This indicates that in almost all of our assembly attempts, the targeted genome was successfully assembled in a single connected component of the graph.

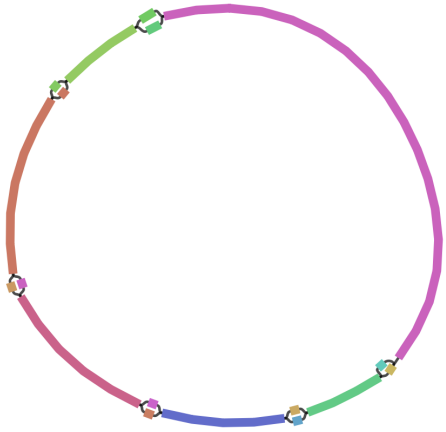
The size of this first component was computed by adding the lengths of every sequence and substracting the overlap between node. It was ranging between 95.7% and 122.7% of the *Buchnera aphidicola str. LSR1* genome, and was of 107% on average. A length smaller than the genome size may indicate missing regions in the assembly, while a larger size is to be expected since polymorphism results in alternative paths in the graph. Since the assembly size does not exceed 122% of the expected genome size, we can assume that the number of sequences assembled apart from the *Buchnera*'s genome is reasonably low. This also illustrates the efficiency and specificity of the targeted assembly approach, since only a small part of the whole dataset was assembled into this graph.



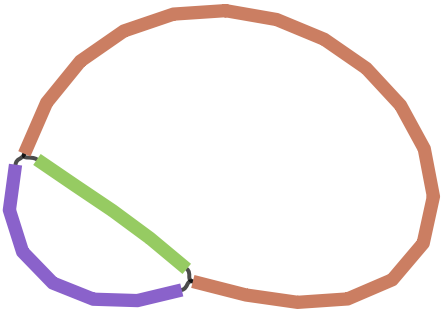
(a) *Buchnera of sample ArPo28 assembled in a single circular contig (642,215 bp)*



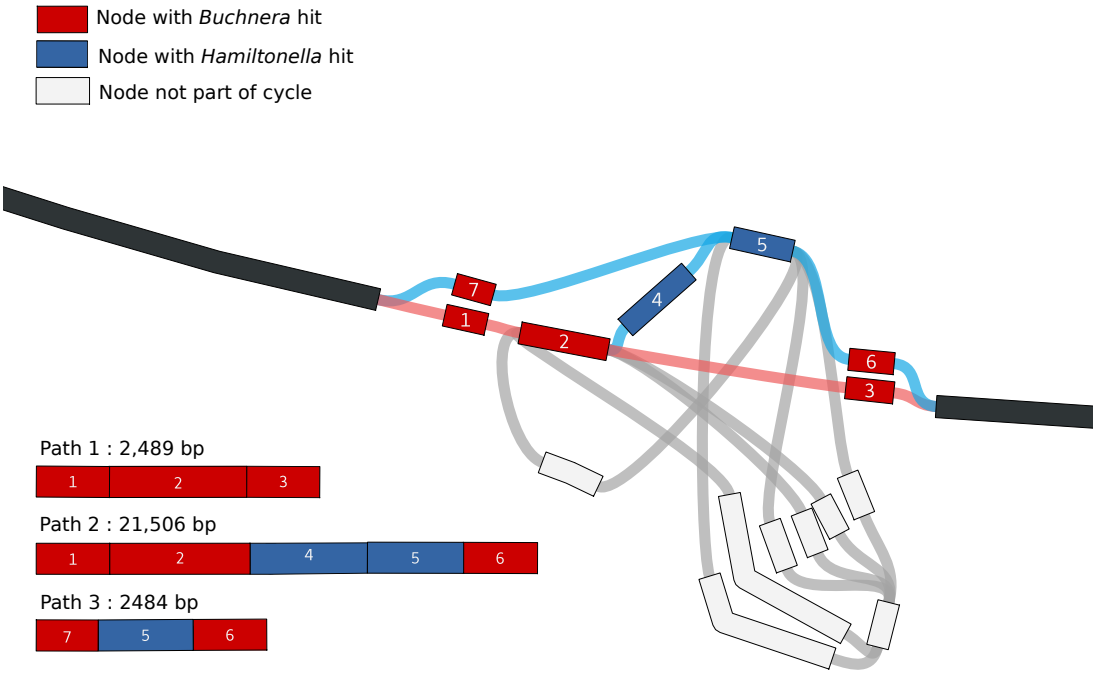
(b) *Buchnera of sample L3Lc03 incompletely assembled in a linear contig (631,066 bp)*



(c) *Buchnera of pooled sample Lc assembled with variable regions*



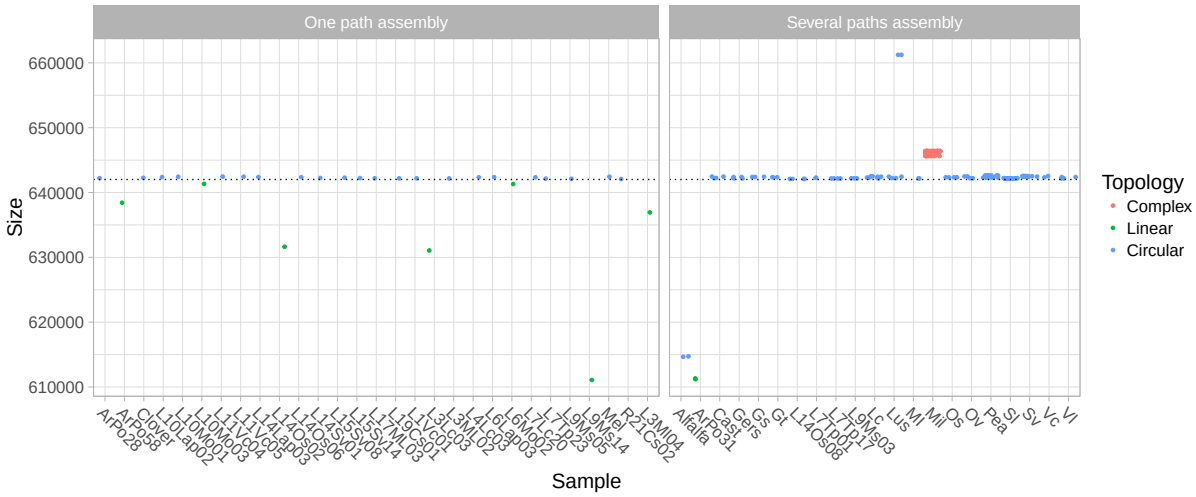
(d) *APSE of pooled sample OV assembled with two coexisting phage variants*



(e) *Closeup of a Hamiltonella contaminated region of the graph of Lus Buchnera assembly*

Figure 7 – ***Examples of genome graphs generated by MindTheGap***

Figure 8 – ***Length of the paths generated from MindTheGap genome graphs.*** *The horizontal line indicates the Buchnera aphidicola LSR1 genome length*



Altogether, these results indicate that the targeted *Buchnera* genome was most of the time assembled in the largest component. This component is expected to be a good representation of the *Buchnera* genome and its intra-sample polymorphism.

Assembly graph can be converted to whole-genome assembly with more or less ease

The genome graph includes possible structural variations or variant-rich regions, and therefore cannot be considered as a conventional genome assembly. For all samples, the component(s) associated to *Buchnera* were analyzed in order to enumerate all possible paths traversing them.

More than half of the samples are assembled in a single linear path In a first case, the graph was traversed by a single linear path. This concerned 28 samples out of the 50 used. In that case, the conversion of the graph into a genome assembly is straightforward. The assembly size is shown in the left pane of Figure 8. It appears to be well correlated with the topology observed for the graph. Circular graphs are obtained for most of the samples (see Figure 7a), and show a homogeneous assembly size comprised between 642,082 and 642,457 bp., which is in the range of other *Buchnera* genomes from *A. pisum* [Degnan et al., 2011]. The genome length of the reference genome of *Buchnera aphidicola LSR1* is 642,011 bp. On the other hand, seven samples were assembled into linear gapfilling graphs (see Figure 7b), and the resulting assembly were significantly smaller than other assemblies (between 95 and 100% of *Buchnera*). Linear graphs are associated with incomplete assembly, where all gaps within the assembly could not be filled. Accordingly, their size is smaller than the expected genome size. An extreme example is the L3ML04 sample which was assembled into two linear contigs. These samples are not associated to the presence of left-over contigs not used by the gapfilling step, indicating that the missing region was not assembled at all.

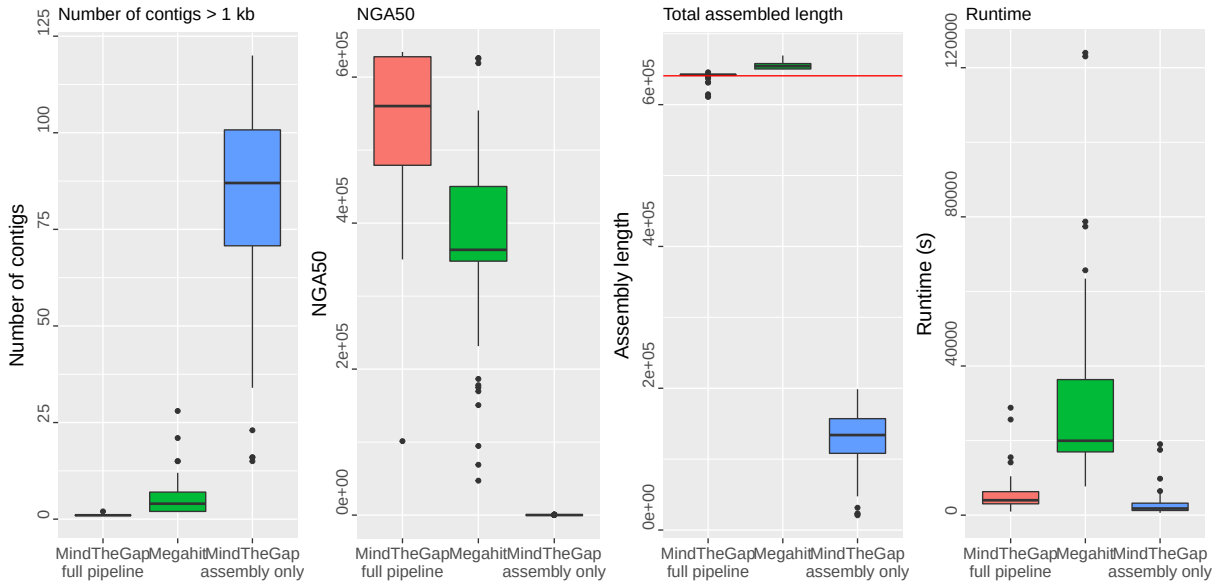
Graph traversal results in homogeneous assemblies accounting for intra-sample genomic variations Some samples could not be assembled in a single sequence, but in several alternative paths. The lengths of all these paths are represented in the right pane of Figure 8. A clustering based on a 99.9% ANI threshold was used to distinguish samples for which paths with different genomic structures may have arisen. These graphs were manually inspected to analyze the reasons for the sequence differences. First, intra-sample variability causes bubbles in the assembly graph, and therefore alternative paths. Several bubbles were detected among the graphs, containing either short polymorphism or insertions up to 2 kilobases, that can cause a significant ANI difference. This pattern is illustrated by Figure 7c. Second, a particular region of the graph may be disrupted by shared sequences originating from other organisms. This was caused by the assembly of non-*Buchnera* contigs during the assembly phase and mapping with the 16S/23S rRNA regions of both *Buchnera* and other related symbiont genomes. This frequently resulted in larger than expected assemblies, incorporating other symbiont sequences in the 16S/23S region. An example of this situation is illustrated in Figure 7e, showing sample *Lus* assembled in several paths, including or not *Hamiltonella defensa* contigs. In most cases, non-*Buchnera* contigs were present in the graph as dead-ends. Since the path traversal algorithm prioritizes circular paths, they were not taken into account, but their presence in the graph remains worrying. For the *Mil* sample however, no circular path was found in the graph, and this assumption could not help to generate a smaller number of paths. The highly fragmented regions constituted of mixed origins contigs was traversed by the algorithm, resulting in a high number of paths with a larger than average length.

Comparison with metagenomic assembly

Assembly statistics were computed for all samples for the *MindTheGap* pipeline, a MegaHit metagenomic assembly followed by contig filtering and the assembly step of the pipeline prior to the gapfilling.

MindTheGap outperforms the metagenomic assembly in most cases. Reference-guided assembly enables a one-contig assembly in all cases but one, while *MegaHit* outputs a single contig for only 70% of samples. Accordingly, the NGA50 is significantly higher for *MindTheGap*. This indicates both a lower number of contigs, and a lower number of missassemblies when aligning contigs on the reference genome. The average assembly size for *Megahit* exceeds the expected genome length. An explanation for this could be that highly polymorphic regions may be assembled into distinct contigs by the metagenomic assembler, while *MindTheGap* merges them, or represents them as bubbles in the genome graph. *MindTheGap* is also significantly faster than Megahit. The average runtime of *MindTheGap* is 95 minutes, which is 5.5 times inferior to *MegaHit* runtime (525 minutes). This has to be mitigated by the fact that *MegaHit* also produces contigs for not target organisms (such as *A. pisum* or secondary symbionts in that case).

Figure 9 – ***Comparison of assembly statistics for MindTheGap pipeline, Megahit assembly and assembly step of the pipeline*** *Horizontal red line indicates Buchnera aphidicola LSR1 genome length*



Gapfilling significantly improves the assembly of mapped reads

The assembly metrics of the full *MindTheGap* pipeline were also compared to those of the first step (Assembly of reads recruited by mapping). After removing contigs smaller than 1 kb, 127 kb were assembled in 81 contigs on average using the sole assembly step. This represents 19% of the size assembled using the full pipeline. The first step represents 52% of the total runtime of the pipeline. This highlights the efficiency of the reference-guided approach, which significantly improves the assembly size and contiguity in a moderate time.

4.4.2 Assembly of structural variants of pea aphid bacteriophages

To illustrate the ability of *MindTheGap* to detect and assemble structural variants of a genome, it was applied to the very appropriate genome of the bacteriophage APSE. As shown in figure 6, a region of its genome is a cassette of virulence genes, that may be substituted between APSE variants. Reference-guided assembly was used to assemble APSE genomes from 18 pea aphid datasets in which APSE was present, using APSE-1 as a guide genome.

Most samples were assembled into complete APSE genomes

9 samples were assembled into a unique circular contig. For 7 other samples, the graph showed two alternative paths, matching with the virulence cassette known for *APSE* (see Figure 7d for an example). These samples were assembled into two different phage genomes. For the *Gers* sample, 3 alternative sequences were detected. The inspection of the sequences revealed that one of them is mostly a repeated version of one of the other variants, matching either with a tandem repeat, or a repeat in the sequence confusing the assembly. Due to the low coverage of this sample (29X), this situation could not be resolved using more stringent

parameters. Finally, for the *Cast* sample, 4 sites of structural variation with between two and three variants were observed. In that case, several strains have been assembled in the genome graph, but path enumeration may result in chimeric strain genomes. This level of structural variation appeared to be intractable with our current approach, and requires additional extrinsic information.

Identification of phage variants from virulence cassette annotation

Overall, 25 circular APSE genomes were assembled (9 samples with unique variant, and 8 with two variants), with a length ranging between 36 and 41 kb, which is in the range of previously described APSE genomes (36-39 kb) [Degnan and Moran, 2008]. In order to characterize those variants, a first annotation of these genomes has been attempted using the already known gene repertoire of APSE. Correspondingly with the previous knowledge of this model, the main source of genomic variation is the virulence cassette located between the genes P5 and P14 [Degnan and Moran, 2008]. Based on the cassette genomic content, each phage genome was assigned to a phage variant, as summarized in table 4. We also included for 5 samples the protection phenotype observed against the parasitoid *Aphidius ervi*. 11 genomes were attributed to the APSE3 variant, 5 to other already known APSE variants (APSE1, APSE4 and APSE7), and 7 to a novel variant. A result of the gene alignment for each of the phage variants is showed in Figure 10. APSE-1 and APSE-4 variants share close virulence cassettes but differ by the insertion of the gene G in the region between genes P41 and P45. APSE3 carries a distinctive cassette, with the presence of genes Y, Z, E and F, and was found in most of the samples. The G protein gene was also found is some of the APSE-3 genomes, and this situation has not been described before. In previous work, variants APSE-6 and APSE-7 were isolated from other insects than the pea aphid. We confirmed that APSE-6 is not found, but the Clover sample shows some sequence similarity with most genes of the APSE-7 cassette, indicating that this variant may exist in the pea aphid.

Interestingly, we found a group of 8 samples with a novel variant. These genomes have a similar gene content that has not been described before, and contain a 5kb region not matching with any known APSE gene.

The guide tree obtained from the Mauve alignment (Figure 11) reflects this diversity, and represents the distances between the different phage variants. Samples with the same variants are clustered together. The relative positions of phage variants indicate their similarity. A first cluster in the tree is made of APSE-3 genomes, a second contains a APSE-7 variant and the novel variant, and a last one consists in APSE-1 and APSE-4 genomes.

Comparison with other bioinformatic tools

MindTheGap was compared to two other tools that can be used to assemble structural variants from a known genome : Pilon and MITObim. Pilon is designed to detect discrepancies between a reference genome and a read set, and is able to locally assemble structural variants. MITObim can be used to assemble novel mitochondrial genomes. It uses a remote reference genome, identifies conserved regions by mapping reads, and extend them by iteratively

Figure 10 – ***Representation of APSE genes alignment on MindTheGap assembled genomes.*** *One representative of each variant is shown. Genes homologous between several variants may show multiple hits.*

Figure 11 – *Clustering of the assembled APSE genomes.* *Built from Mauve multiple alignment guide tree*

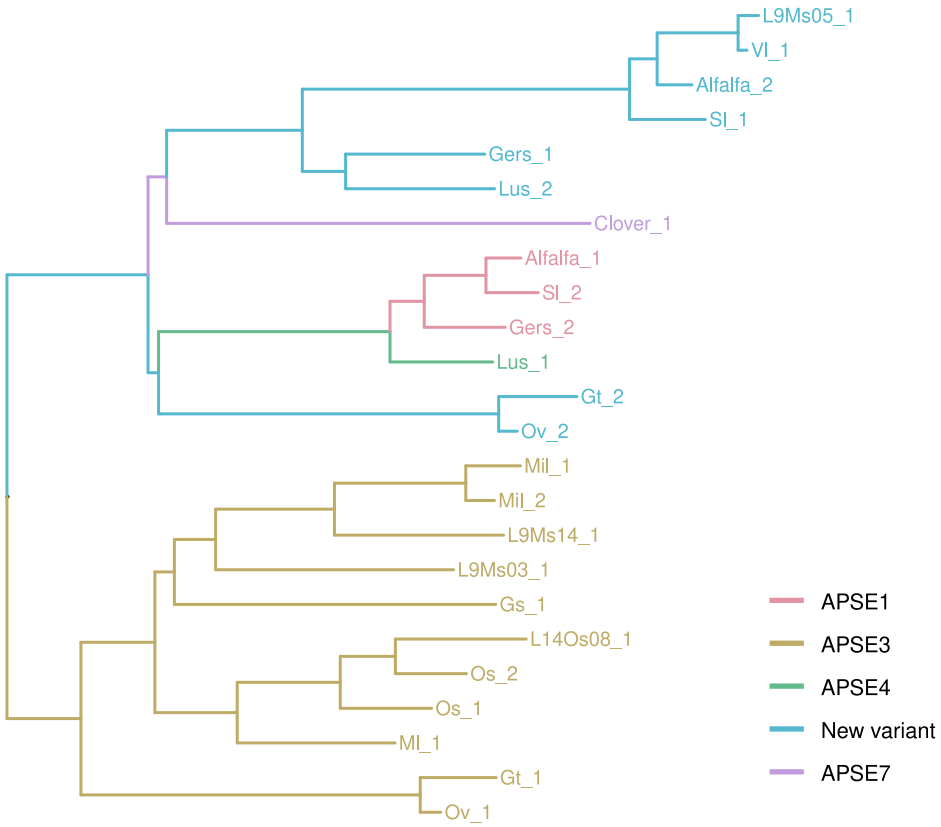


Table 4 – ***Assignation of assembled genomes to APSE variants.*** *The suffix "_2" indicates a second variant detected within the sample. Annotation was performed based on blast hits, and the cassette contents presented in [Degnan and Moran, 2008].*

Clone or Pool	Virulence cassette	Phage variant	Comments	Protective phenotype
Alfalfa_1	P7, P8, P9, P10, P11, P12, P13	APSE1		Uncharacterized
Sl_2	P7, P8, P9, P10, P11, P12, P13	APSE1		Uncharacterized
Gers_2	P7, P8, P9, P10, P11, P12, P13	APSE1	G (partial hit)	Uncharacterized
L9Ms03	Y,Z,E,F	APSE3	G (partial hit)	Uncharacterized
L9Ms14	Y,Z,E,F	APSE3		Uncharacterized
L14Os08	Y,Z,E,F	APSE3	G (partial hit)	Low protection
Gs	Y,Z,E,F + G	APSE3	Gene G	No protection
Gt_1	Y,Z,E,F + G	APSE3	Gene G	Uncharacterized
Os_1	Y,Z,E,F + G	APSE3	Gene G	Low protection
Os_2	Y,Z,E,F	APSE3	G (partial hit)	Low protection
Ov_1	Y,Z,E,F	APSE3		Uncharacterized
Mil_1	Y,Z,E,F	APSE3		Uncharacterized
Mil_2	Y,Z,E,F	APSE3	∼2kb insertion near gene p46	Uncharacterized
Ml	Y,Z,E,F + G	APSE3	Gene G	Uncharacterized
Mel	Y,Z,E,F + G	APSE3	Gene G	Uncharacterized
Lus_1	P7,P8,P9,P10, P11,P12,p13 + G	APSE4		Uncharacterized
Clover	cdtB, X, E, P11,P12,P13	APSE7		Uncharacterized
Alfalfa_2	e, P11, P12,P13	New variant		Uncharacterized
Vl	e, P11, P12,P13	New variant		Uncharacterized
L9Ms05	E,P11,P12,P13	New variant		Uncharacterized
Gt_2	P11,P12,P13	New variant		High protection
Sl_1	e, P11, P12,P13	New variant		Uncharacterized
Lus_2	e, P11, P12,P13	New variant		Uncharacterized
Ov_2	P11,P12,P13	New variant		Uncharacterized
Gers_1	D,E,P11,P12,P13	New variant		Uncharacterized

selecting overlapping reads. The *Gt* sample was selected to evaluate the performance of these alternative tools, because it is highly covered and carries two coexisting APSE variants.

While the whole *MindTheGap* pipeline took one hour, MITObim reached the maximum number of iterations allowed by the parameters in 6h33'. The initial contig was extended from 36,683 to 37,460 bp, while the length of the two variants assembled by *MindTheGap* is 38,031 bp (novel variant) and 40,936 bp (APSE-3 variant). Blast alignment of the contig produced by MITObim indicates that the assembled cassette is a part of APSE-3 cassette. At the end of the process, only 18% of the Gt virulence cassette had been assembled. On each iteration, contigs were extended by less than a read length, leading to a slow and incomplete assembly. In addition, MITObim is not designed to return several genomes, and only the APSE3 related cassette was assembled, while *MindTheGap* yielded two alternative sequences.

When applied to the Gt sample and the APSE1 genome, Pilon correctly identifies a 14kb uncovered region on the reference genome, as long with 300 SNPs and 5 small insertions. However, Pilon was unable to correct this large structural variation, which matches with the virulence cassette.

4.5 Discussion

4.5.1 Guided assembly for metagenomic datasets

Starting from the observation than both reference-based assignation and *de novo* assembly are inadequate to study some aspects of the metagenomic diversity, we present in the present work an hybrid method under the term of reference-guided assembly. This method was designed to assemble the genome of a single species of interest and its structural variants from metagenomic datasets. For this task, it outperforms both reference-based approaches and *de novo* assemblers. Reference based assignation suffers from the low number of sequenced and assembled bacterial genome, and is unable to identify new genomes. In *de novo* approaches, the assembly is performed prior to contig binning or mapping. This can be described as an *Assembly first* approach. Here, we present a *Mapping first* approach, that lightens the computational burden of metagenomic assembly, at the cost of a single genome assembly. To our knowledge, it is the first reference-based assembly approach suitable for metagenomic data.

MindTheGap was successfully applied to two organisms of the pea aphid complex. We hope it could be applied both to secondary symbionts of the pea aphid, and to other microbial communities. Structural variations of pea aphid symbionts have hardly been studied, however recent studies show that *Hamiltonella*'s genomic structure show profound rearrangements between strains [Chevignon et al., 2018]. Similarly, the study of SNPs of *Regiella insecticola* revealed two different strains, which may also differ by their genomic structure.

Beyond the pea aphid complex, *MindTheGap* may also be applied to a wide range of assembly issues. The targeted assembly approach reduces the number of sequences to assemble, and thus simplifies the assembly problem. This approach may therefore be suitable

for large and complex communities such as the human microbiome. Here, *MindTheGap* was presented as a complete pipeline from reads to contigs, but the second step of the pipeline can be associated to any other assembler. In this manner, *MindTheGap* can be used as a finishing tool for previous assemblies. In a metagenomic context, the targeted assembly may be a way to increase the contiguity of assemblies by joining metagenomic contigs coming from the same species.

However, the method has to be improved to give satisfying results for more complex genomes and communities. First, a wide range of parameters are applicable to the pipeline, including mapping sensitivity, assembly and gapfilling kmer sizes and minimum abundances. During the *Buchnera* genome reconstruction, inclusion in the assembly of contigs from another symbiont was observed in some cases. This indicates the necessity of either a more careful mapping, or a contig filtering based on reference homology or coverage. The importance of the mapping stringency has not been explored yet, but we believe that it is an important parameter, that should depend on both the distance between the target genome and the guide genome, and the complexity of the community. Assembly parameters have to be chosen carefully, in order to output reliable contigs with a minimum size of the target genome covered. Gapfilling may be less subject to the choice of parameters, but different parameters may result in different assemblies, especially when rare structural variants or repetitions are involved. Overall, despite this high number of parameters, *MindTheGap* was run with success on *Buchnera* with only two parameter sets, but some samples assembled in short contigs may benefit from some finer tuning. Better understanding of these parameters, and some automatic inference of the best values would also benefit to researchers applying *MindTheGap* to other communities.

The graph analysis may also be subject to enhancements. The *adhoc* procedures developed to output circular genomes of *Buchnera* or virulence cassettes of *APSE* may not be suitable for every assembly. Genome graphes are a promising tool for assembly troubleshooting and representation of structural variations, but converting them to regular "linear" assemblies without human guidance is challenging. To scale up to larger problems, both the path enumeration and the hypothesis of circular target genomes have to be replaced by a more effective approach. In addition, some structural variations may be impossible to solve using only the genome graph. Read depth may be used as a tool for both graph traversal and phasing of structural variants. However, it is subject to important variations depending on many other factors (GC content, homology with other organisms...) and may be delicate to use. Alternatively, using information with a longer range than a kmer size, for instance by mapping paired-end reads on graph paths, may help to produce an assembly, but would require a costly extra-step of assembly.

4.5.2 Structural variations in symbiont genomes of the pea aphid complex

However,

By applying *MindTheGap* to pea aphid metagenomic samples, we obtained first results on the structural diversity of some organisms of its microbiota. Previous works showed that *B.*

aphidicola genomes could vary between and within aphid species, with potential functional consequences for the host [Chong and Moran, 2018, Vogel and Moran, 2011]. Accordingly, we assembled genomes with variations in size and sequence across the samples analyzed here. The genomic structure of *Buchnera* is very conserved, and, as expected, we did not find any structural rearrangements or large genomic variations. We did find some variations in the genome size, associated with small insertions or deletions. The nature of these variations have to be further investigated, but they may be linked to the pseudogeneisation that the genome of *Buchnera* is undergoing [Gil et al., 2002].

It is known that the bacteriophage APSE is involved in the protection of the pea aphid against parasitoids [Degnan and Moran, 2008], but also that different APSE variants have different effects on the protection phenotype. Here, we generated complete APSE genomes for 17 samples. The gene content of the assembled genomes was determined by comparing the sequences with those of known APSE genes. This approach was able to assign APSE variants to samples, and highlight a novel variant. However, a better approach, with *ab initio* annotation of the virulence cassette and function prediction, may be more appropriate to investigate the functionnal role of the novel variant, and make hypotheses on its functional role.

Moreover, genomic analyses may not be enough to investigate the role of APSE variants for the pea aphid. The protection phenotype of only a few samples analyzed in the present study has been determined. A wider study, encompassing phenotypic characterization and -omics data would be highly valuable to precise the role of APSE.

4.6 Conclusion

In the present work, we present an hybrid strategy between *de novo* assembly and reference-based techniques for the characterization of microbial genomes in metagenomic datasets. Compared to metagenomic assembly, this approach is very efficient to assemble a single genome. It yields a genome graph representation, which is valuable to represent the structural diversity of genomes, rarely examined in metagenomic datasets.

We applied this method based on *MindTheGap* on 50 metagenomic samples of the pea aphid. As an targeted assembler, *MindTheGap* shows excellent results when applied to the primary symbiont *Buchnera aphidicola*. In addition, we were able to assemble structural variants of the bacteriophage APSE, including a novel variant that is yet to be characterized. *MindTheGap* may also be applied to more symbionts of the pea aphid to give more insights on the genomic structural diversity of its microbiota. More detailed analyses, including functional and metabolic approaches, are necessary to better understand the interactions between the pea aphid and its symbionts.

Finally, the method detailed in the present study is not restricted to the pea aphid holobiont, and may be useful in any metagenomic assembly problem. However, the method has not been tested yet with more complex genomes or microbial communities, and further development may be required to be applicable to any metagenomic problem.

Chapitre 5

Discussion et perspectives

5.1 Rappel des objectifs de la thèse

Ces travaux de thèse ont porté sur l'étude par des données métagénomiques du microbiote du puceron du pois. Bien qu'étant un modèle d'étude pour les relations hôte-symbiotes chez les insectes, les communautés microbiennes associées à ce complexe restaient mal décrites au niveau génomique. L'objectif de cette thèse était de tirer profit des avantages offerts par la métagénomique plein-génome en exploitant principalement un grand jeu de données de séquençage métagénomique de pucerons du pois. L'holobionte du puceron du pois a ainsi été caractérisé sur les trois niveaux suivants :

1. Un premier objectif était l'inventaire des organismes impliqués dans cette communauté symbiotique. Bien que les principaux symbiotes du puceron du pois aient déjà été largement identifiés, il restait possible d'identifier de nouveaux partenaires plus discrets, tels que des virus ou plus transitoires, comme certaines bactéries intestinales.

2. Les données métagénomiques offrent la possibilité de dépasser ce simple inventaire d'espèces microbiennes, en offrant une haute résolution (par le séquençage de variants génomiques) et ceci sur tout le génome. Le deuxième objectif était donc d'utiliser cette information pour mieux caractériser et comprendre la diversité génomique de cette communauté.

3. Enfin, un troisième objectif était de comprendre comment cette diversité s'organise au niveau populationnel, notamment entre les différents biotypes de puceron, et éventuellement d'appréhender les mécanismes évolutifs ayant façonné cette diversité.

La métagénomique est une discipline relativement récente, qui a longtemps été réservée à des communautés modèles telles que le microbiote humain. Ainsi, répondre à ces questions a nécessité à la fois l'inventaire des méthodologies existantes, leur adaptation au contexte d'un holobionte, et le développement de méthodes dédiées.

C'est sur ces deux volets, que sont la caractérisation métagénomique du microbiote du puceron du pois et le développement de méthodes nécessaires à cette tâche, que s'est orientée cette thèse.

5.2 Quelles méthodologies adaptées à la problématique de thèse ?

Une première étape de ces travaux de thèse a été un travail de revue bibliographique, consistant à identifier les outils bioinformatiques existants, et comment ils peuvent répondre aux objectifs fixés.

Les communautés les plus fréquemment étudiées en métagénomique sont des communautés environnementales (sols, océans), et le microbiome humain. La complexité des communautés rend difficile l'étude des génomes individuels des microbes présents en l'état actuel des technologies de séquençage. S'il est aujourd'hui possible de reconstituer partiellement des génomes à partir de ces communautés (les Metagenome Assembled Genomes ou MAG), l'assemblage complet et systématique de génomes bactériens à partir de données métagénomiques reste un défi impossible à relever pour le moment. Ainsi, les travaux et les méthodologies autour de ces communautés complexes se concentrent le plus souvent sur leur caractérisation taxonomique à un niveau grossier (genre ou espèce), et sur l'identification de certains gènes importants d'un point de vue fonctionnel. Un autre objet d'étude fréquent de la métagénomique est le microbiote humain, qui est quant à lui très bien décrit. Dans ce contexte, des méthodes basées sur des bases de données de référence permettent de bien décrire la diversité de ces systèmes.

La problématique de cette thèse est la description fine de la communauté bactérienne associée au puceron du pois, qui ne dispose pas de ressources génomiques aussi abondantes que d'autres communautés plus étudiées. Ainsi, les méthodes *de novo* sont incapables de répondre aux questions biologiques que nous posons sur notre modèle, tandis que les méthodes reposant fortement sur des génomes de référence n'y sont pas applicables. Une solution est l'approche proposée par des outils comme StrainPhlan [Truong et al., 2017] ou ConStrains [Luo et al., 2015], qui identifient des souches bactériennes à partir de variants détectés par alignement sur un génome de référence. Ils sont cependant restreints à une liste de gènes limitée.

L'état de l'art décrit dans le deuxième chapitre a ainsi permis d'identifier à la fois l'intérêt et les limites de recourir à des génomes de référence pour caractériser la diversité métagénomique. Les travaux présentés ensuite vont dans le même sens, avec d'abord la caractérisation de souches par alignement sur des références, puis une méthode permettant l'assemblage de génomes à partir d'une référence inappropriée et la détection de variations structurales.

5.3 Comment étudier la diversité génomique d'un holobionte ?

La majorité des holobiontes se distingue des communautés les plus fréquemment étudiées en métagénomique. Il s'agit d'une communauté dominée par quelques espèces bactériennes déjà connues, pour lesquelles des génomes de référence sont disponibles de manière incomplète. En revanche, les variations génomiques intra-spécifiques de ces symbiotes n'ont jamais été examinées. Cela constitue donc un cas finalement peu rencontré dans les études

métagénomiques, pour lequel peu de méthodes ont été développées. Le puceron du pois est un bon exemple de cette situation, puisqu'il est associé à un faible nombre de symbiotes.

À travers les contributions de cette thèse, et l'expérience acquise sur l'holobionte du puceron du pois, nous proposons une démarche permettant d'étudier les variations intra-spécifiques du microbiote d'un holobionte, représentée par la figure 12 :

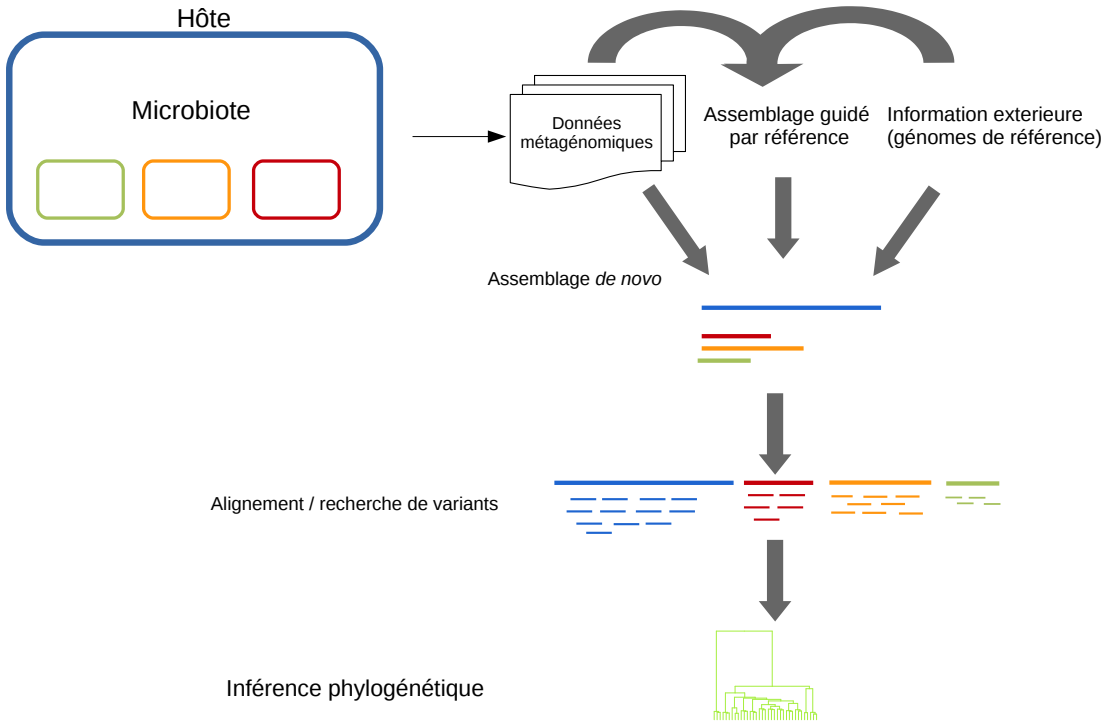
1. Constitution d'un ensemble de génomes de référence, incluant le(s) génome(s) de l'hôte, à l'aide de connaissances *a priori* ou par assemblage. Dans ce cas, la méthode d'assemblage guidé développée autour de *MindTheGap* peut être utilisée si un génome proche de celui à assembler est disponible.
2. Les génomes de référence ainsi collectés peuvent être utilisés pour l'assignation des lectures à un organisme de l'holobionte, par l'alignement des lectures métagénomiques sur ces génomes de référence. Cela permet de constituer des jeux de lectures spécifiques à chaque symbiote. La proportion de lectures s'alignant peut être utilisée comme un indicateur de la qualité de l'alignement et de la complétude du jeu de génomes de référence.
3. L'étude des variations intra-spécifiques devient alors possible en appliquant des méthodes classiques de détection de variants.
4. Enfin, les différents génotypes détectés peuvent être comparés par des approches phylogénétiques.

Cette démarche s'est révélée très satisfaisante chez le puceron du pois, en fournissant des résultats importants sur la diversité génomique et l'histoire évolutive des associations hôte-microbiote au sein du complexe. Elle pourrait vraisemblablement être appliquée à d'autres holobiontes. Le puceron du pois est un modèle particulièrement bien décrit, et la constitution d'un ensemble de génomes de référence n'a pas été trop difficile. Pour des holobiontes moins bien décrits, la détection et l'assemblage des symbiotes seraient une étape plus difficile. L'usage de *MindTheGap* serait alors tout à fait pertinent pour fournir des génomes de qualité suffisante à l'alignement de la majorité des lectures sans erreur d'assignation. Outre cela, certains problèmes méthodologiques, énumérés ci-dessous, pourraient être rencontrés dans d'autres systèmes, et sont liés notamment au recours à des génomes de référence et la difficulté de la caractérisation de la diversité intra-échantillon.

Recours à un génome de référence

L'étape clé de cette démarche est l'alignement sur des génomes de référence, permettant de passer d'un cadre métagénomique à un jeu de lectures provenant d'un même organisme. Cette étape s'accommode mal du *continuum* de diversité qui peut exister dans certaines communautés. Des microbes distincts peuvent partager des régions génomiques proches, par exemple suite à un transfert horizontal de gène(s). Dans ce cas, des lectures d'un autre symbiote peuvent être alignées sur un génome de référence par erreur, en particulier si le symbiote "contaminant" est absent du jeu de génomes de référence. Ces erreurs se traduiraient

Figure 12 – *Démarche permettant l'analyse métagénomique d'un holobionte*
La démarche proposée ici repose sur l'obtention de génomes de référence. Ils peuvent déjà exister dans les bases de données de référence, être assemblés de novo à partir de données métagénomique, ou bien être assemblés à partir d'un génome existant mais distant, tel que proposé par notre méthode d'assemblage guidé.
L'alignement des lectures sur ces génomes permet ensuite de s'abstraire du cadre métagénomique, et de réaliser des analyses classiques, telles que la détection de variants et l'inférence phylogénétique.



alors par des variants artefactuels qui perturberaient les résultats de l'analyse phylogénétique. La solution retenue pour le puceron du pois a été d'exclure de l'analyse de variants toutes les régions homologues entre plusieurs génomes de référence, ce qui serait plus difficilement applicable dans des communautés plus grandes et non intégralement assemblées.

À l'inverse, pour un même génome de référence de symbiote, des génotypes très différents peuvent être rencontrés. Notre étude met en évidence la coexistence de deux génotypes de *Regiella*, qui ont été traités simultanément. Ils auraient également pu être traités de manière disjointe, en construisant un jeu de lecture pour chaque souche de *Regiella*. En fonction du degré de divergence entre les deux souches, une telle analyse aurait pu mettre en lumière des aspects différents le la diversité de ce symbiote.

Enfin, si notre méthode d'assemblage guidé permet d'observer les variations structurales du génome d'un symbiote, le reste du processus est mal adaptée à cette diversité. Nos travaux permettant l'inférence de scénarios évolutifs des associations hôte-microbiote reposent exclusivement sur l'alignement avec un génome de référence, et s'accommoderaient mal d'un éventuel polymorphisme structural. Le bactériophage APSE, de par sa cassette de virulence variable, est par exemple peu adapté à notre démarche basée sur l'alignement sur un génome de référence.

Difficultés de la caractérisation de la diversité intra-échantillon

Un des résultats apportés par ces travaux est l'étude du polymorphisme intra-individuel. En particulier, nos résultats indiquent la coexistence de deux souches distinctes du symbiote *Regiella*. Ce résultat intéressant, permis par les données métagénomiques, est cependant dû à des conditions très favorables. Ces souches sont présentes seulement au nombre de deux, se distinguent par un très grand nombre de variants, sont toutes deux présentes avec une couverture forte, et des abondances relatives éloignées du ratio 50%-50%. À l'exception de ce cas particulier, nous avons traité les génotypes de chaque échantillon comme un consensus, en sélectionnant les allèles les plus abondants de chaque variant. Ainsi, une part non négligeable du polymorphisme n'a pas été considérée. En particulier, les séquençage de *pools* abritent une importante diversité génomique, qui est difficile à analyser à l'échelle des génotypes individuels qui la composent.

Une opportunité pour améliorer la finesse des génotypes identifiés serait de considérer ces variations intra-échantillons. Pour ce faire, une solution serait d'employer des méthodes statistiques comparant ces variants et leurs abondances à travers différents échantillons. La situation décrite pour *Regiella*, à savoir deux souches fortement distinctes, ne semble pas être un cas général, mais il serait possible d'identifier des allèles ou groupes d'allèles partageant la même répartition au sein de la population.

5.4 Alignement guidé par référence

L'assemblage *de novo* de génomes à partir de données métagénomiques est un problème difficile et loin d'être complétement résolu. Au cours de ces travaux de thèse, nous avons

proposé une méthode qui facilite la reconstruction de génomes bactériens à partir de données métagénomiques, en utilisant un génome de référence distant comme une amorce à l'assemblage. Par rapport aux méthodes d'assemblage métagénomique traditionnelles, cet apport d'information extérieure améliore à la fois la qualité de l'assemblage et diminue les ressources informatiques et le temps de calcul nécessaires. Dans le cas de l'assemblage du génome de *Buchnera aphidicola*, les performances sont supérieures à celles obtenues avec un assembleur métagénomique, notamment dans les échantillons de *pools* montrant davantage de polymorphisme.

Outre les performances d'assemblage, cette méthode est conçue pour prendre en compte les variations structurales des génomes assemblés. Le résultat de l'assemblage n'est pas une séquence linéaire mais un graphe permettant d'identifier d'éventuels variants structuraux. Cette propriété est mise en relief par l'application de la méthode à l'assemblage du bactériophage APSE, pour lequel plusieurs variants d'une cassette de virulence ont été identifiés. À notre connaissance, aucune autre méthode existante n'est aussi adaptée à ce cas de figure, où différentes insertions de plusieurs kilobases peuvent coexister. Cette représentation des génomes sous forme de graphe est particulièrement pertinente en métagénomique, où un fort polymorphisme structural est observable. Ces résultats invitent à poursuivre le développement de *MindTheGap* pour étendre ses applications et améliorer ses performances.

Incorporation du polymorphisme ponctuel

MindTheGap était destiné dans un premier temps à la détection d'insertions structurales. Nous en présentons ici une version étendue, intégrée dans une méthode d'assemblage guidé. Dans ce contexte, *MindTheGap* est utilisé pour réaliser un assemblage local entre des contigs, ce qui permet de détecter soit des arrangements différents entre des contigs, soit des insertions. La diversité détectée de cette manière n'est cependant pas exhaustive. Premièrement, les contigs sont utilisés tels quels dans le graphe d'assemblage produit par la méthode. Ni le polymorphisme ponctuel, ni les éventuelles insertions ou déletions au sein de ces contigs ne sont représentés. En dehors des contigs, seules les séquences de *gapfilling* suffisamment distinctes sont retournées, et le polymorphisme ponctuel est donc également ignoré. Seule une accumulation importante de ces variants peut se traduire par deux séquences suffisamment différentes qui seront alors retournées

Une solution plus globale pour intégrer le polymorphisme ponctuel à notre approche pourrait être de réaligner les lectures sur l'assemblage final. Ces variants pourraient être intégrés au graphe d'assemblage, qui représenterait alors l'intégralité de la diversité observée sur ce génome, mais serait aussi plus difficile à analyser.

Résolution du graphe

Un des atouts de *MindTheGap* est de pouvoir aboutir à un génome non linéaire, où les variations structurales sont représentées. Un tel graphe est pourtant encore difficile à analyser pour le moment, et le développement de méthodes plus perfectionnées pour le parcourir et

en extraire des génomes complets est nécessaire.

Actuellement, l'outil ne permet pas de phaser deux variants structuraux situés à des endroits différents du génome. Dans certains cas d'assemblage, il est possible d'utiliser la différence de couverture entre différents variants pour identifier ceux qui proviennent d'un même génotype. Néanmoins, cette approche n'est pas implémentée dans *MindTheGap*. Il est à noter qu'en dehors des cas les plus simples, la couverture est une information difficile à utiliser du fait de sa variabilité. En complément, utiliser l'information de lectures pairées peut aussi permettre de résoudre certains problèmes du graphe, comme des répétitions ou des variants situés dans des régions proches du génome. Cela nécessiterait également de réaligner les lectures sur le graphe.

Autres applications de *MindTheGap*

Nous avons employé *MindTheGap* pour assembler un génome d'intérêt à partir de données métagénomiques du puceron du pois, mais nous pouvons entrevoir d'autres utilisations de cet outil.

La méthode peut tout d'abord être employée comme un outil d'assemblage métagénomique dans d'autres systèmes que le puceron du pois. Les premiers travaux sur le microbiome digestif humain montrent des résultats encourageants. Les résultats doivent cependant être approfondis et comparés à d'autres méthodes d'assemblage métagénomique.

MindTheGap peut aussi être employé pour améliorer un assemblage existant. Il est alors utilisé comme un logiciel de *gapfilling* permettant d'assembler les régions manquantes et de réduire la contiguité d'un assemblage. Nous l'avons utilisé à cet effet, nottamment pour améliorer l'assemblage de *Rickettsia sp.*.

Nous avons ici employé *MindTheGap* sur un échantillon à la fois, en produisant les contigs et les insertions à partir du même jeu de lectures. Dans la littérature, il est fréquent de tirer profit de l'existence de plusieurs jeux de données au contenu différent pour reconstituer des souches microbiennes [Luo et al., 2015, Cleary et al., 2015]. D'une manière un peu différente, l'information de plusieurs jeux de données métagénomiques pourrait être utilisée pour construire un génome de base (*core genome*) pour un microbe de la communauté. Dans ce cas, les contigs initiaux peuvent être construits à partir de sous-ensembles de lectures voire de kmers, sélectionnés par intersection de différents jeux de données. Ainsi, dans le cadre de projets de séquençage d'un holobionte, le génome de l'hôte ou de symbiotes obligatoires peut être en partie reconstruit en sélectionnant les kmers présents dans un grand nombre de ces jeux de données. À la condition qu'une information préalable sur la présence et l'absence des symbiotes facultatifs soit disponible, il est également possible de sélectionner ces génomes, en choisissant les kmers présents dans les jeux de données infectés, mais pas dans d'autres. En pratique, une telle sélection pourrait s'opérer en suivant les principes des algorithmes de métagénomique comparative *de novo*, tels que Commet [Maillet et al., 2014] au niveau des lectures ou Simka [Benoit et al., 2016] au niveau des kmers. L'assemblage construit à partir de ces séquences représenterait les régions conservées de cet organisme à travers tout le jeux de données. L'emploi de *MindTheGap* à partir de ces contigs et de chaque échantillon

pris séparément permettrait de reconstruire le reste du génome. La comparaison des graphes d'assemblage entre différents échantillons correspondant à différentes conditions en ferait un outil très puissant de génomique comparative entre différent métagénomes. Cette approche permettrait en outre de se passer du génome de référence nécessaire à l'assemblage guidé. Bien sûr de nombreuses questions restent à résoudre quant à cette application, notamment sur les conditions de la sélection des lectures communes. Il s'agit néanmoins d'une perspective intéressante pour l'application de *MindTheGap*.

Une autre piste permettant d'employer *MindTheGap* est l'assemblage de données *10X Genomics*. Cette technologie commercialisée par la société *10X Genomics* repose sur la notion de lectures liées (*linked*). Les lectures provenant d'un même fragment d'ADN sont encapsulées ensemble, séquencées par la technologie *Illumina*, et un marqueur permettant de les identifier est ajouté aux séquences. Il est alors possible d'identifier les lectures provenant d'une même région génomique, ce qui facilite des tâches telles l'assemblage ou la reconstruction d'haplotypes. En pratique, les assemblages 10X sont parfois incomplets, et les régions que l'assembleur n'a pu reconstruire contiennent des répétitions de N dont la taille peut atteindre plusieurs dizaines de kilobases. Utiliser *MindTheGap* permettrait d'assembler ces régions manquantes et de reconstituer un meilleur assemblage.

5.5 Apports sur la diversité et l'évolution des associations hôte-microbiote au sein du complexe du puceron du pois

La démarche proposée pour l'analyse métagénomique d'holobiontes a été appliquée au puceron du pois. Elle a permis d'étudier à différentes échelles les multiples aspects de la diversité de cette communauté , ayant chacun un impact sur le fonctionnement et l'évolution de l'holobionte.

En construisant un ensemble de génomes de référence contre lequel la grande majorité des lectures s'alignent, nous avons réalisé un inventaire exhaustif du microbiote du puceron du pois. Nos résultats sont conformes avec les connaissances préalables sur le complexe, et indiquent que le puceron du pois est associé à un nombre réduit de symbiotes bactériens, dont un symbiote obligatoire. L'exploration de ces jeux de données métagénomiques n'a pas permis d'identifier de nouveaux membres de ce microbiote, que ce soit des bactéries présentes transitoirement ou des virus.

Nous avons ensuite décrit la communauté symbiotique du puceron du pois à une échelle intra-spécifique jusque là peu étudiée. L'ampleur et la forme prise par cette diversité intra-spécifique est très variable selon les symbiotes, notamment en fonction du mode de transmission. Cela nous a mené à proposer différents scénarios évolutifs. Nous avons confirmé la différence fondamentale de transmission entre le symbiote obligatoire *Buchnera aphidicola*, systématiquement hérité verticalement, et les symbiotes secondaires qui peuvent connaitre des événements de transfert horizontal. Parmi ces symbiotes secondaires, les différents scénarios décrits s'associent à des hypothèses à la fois sur les modalités de l'entrée dans le complexe et

la fréquence des transferts horizontaux.

À partir de ces résultats, plusieurs pistes permettraient d'étendre significativement nos connaissances sur la communauté symbiotique du puceron du pois.

Vers la métagénomique des populations

Les hypothèses faites sur les scénarios d'association symbiotique sont fragilisées par la taille réduite de l'échantillon. À titre de comparaison, [Oliver et al., 2010] utilisent plus d'un millier d'individus pour caractériser les dynamiques symbiotiques à l'aide de séquences de gènes de ménage. Bien que la résolution permise par la métagénomique plein-génome soit supérieure, l'usage de seulement 50 échantillons ne permet pas de complètement caractériser une population complexe constituée de 15 biotypes. Les séquençages de *pools* permettent en principe d'atténuer cette limitation, en séquençant un grand nombre d'individus à la fois. En pratique, leur analyse est difficile, et ne permet pas d'obtenir des résultats plus fins que la présence ou l'absence de symbiotes secondaires. Les résultats présentés sont encourageants, mais bénéficieraient donc d'étendre l'analyse à un plus grand nombre d'individus, pour décrire avec plus de confiance des dynamiques évolutives.

À l'aide de cet échantillon réduit, nous avons essentiellement tiré des conclusions qualitatives sur l'histoire évolutive des associations symbiotiques. Pourtant, la métagénomique plein-génome nous permet d'obtenir une information quantitative, exempte des biais d'amplification fréquents en *barcoding*. Sous réserve de certaines conditions expérimentales lors de la préparation des librairies, les couvertures de symbiotes peuvent être comparées entre elles pour donner une mesure de leur abondance. La fréquence des variants, qui a été considérée comme binaire dans nos travaux, pourrait être utilisée de manière plus précise, éventuellement en calculant des fréquences alléliques à l'échelle de biotypes. Enfin, la fréquence de certains éléments génomiques pourrait être comparée à celle du symbiote associé. En particulier, les plasmides de *Buchnera* et le bactériophage APSE associé à *Hamiltonella* sont fortement liés au métabolisme de l'hôte (fourniture d'acides aminées et résistance à des ennemis naturels). Le nombre de copies de ces éléments pourrait être un indicateur de leur importance pour l'holobionte dans une population donnée.

Vers la métagénomique fonctionnelle

La principale promesse de la métagénomique plein-génome est d'accéder au potentiel fonctionnel complet d'une communauté microbienne. En raison de la complexité de ces données et de la faible annotation des génomes bactériens, il s'agit en réalité d'un objectif difficile à atteindre. Nous avons caractérisé un grand nombre de variants génomiques symbiotiques, distribués de manière hétérogène dans les différents biotypes d'hôte. Un prolongement logique de ces travaux serait de tenter de mesurer l'impact fonctionnel de ces variants. Pour cela, il est d'abord nécessaire d'annoter ces variants, en distinguant ceux situés dans des régions codantes, et ceux ayant un effet sur la séquence protéique. En complément, les données métagénomiques peuvent être associées à d'autres types de données. Il peut s'agir de données métatranscriptomiques, de données métabolomiques, ou

plus simplement de mesures phénotypiques. Cette analyse intégrée permettrait à la fois de mieux décrire les interactions métabolomiques entre l'hôte et ses symbiotes, et de comprendre les mécanismes de sélection aboutissant à la diversité observée à présent. Un objectif pourrait être la reconstruction conjointe des réseaux métaboliques de l'hôte et de ses symbiotes, de manière similaire à ce qu a été fait chez *Bemisia tabacii* [Opatovsky et al., 2018]. Un tel travail a permis de mettre en évidence le rôle métabolique de symbiotes, ainsi que les interactions entre eux et avec l'hôte.

Outre l'aspect métabolique, certains autres aspects phénotypiques de l'association symbiotiques pourraient être éclaircis par cette caractérisation fonctionnelle.

Par rapport aux autres symbiotes, la structuration de la diversité observée pour *Regiella* est remarquable, avec deux souches très distinctes et séparées par plusieurs dizaines de milliers de variants. Les rôles de ce symbiote pour son hôte son multiples (résistance à des parasitoïdes ou pathogènes fongiques, et adaptation à la plante-hôte), et ces nombreux variants qui covarient au sein du complexe pourraient les expliquer.

Le rôle du bactériophage APSE dans la résistance au parasitoïde *Aphidius ervi* a été démontré [Oliver et al., 2009], mais cette protection peut varier en fonction des populations d'hôtes [Leclair et al., 2016], et le mécanisme sous-jacent est encore mal connu [Brandt et al., 2017]. À l'aide de *MindTheGap*, nous avons assemblé et identifié différents variants de ce phage, dont certains n'avaient jamais été décrits auparavant, et qui se différencient par leur cassette de virulence. Ce travail ouvre la voie à une meilleure description et compréhension des modalités de cette protection symbiotiques.

Table des figures

Bibliographie

[Ahn et al., 2015] Ahn, T. H., Chai, J., and Pan, C. (2015). Sigma : Strain-level inference of genomes from metagenomic analysis for biosurveillance. Bioinformatics, 31(2) :170–177.

[Albertsen et al., 2013] Albertsen, M., Hugenholtz, P., Skarshewski, A., Nielsen, K. L., Tyson, G. W., and Nielsen, P. H. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. Nature Biotechnology, 31(6) :533–538.

[Alneberg et al., 2014a] Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014a). Binning metagenomic contigs by coverage and composition. Nature Methods, 11(11) :1144–1146.

[Alneberg et al., 2014b] Alneberg, J., Bjarnason, B. S., De Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F., and Quince, C. (2014b). Binning metagenomic contigs by coverage and composition. Nature Methods, 11(11) :1144–1146.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, W. E., and Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215 :402–410.

[Angly et al., 2009] Angly, F. E., Willner, D., Prieto-Davó, A., Edwards, R. A., Schmieder, R., Vega-Thurber, R., Antonopoulos, D. A., Barott, K., Cottrell, M. T., Desnues, C., Dinsdale, E. A., Furlan, M., Haynes, M., Henn, M. R., Hu, Y., Kirchman, D. L., McDole, T., McPherson, J. D., Meyer, F., Miller, R. M., Mundt, E., Naviaux, R. K., Rodriguez-Mueller, B., Stevens, R., Wegley, L., Zhang, L., Zhu, B., and Rohwer, F. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Computational Biology, 5(12) :e1000593.

[Arumugam et al., 2011] Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J. M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., De Vos, W. M., Brunak, S., Doré, J., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. Nature, 473(7346) :174–180.

[Bankevich et al., 2012] Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes : A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. Journal of Computational Biology, 19(5) :455–477.

[Bateman et al., 2017] Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., Bely, B., Bingley, M., Bonilla, C., Britto, R., Bursteinas, B., Bye-AJee, H., Cowley,

A., Da Silva, A., De Giorgi, M., Dogan, T., Fazzini, F., Castro, L. G., Figueira, L., Garmiri, P., Georghiou, G., Gonzalez, D., Hatton-Ellis, E., Li, W., Liu, W., Lopez, R., Luo, J., Lussi, Y., MacDougall, A., Nightingale, A., Palka, B., Pichler, K., Poggioli, D., Pundir, S., Pureza, L., Qi, G., Rosanoff, S., Saidi, R., Sawford, T., Shypitsyna, A., Speretta, E., Turner, E., Tyagi, N., Volynkin, V., Wardell, T., Warner, K., Watkins, X., Zaru, R., Zellner, H., Xenarios, I., Bougueleret, L., Bridge, A., Poux, S., Redaschi, N., Aimo, L., ArgoudPuy, G., Auchincloss, A., Axelsen, K., Bansal, P., Baratin, D., Blatter, M. C., Boeckmann, B., Bolleman, J., Boutet, E., Breuza, L., Casal-Casas, C., De Castro, E., Coudert, E., Cuche, B., Doche, M., Dornevil, D., Duvaud, S., Estreicher, A., Famiglietti, L., Feuermann, M., Gasteiger, E., Gehant, S., Gerritsen, V., Gos, A., Gruaz-Gumowski, N., Hinz, U., Hulo, C., Jungo, F., Keller, G., Lara, V., Lemercier, P., Lieberherr, D., Lombardot, T., Martin, X., Masson, P., Morgat, A., Neto, T., Nouspikel, N., Paesano, S., Pedruzzi, I., Pilbout, S., Pozzato, M., Pruess, M., Rivoire, C., Roechert, B., Schneider, M., Sigrist, C., Sonesson, K., Staehli, S., Stutz, A., Sundaram, S., Tognolli, M., Verbregue, L., Veuthey, A. L., Wu, C. H., Arighi, C. N., Arminski, L., Chen, C., Chen, Y., Garavelli, J. S., Huang, H., Laiho, K., McGarvey, P., Natale, D. A., Ross, K., Vinayaka, C. R., Wang, Q., Wang, Y., Yeh, L. S., and Zhang, J. (2017). UniProt : The universal protein knowledgebase. Nucleic Acids Research, 45(D1) :D158–D169.

[Benoit et al., 2016] Benoit, G., Peterlongo, P., Mariadassou, M., Drezen, E., Schbath, S., Lavenier, D., and Lemaitre, C. (2016). Multiple Comparative Metagenomics using Multiset k-mer Counting. PeerJ Computer Science, 2 :e94.

[Borenstein et al., 2008] Borenstein, E., Kupiec, M., Feldman, M. W., and Ruppin, E. (2008). Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences, 105(38) :14482–14487.

[Bork et al., 2015] Bork, P., Bowler, C., De Vargas, C., Gorsky, G., Karsenti, E., and Wincker, P. (2015). Tara Oceans studies plankton at Planetary scale. Science, 348(6237) :873.

[Brandt et al., 2017] Brandt, J. W., Chevignon, G., Oliver, K. M., and Strand, M. R. (2017). Culture of an aphid heritable symbiont demonstrates its direct role in defence against parasitoids. Proceedings of the Royal Society B : Biological Sciences, 284(1866) :20171925.

[Bright and Bulgheresi, 2010] Bright, M. and Bulgheresi, S. (2010). A complex journey : Transmission of microbial symbionts. Nature Reviews Microbiology, 8(3) :218–230.

[Brooks et al., 2016] Brooks, A. W., Kohl, K. D., Brucker, R. M., van Opstal, E. J., and Bordenstein, S. R. (2016). Phylosymbiosis : Relationships and Functional Effects of Microbial Communities across Host Evolutionary History. PLoS Biology, 14(11) :e2000225.

[Brown, 2015] Brown, C. T. (2015). Strain recovery from metagenomes. Nature Publishing Group, 33(10) :1041–1043.

[Buchfink et al., 2014] Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1) :59–60.

[Buchner, 1965] Buchner, P. (1965). Endosymbiosis of Animals with Plant Microorganisms. John Wiley & Sons, New York.

[Burke and Moran, 2011] Burke, G. R. and Moran, N. A. (2011). Massive genomic decay in Serratia symbiotica, a recently evolved symbiont of aphids. Genome Biology and Evolution, 3(1) :195–208.

[Burton et al., 2014] Burton, J. N., Liachko, I., Dunham, M. J., and Shendure, J. (2014). Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. G3&amp ;#58 ; Genes | Genomes | Genetics, 4(7) :1339–1346.

[Buza et al., 2015] Buza, K., Wilczynski, B., and Dojer, N. (2015). RECORD : Reference-assisted genome assembly for closely related genomes. International Journal of Genomics, 2015 :1–10.

[Caporaso et al., 2010] Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature methods, 7(5) :335–6.

[Carding et al., 2015] Carding, S., Verbeke, K., Vipond, D. T., Corfe, B. M., and Owen, L. J. (2015). Dysbiosis of the gut microbiota in disease. Microbial Ecology in Health & Disease, 26(0).

[Cecchini et al., 2013] Cecchini, D. A., Laville, E., Laguerre, S., Robe, P., Leclerc, M., Doré, J., Henrissat, B., Remaud-Siméon, M., Monsan, P., and Potocki-Véronèse, G. (2013). Functional Metagenomics Reveals Novel Pathways of Prebiotic Breakdown by Human Gut Bacteria. PLoS ONE, 8(9) :e72766.

[Chevignon et al., 2018] Chevignon, G., Boyd, B. M., Brandt, J. W., Oliver, K. M., and Strand, M. R. (2018). Culture-Facilitated Comparative Genomics of the Facultative Symbiont Hamiltonella defensa. Genome Biology and Evolution, 10(3) :786–802.

[Chikhi and Rizk, 2012] Chikhi, R. and Rizk, G. (2012). Space-efficient and exact de Bruijn graph representation based on a bloom filter. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7534 LNBI(1) :236–248.

[Cho and Blaser, 2012] Cho, I. and Blaser, M. J. (2012). The human microbiome : At the interface of health and disease. Nature Reviews Genetics, 13(4) :260–270.

[Chong and Moran, 2018] Chong, R. A. and Moran, N. A. (2018). Evolutionary loss and replacement of Buchnera, the obligate endosymbiont of aphids. ISME Journal, 12(3) :898–908.

[Chrostek et al., 2017] Chrostek, E., Pelz-Stelinski, K., Hurst, G. D., and Hughes, G. L. (2017). Horizontal transmission of intracellular insect symbionts via plants. Frontiers in Microbiology, 8(NOV) :2237.

[Cleary et al., 2015] Cleary, B., Brito, I. L., Huang, K., Gevers, D., Shea, T., Young, S., and Alm, E. J. (2015). Detection of low-abundance bacterial strains in metagenomic datasets by eigengenome partitioning. Nature Biotechnology, 33(10) :1053–1060.

[Darling et al., 2004] Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve : Multiple alignment of conserved genomic sequence with rearrangements. Genome Research, 14(7) :1394–1403.

[David et al., 2014] David, L. A., Maurice, C. F., Carmody, R. N., Gootenberg, D. B., Button, J. E., Wolfe, B. E., Ling, A. V., Devlin, A. S., Varma, Y., Fischbach, M. A., Biddinger, S. B., Dutton, R. J., and Turnbaugh, P. J. (2014). Diet rapidly and reproducibly alters the human gut microbiome. Nature, 505(7484) :559–563.

[Dawkins, 1999] Dawkins, R. (1999). The extended phenotype : the long reach of the gene.

[De Bary, 1879] De Bary, A. (1879). Die erscheinung der symbiose. Verlag von Karl J. Tr{ü}bner.

[Dedryver et al., 2010] Dedryver, C. A., Le Ralec, A., and Fabre, F. (2010). The conflicting relationships between aphids and men : A review of aphid damage and control strategies. Comptes Rendus - Biologies, 333(6-7) :539–553.

[Degnan and Moran, 2008] Degnan, P. H. and Moran, N. A. (2008). Diverse phage-encoded toxins in a protective insect endosymbiont. Applied and Environmental Microbiology, 74(21) :6782–6791.

[Degnan et al., 2011] Degnan, P. H., Ochman, H., and Moran, N. A. (2011). Sequence conservation and functional constraint on intergenic spacers in reduced genomes of the obligate symbiont Buchnera. PLoS Genetics, 7(9) :e1002252.

[Delmont et al., 2011] Delmont, T. O., Robe, P., Cecillon, S., Clark, I. M., Constancias, F., Simonet, P., Hirsch, P. R., and Vogel, T. M. (2011). Accessing the soil metagenome for studies of microbial diversity. Applied and Environmental Microbiology, 77(4) :1315–1324.

[Denisov et al., 2008] Denisov, G., Walenz, B., Halpern, A. L., Miller, J., Axelrod, N., Levy, S., and Sutton, G. (2008). Consensus generation and variant detection by Celera Assembler. Bioinformatics, 24(8) :1035–1040.

[DeSantis et al., 2006] DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and Environmental Microbiology, 72(7) :5069–5072.

[Donia et al., 2014] Donia, M., Cimermancic, P., Schulze, C., Wieland Brown, L., Martin, J., Mitreva, M., Clardy, J., Linington, R., and Fischbach, M. (2014). A Systematic Analysis of Biosynthetic Gene Clusters in the Human Microbiome Reveals a Common Family of Antibiotics. Cell, 158(6) :1402–1414.

[Douglas, 2008] Douglas, A. E. (2008). Conflict, cheats and the persistence of symbioses. New Phytologist, 177(4) :849–858.

[Douglas and Prosser, 1992] Douglas, A. E. and Prosser, W. A. (1992). Synthesis of the essential amino acid tryptophan in the pea aphid (Acyrthosiphon pisum) symbiosis. Journal of Insect Physiology, 38(8) :565–568.

[Douglas and Werren, 2016] Douglas, A. E. and Werren, J. H. (2016). Holes in the hologenome : Why host-microbe symbioses are not holobionts. mBio, 7(2) :e02099.

[Drezen et al., 2014] Drezen, E., Rizk, G., Chikhi, R., Deltel, C., Lemaitre, C., Peterlongo, P., and Lavenier, D. (2014). GATB : Genome Assembly & Analysis Tool Box. Bioinformatics (Oxford, England), 30(20) :2959–2961.

[Dykhuizen, 2005] Dykhuizen, D. (2005). Species Numbers in Bacteria. Proceedings. California Academy of Sciences, 56(6 Suppl 1) :62–71.

[Esposito and Kirschberg, 2014] Esposito, A. and Kirschberg, M. (2014). How many 16S-based studies should be included in a metagenomic conference ? It may be a matter of etymology. FEMS Microbiology Letters, 351(2) :145–146.

[Finn et al., 2014] Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L., Tate, J., and Punta, M. (2014). Pfam : The protein families database.

[Francis et al., 2013] Francis, O. E., Bendall, M., Manimaran, S., Hong, C., Clement, N. L., Castro-Nallar, E., Snell, Q., Schaalje, G. B., Clement, M. J., Crandall, K. A., and Johnson, W. E. (2013). Pathoscope : Species identification and strain attribution with unassembled sequencing data. Genome Research, 23(10) :1721–1729.

[Funk et al., 2001] Funk, D. J., Wernegreen, J. J., and Moran, N. A. (2001). Intraspecific variation in symbiont genomes : Bottlenecks and the aphid-Buchnera association. Genetics, 157(2) :477–489.

[Gan et al., 2011] Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., Lyngsoe, R., Schultheiss, S. J., Osborne, E. J., Sreedharan, V. T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E. J., Harberd, N. P., Kemen, E., Toomajian, C., Kover, P. X., Clark, R. M., Rätsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature, 477(7365) :419–423.

[Garza and Dutilh, 2015] Garza, D. R. and Dutilh, B. E. (2015). From cultured to uncultured genome sequences : Metagenomics and modeling microbial ecosystems. Cellular and Molecular Life Sciences, 72(22) :4287–4308.

[Gehrer and Vorburger, 2012] Gehrer, L. and Vorburger, C. (2012). Parasitoids as vectors of facultative bacterial endosymbionts in aphids. Biology Letters, 8(4) :613–615.

[Gil et al., 2002] Gil, R., Sabater-Munoz, B., Latorre, A., Silva, F. J., and Moya, A. (2002). Extreme genome reduction in Buchnera spp. : Toward the minimal genome needed for symbiotic life. Proceedings of the National Academy of Sciences, 99(7) :4454–4458.

[Goldfeder et al., 2017] Goldfeder, R. L., Wall, D. P., Khoury, M. J., Ioannidis, J. P. A., and Ashley, E. A. (2017). Human Genome Sequencing at the Population Scale : A Primer on High-Throughput DNA Sequencing and Analysis. American Journal of Epidemiology, 186(8) :1000–1009.

[Gonnella and Kurtz, 2017] Gonnella, G. and Kurtz, S. (2017). GfaPy : A flexible and extensible software library for handling sequence graphs in Python. Bioinformatics, 33(19) :3094–3095.

[Goodwin et al., 2016] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age : Ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6) :333–351.

[Guay et al., 2009] Guay, J. F., Boudreault, S., Michaud, D., and Cloutier, C. (2009). Impact of environmental stress on aphid clonal resistance to parasitoids : Role of Hamiltonella defensa bacterial symbiosis in association with a new facultative symbiont of the pea aphid. Journal of Insect Physiology, 55(10) :919–926.

[Gurevich et al., 2013] Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST : Quality assessment tool for genome assemblies. Bioinformatics, 29(8) :1072–1075.

[Guyomar et al., 2018] Guyomar, C., Legeai, F., Jousselin, E., Mougel, C., Lemaitre, C., and Simon, J.-C. (2018). Multi-scale characterization of symbiont diversity in the pea aphid complex through metagenomic approaches. Microbiome, in press.

[Hahn et al., 2013] Hahn, C., Bachmann, L., and Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads - A baiting and iterative mapping approach. Nucleic Acids Research, 41(13) :e129.

[Handelsman et al., 1998] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes : A new frontier for natural products. Chemistry and Biology, 5(10) :R245–R249.

[Henry et al., 2013] Henry, L. M., Peccoud, J., Simon, J. C., Hadfield, J. D., Maiden, M. J. C., Ferrari, J., and Godfray, H. C. J. (2013). Horizontally transmitted symbionts and host colonization of ecological niches. Current Biology, 23(17) :1713–1717.

[Herre et al., 1999] Herre, E. A., Knowlton, N., Mueller, U. G., and Rehner, S. A. (1999). The evolution of mutualisms : Exploring the paths between conflict and cooperation. Trends in Ecology and Evolution, 14(2) :49–53.

[Hoff et al., 2009] Hoff, K. J., Lingner, T., Meinicke, P., and Tech, M. (2009). Orphelia : Predicting genes in metagenomic sequencing reads. Nucleic Acids Research, 37(SUPPL. 2).

[Huson et al., 2016] Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H. J., and Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLoS Computational Biology, 12(6) :e1004957.

[Huttenhower et al., 2012] Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., Fitzgerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayvaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S., Chen, I. M. A., Chen, L., Chhibba,

S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., Desantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Michael Dunne, W., Scott Durkin, A., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. A., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang, K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C. C., Lozupone, C. A., Dwayne Lunsford, R., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., Oglaughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y. H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Methé, B. A., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K., and White, O. (2012). Structure, function and diversity of the healthy human microbiome. Nature, 486(7402) :207–214.

[Imelfort et al., 2014] Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P., and Tyson, G. W. (2014). GroopM : an automated tool for the recovery of population genomes from related metagenomes. PeerJ, 2 :e603.

[Kanehisa, 2004] Kanehisa, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Research, 32(90001) :277D–280.

[Keegan et al., 2016] Keegan, K. P., Glass, E. M., and Meyer, F. (2016). MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In Methods in molecular biology (Clifton, N.J.), volume 1399, pages 207–233.

[Kim et al., 2016] Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge : Rapid and sensitive classification of metagenomic sequences. Genome Research, 26(12) :1721–1729.

[Klein et al., 2011] Klein, J. D., Ossowski, S., Schneeberger, K., Weigel, D., and Huson, D. H. (2011). LOCAS – A Low Coverage Assembly Tool for Resequencing Projects. PLoS ONE, 6(8) :e23455.

[Koren et al., 2012] Koren, O., Goodrich, J. K., Cullender, T. C., Spor, A., Laitinen, K., Kling Bäckhed, H., Gonzalez, A., Werner, J. J., Angenent, L. T., Knight, R., Bäckhed, F., Isolauri, E., Salminen, S., and Ley, R. E. (2012). Host remodeling of the gut microbiome and metabolic changes during pregnancy. Cell, 150(3) :470–480.

[Koren et al., 2017] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu : Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Research, 27(5) :722–736.

[Kreimer et al., 2012] Kreimer, A., Doron-Faigenboim, A., Borenstein, E., and Freilich, S. (2012). NetCmpt : A network-based tool for calculating the metabolic competition between bacterial species. Bioinformatics, 28(16) :2195–2197.

[Laczny et al., 2015] Laczny, C. C., Sternal, T., Plugaru, V., Gawron, P., Atashpendar, A., Margossian, H. H., Coronado, S., der Maaten, L. V., Vlassis, N., and Wilmes, P. (2015). VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome, 3(1).

[Landman et al., 2014] Landman, S. R., Hwang, T. H., Silverstein, K. A., Li, Y., Dehm, S. M., Steinbach, M., and Kumar, V. (2014). SHEAR : Sample heterogeneity estimation and assembly by reference. BMC Genomics, 15(1) :84.

[Leclair et al., 2017] Leclair, M., Polin, S., Jousseaume, T., Simon, J. C., Sugio, A., Morlière, S., Fukatsu, T., Tsuchida, T., and Outreman, Y. (2017). Consequences of coinfection with protective symbionts on the host phenotype and symbiont titres in the pea aphid system. Insect Science, 24(5) :798–808.

[Leclair et al., 2016] Leclair, M., Pons, I., Mahéo, F., Morlière, S., Simon, J. C., and Outreman, Y. (2016). Diversity in symbiont consortia in the pea aphid complex is associated with large phenotypic variation in the insect host. Evolutionary Ecology, 30(5) :925–941.

[Levy et al., 2015] Levy, R., Carr, R., Kreimer, A., Freilich, S., and Borenstein, E. (2015). NetCooperate : A network-based tool for inferring host-microbe and microbe-microbe cooperation. BMC Bioinformatics, 16(1).

[Li et al., 2015] Li, D., Liu, C. M., Luo, R., Sadakane, K., and Lam, T. W. (2015). MEGAHIT : An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics, 31(10) :1674–1676.

[Li and Durbin, 2009] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14) :1754–1760.

[Liu and Pop, 2010] Liu, B. and Pop, M. (2010). Identifying differentially abundant metabolic pathways in metagenomic datasets. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6053 LNBI(Suppl 2) :101–112.

[Luo et al., 2015] Luo, C., Knight, R., Siljander, H., Knip, M., Xavier, R. J., and Gevers, D. (2015). ConStrains identifies microbial strains in metagenomic datasets. Nature Biotechnology, 33(10) :1045–1052.

[Maillet et al., 2014] Maillet, N., Collet, G., Vannier, T., Lavenier, D., and Peterlongo, P. (2014). Commet : Comparing and combining multiple metagenomic datasets. In Proceedings - 2014 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2014, pages 94–98. IEEE.

[Mande et al., 2012] Mande, S. S., Mohammed, M. H., and Ghosh, T. S. (2012). Classification of metagenomic sequences : Methods and challenges. Briefings in Bioinformatics, 13(6) :669–681.

[Mandrioli and Manicardi, 2013] Mandrioli, M. and Manicardi, G. (2013). Evolving aphids : one genome-one organism insects or holobionts ? Invertebrate Survival Journal, 10 :1–6.

[Manzano-Marín and Latorre, 2016] Manzano-Marín, A. and Latorre, A. (2016). Snapshots of a shrinking partner : Genome reduction in Serratia symbiotica. Scientific Reports, 6(1) :32590.

[Marchesi and Ravel, 2015] Marchesi, J. R. and Ravel, J. (2015). The vocabulary of microbiome research : a proposal. Microbiome, 3(1) :31.

[Margulis, 1976] Margulis, L. (1976). Genetic and evolutionary consequences of symbiosis. Experimental Parasitology, 39(2) :277–349.

[Margulis et al., 1991] Margulis, L., Fester, R., and Fester, R. (1991). Symbiosis as a source of evolutionary innovation : speciation and morphogenesis. MIT Press.

[Martin and Schwab, 2012] Martin, B. D. and Schwab, E. (2012). Current usage of symbiosis and associated terminology. International Journal of Biology, 5(1) :32.

[McDowall and Hunter, 2011] McDowall, J. and Hunter, S. (2011). InterPro protein classification. Methods in molecular biology (Clifton, N.J.), 694 :37–47.

[Menzel et al., 2016] Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nature Communications, 7.

[Mira and Moran, 2002] Mira, A. and Moran, N. A. (2002). Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. Microbial Ecology, 44(2) :137–143.

[Mittler, 1971] Mittler, T. E. (1971). Some effects on the aphid Myzus persicae of ingesting antibiotics incorporated into artificial diets. Journal of Insect Physiology, 17(7) :1333–1347.

[Miura et al., 2003] Miura, T., Braendle, C., Shingleton, A., Sisk, G., Kambhampati, S., and Stern, D. L. (2003). A Comparison of Parthenogenetic and Sexual Embryogenesis of the Pea Aphid Acyrthosiphon pisum (Hemiptera : Aphidoidea). Journal of Experimental Zoology Part B : Molecular and Developmental Evolution, 295(1) :59–81.

[Montllor et al., 2002] Montllor, C. B., Maxmen, A., and Purcell, A. H. (2002). Facultative bacterial endosymbionts benefit pea aphids Acyrthosiphon pisum under heat stress. Ecological Entomology, 27(2) :189–195.

[Moran, 1996] Moran, N. A. (1996). Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proceedings of the National Academy of Sciences, 93(7) :2873–2878.

[Moran and Baumann, 2000] Moran, N. A. and Baumann, P. (2000). Bacterial endosymbionts in animals. Current Opinion in Microbiology, 3(3) :270–275.

[Moran and Dunbar, 2006] Moran, N. A. and Dunbar, H. E. (2006). Sexual acquisition of beneficial symbionts in aphids. Proceedings of the National Academy of Sciences, 103(34) :12803–12806.

[Moran et al., 2008] Moran, N. A., McCutcheon, J. P., and Nakabachi, A. (2008). Genomics and Evolution of Heritable Bacterial Symbionts. Annual Review of Genetics, 42(1) :165–190.

[Moran et al., 1993] Moran, N. A., Munson, M. A., Baumann, P., and Ishikawa, H. (1993). A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proceedings of the Royal Society B : Biological Sciences, 253(1337) :167–171.

[Moran and Sloan, 2015] Moran, N. A. and Sloan, D. B. (2015). The Hologenome Concept : Helpful or Hollow ? PLoS Biology, 13(12) :e1002311.

[Morrison and Peairs, 1998] Morrison, W. P. and Peairs, F. B. (1998). Response model concept and economic impact. Response model for an introduced pest—the Russian wheat aphid. Lanham, MD : Entomological Society of America.

[Moya et al., 2008] Moya, A., Peretó, J., Gil, R., and Latorre, A. (2008). Learning how to live together : Genomic insights into prokaryote-animal symbioses. Nature Reviews Genetics, 9(3) :218–229.

[Nakabachi, 2015] Nakabachi, A. (2015). Horizontal gene transfers in insects. Current Opinion in Insect Science, 7 :24–29.

[Namiki et al., 2012] Namiki, T., Hachiya, T., Tanaka, H., and Sakakibara, Y. (2012). MetaVelvet : An extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research, 40(20) :e155.

[Nault, 1997] Nault, L. R. (1997). Arthropod Transmission of plant viruses : A new synthesis. Annals of the Entomological Society of America, 90(5) :521–541.

[Nikoh et al., 2010] Nikoh, N., McCutcheon, J. P., Kudo, T., Miyagishima, S. Y., Moran, N. A., and Nakabachi, A. (2010). Bacterial genes in the aphid genome : Absence of functional gene transfer from Buchnera to its host. PLoS Genetics, 6(2) :e1000827.

[Nishiguchi et al., 1998] Nishiguchi, M. K., Ruby, E. G., and McFall-Ngai, M. J. (1998). Competitive dominance among strains of luminous bacteria provides an unusual form of evidence for parallel evolution in sepiolid squid-vibrio symbioses. Applied and Environmental Microbiology, 64(9) :3209–3213.

[Nurk et al., 2016] Nurk, S., Meleshko, D., Korobeynikov, A., and Pevzner, P. (2016). metaSPAdes : a new versatile de novo metagenomics assembler. aRxiv, (2004) :arXiv :1604.03071.

[Oerke et al., 1994] Oerke, E. C., Dehne, H. W., Schonbeck, F., and Weber, A. (1994). Estimated crop losses in wheat. Crop production and crop protection : estimated losses in major food and cash crops. Amsterdam : Elsevier, pages 179–296.

[Oliver et al., 2010] Oliver, K. M., Degnan, P. H., Burke, G. R., and Moran, N. A. (2010). Facultative Symbionts in Aphids and the Horizontal Transfer of Ecologically Important Traits. Annual Review of Entomology, 55(1) :247–266.

[Oliver et al., 2009] Oliver, K. M., Degnan, P. H., Hunter, M. S., and Moran, N. A. (2009). Bacteriophages encode factors required for protection in a symbiotic mutualism. Science, 325(5943) :992–994.

[Oliver et al., 2005] Oliver, K. M., Moran, N. a., and Hunter, M. S. (2005). Variation in resistance to parasitism in aphids is due to symbionts not host genotype. Proceedings of the National Academy of Sciences of the United States of America, 102(36) :12795–800.

[Oliver et al., 2003] Oliver, K. M., Russell, J. A., Moran, N. A., and Hunter, M. S. (2003). Facultative bacterial symbionts in aphids confer resistance to parasitic wasps. Proceedings of the National Academy of Sciences, 100(4) :1803–1807.

[Oliver et al., 2014] Oliver, K. M., Smith, A. H., and Russell, J. A. (2014). Defensive symbiosis in the real world - advancing ecological studies of heritable, protective bacteria in aphids and beyond. Functional Ecology, 28(2) :341–355.

[O'Mahony et al., 2015] O'Mahony, S. M., Clarke, G., Borre, Y. E., Dinan, T. G., and Cryan, J. F. (2015). Serotonin, tryptophan metabolism and the brain-gut-microbiome axis. Behavioural Brain Research, 277 :32–48.

[Ondov et al., 2016] Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. (2016). Mash : Fast genome and metagenome distance estimation using MinHash. Genome Biology, 17(1) :132.

[Opatovsky et al., 2018] Opatovsky, I., Santos-Garcia, D., Ruan, Z., Lahav, T., Ofaim, S., Mouton, L., Barbe, V., Jiang, J., Zchori-Fein, E., and Freilich, S. (2018). Modeling trophic dependencies and exchanges among insects' bacterial symbionts in a host-simulated environment. BMC Genomics, 19(1) :402.

[Ounit et al., 2015] Ounit, R., Wanamaker, S., Close, T. J., and Lonardi, S. (2015). CLARK : fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics, 16(1) :236.

[Panina et al., 2001] Panina, E. M., Vitreschak, A. G., Mironov, A. A., and Gelfand, M. S. (2001). Regulation of aromatic amino acid biosynthesis in gamma-proteobacteria. Journal of molecular microbiology and biotechnology, 3(4) :529–543.

[Parks et al., 2015] Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM : Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7) :1043–1055.

[Paten et al., 2017] Paten, B., Novak, A. M., Eizenga, J. M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Research, 27(5) :665–676.

[Peabody et al., 2015] Peabody, M. a., Van Rossum, T., Lo, R., and Brinkman, F. S. L. (2015). Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. BMC bioinformatics, 16(1) :363.

[Peccoud et al., 2014] Peccoud, J., Bonhomme, J., Mahéo, F., de la Huerta, M., Cosson, O., and Simon, J. C. (2014). Inheritance patterns of secondary symbionts during sexual reproduction of pea aphid biotypes. Insect Science, 21(3) :291–300.

[Peccoud et al., 2015] Peccoud, J., Mahéo, F., de la Huerta, M., Laurence, C., and Simon, J. C. (2015). Genetic characterisation of new host-specialised biotypes and novel associations with bacterial symbionts in the pea aphid complex. Insect Conservation and Diversity, 8(5) :484–492.

[Peccoud et al., 2009a] Peccoud, J., Ollivier, A., Plantegenest, M., and Simon, J.-C. (2009a). A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proceedings of the National Academy of Sciences of the United States of America, 106(18) :7495–500.

[Peccoud et al., 2009b] Peccoud, J., Simon, J.-c., Mclaughlin, H. J., and Moran, N. A. (2009b). Post-Pleistocene radiation of the pea aphid complex revealed by rapidly evolving endosymbionts. Proceedings of the National Academy of Sciences of the United States of America, 106(38) :16315–16320.

[Peng et al., 2012] Peng, Y., Leung, H. C. M., Yiu, S. M., and Chin, F. Y. L. (2012). IDBA-UD : A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 28(11) :1420–1428.

[Pérez-Brocal et al., 2006] Pérez-Brocal, V., Gil, R., Ramos, S., Lamelas, A., Postigo, M., Michelena, J. M., Silva, F. J., Moya, A., and Latorre, A. (2006). A small microbial genome : The end of a long symbiotic relationship ? Science, 314(5797) :312–313.

[Pruesse et al., 2012] Pruesse, E., Peplies, J., and Glöckner, F. O. (2012). SINA : Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics, 28(14) :1823–1829.

[Quast et al., 2013] Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project : Improved data processing and web-based tools. Nucleic Acids Research, 41(D1) :D590–D596.

[Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The Uncultured Microbial Majority. Annual Review of Microbiology, 57(1) :369–394.

[Richards et al., 2010] Richards, S., Gibbs, R. A., Gerardo, N. M., Moran, N., Nakabachi, A., Stern, D., Tagu, D., Wilson, A. C., Muzny, D., Kovar, C., Cree, A., Chacko, J., Chandrabose, M. N., Dao, M. D., Dinh, H. H., Gabisi, R. A., Hines, S., Hume, J., Jhangian, S. N., Joshi, V., Lewis, L. R., Liu, Y. S., Lopez, J., Morgan, M. B., Nguyen, N. B., Okwuonu, G. O., Ruiz, S. J., Santibanez, J., Wright, R. A., Fowler, G. R., Hitchens, M. E., Lozado, R. J., Moen, C., Steffen, D., Warren, J. T., Zhang, J., Nazareth, L. V., Chavez, D., Davis, C., Lee, S. L., Patel, B. M., Pu, L. L., Bell, S. N., Johnson, A. J., Vattathil, S., Williams, R. L., Shigenobu, S., Dang, P. M., Morioka, M., Fukatsu, T., Kudo, T., Miyagishima, S. Y., Jiang, H., Worley, K. C., Legeai, F.,

Gauthier, J. P., Collin, O., Zhang, L., Chen, H. C., Ermolaeva, O., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Maglott, D., Murphy, T., Pruitt, K., Sapojnikov, V., Souvorov, A., Thibaud-Nissen, F., Câmara, F., Guigó, R., Stanke, M., Solovyev, V., Kosarev, P., Gilbert, D., Gabaldón, T., Huerta-Cepas, J., Marcet-Houben, M., Pignatelli, M., Moya, A., Rispe, C., Ollivier, M., Quesneville, H., Permal, E., Llorens, C., Futami, R., Hedges, D., Robertson, H. M., Alioto, T., Mariotti, M., Nikoh, N., McCutcheon, J. P., Burke, G., Kamins, A., Latorre, A., Ashton, P., Calevro, F., Charles, H., Colella, S., Douglas, A. E., Jander, G., Jones, D. H., Febvay, G., Kamphuis, L. G., Kushlan, P. F., Macdonald, S., Ramsey, J., Schwartz, J., Seah, S., Thomas, G., Vellozo, A., Cass, B., Degnan, P., Hurwitz, B., Leonardo, T., Koga, R., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S. M., De Vos, M., Duncan, E. J., Evans, J., Ghanim, M., Heddi, A., Kaloshian, I., Vincent-Monegat, C., Parker, B. J., Pérez-Brocal, V., Rahbé, Y., Spragg, C. J., Tamames, J., Tamarit, D., Tamborindeguy, C., Vilcinskas, A., Bickel, R. D., Brisson, J. A., Butts, T., Chang, C. C., Christiaens, O., Davis, G. K., Duncan, E., Ferrier, D., Iga, M., Janssen, R., Lu, H. L., McGregor, A., Miura, T., Smagghe, G., Smith, J., Van Der Zee, M., Velarde, R., Wilson, M., Dearden, P., Edwards, O. R., Gordon, K., Hilgarth, R. S., Rider, S. D., Srinivasan, D., Walsh, T. K., Ishikawa, A., Jaubert-Possamai, S., Fenton, B., Huang, W., Rizk, G., Lavenier, D., Nicolas, J., Smadja, C., Zhou, J. J., Vieira, F. G., He, X. L., Liu, R., Rozas, J., Field, L. M., Campbell, P., Carolan, J. C., Fitzroy, C. I., Reardon, K. T., Reeck, G. R., Singh, K., Wilkinson, T. L., Huybrechts, J., Abdel-Latief, M., Robichon, A., Veenstra, J. A., Hauser, F., Cazzamali, G., Schneider, M., Williamson, M., Stafflinger, E., Hansen, K. K., Grimmelikhuijzen, C. J., Price, D. R., Caillaud, M., Van Fleet, E., Ren, Q., Gatehouse, J. A., Brault, V., Monsion, B., Diaz, J., Hunnicutt, L., Ju, H. J., Pechuan, X., Aguilar, J., Cortés, T., Ortiz-Rivas, B., Martínez-Torres, D., Dombrovsky, A., Dale, R. P., Davies, T. G., Williamson, M. S., Jones, A., Sattelle, D., Williamson, S., Wolstenholme, A., Cottret, L., Sagot, M. F., Heckel, D. G., and Hunter, W. (2010). Genome sequence of the pea aphid Acyrthosiphon pisum. PLoS Biology, 8(2) :e1000313.

[Rizk et al., 2014] Rizk, G., Gouin, A., Chikhi, R., and Lemaitre, C. (2014). MindTheGap : integrated detection and assembly of short and long insertions. Bioinformatics, 30(24) :3451–3457.

[Roesch et al., 2007] Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., Daroub, S. H., Camargo, F. A., Farmerie, W. G., and Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. ISME Journal, 1(4) :283–290.

[Rosenberg and Zilber-Rosenberg, 2011] Rosenberg, E. and Zilber-Rosenberg, I. (2011). Symbiosis and development : The hologenome concept. Birth Defects Research Part C - Embryo Today : Reviews, 93(1) :56–66.

[Rosenberg and Zilber-Rosenberg, 2016] Rosenberg, E. and Zilber-Rosenberg, I. (2016). Do microbiotas warm their hosts ? Gut Microbes, 7(4) :283–285.

[Rosenberg and Zilber-Rosenberg, 2018] Rosenberg, E. and Zilber-Rosenberg, I. (2018). The hologenome concept of evolution after 10 years. Microbiome, 6(1) :78.

[Russell et al., 2003] Russell, J. A., Latorre, A., Sabater-Muñoz, B., Moya, A., and Moran, N. A. (2003). Side-stepping secondary symbionts : Widespread horizontal transfer across and beyond the Aphidoidea. Molecular Ecology, 12(4) :1061–1075.

[Russell and Moran, 2006] Russell, J. A. and Moran, N. A. (2006). Costs and benefits of symbiont infection in aphids : variation among symbionts and across temperatures. Proc. Biol. Sci., 273(1586) :603–610.

[Sachs and Simms, 2006] Sachs, J. L. and Simms, E. L. (2006). Pathways to mutualism breakdown. Trends in Ecology and Evolution, 21(10) :585–592.

[Sandström et al., 2001] Sandström, J. P., Russell, J. A., White, J. P., and Moran, N. A. (2001). Independent origins and horizontal transfer of bacterial symbionts of aphids. Molecular Ecology, 10(1) :217–228.

[Sangwan et al., 2016] Sangwan, N., Xia, F., and Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. Microbiome, 4(1) :8.

[Sankar et al., 2015] Sankar, S. A., Lagier, J. C., Pontarotti, P., Raoult, D., and Fournier, P. E. (2015). The human gut microbiome, a taxonomic conundrum. Systematic and Applied Microbiology, 38(4) :276–286.

[Schloss et al., 2009] Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur : Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology, 75(23) :7537–7541.

[Schouls et al., 2003] Schouls, L. M., Schot, C. S., and Jacobs, J. A. (2003). Horizontal Transfer of Segments of the 16S rRNA Genes between Species of the Streptococcus anginosus Group. Journal of Bacteriology, 185(24) :7241–7246.

[Sczyrba et al., 2017] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., Demaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiutė, M., Hansen, L. H., Sørensen, S. J., Chia, B. K., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y. W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H. H., Liao, Y. C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H. P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nature Methods, 14(11) :1063–1071.

[Sedlazeck et al., 2018] Sedlazeck, F. J., Lee, H., Darby, C. A., and Schatz, M. C. (2018). Piercing the dark matter : Bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics, 19(6) :329–346.

[Segata, 2018] Segata, N. (2018). On the Road to Strain-Resolved Comparative Metagenomics. mSystems, 3(2) :e00190–17.

[Segata et al., 2013] Segata, N., Boernigen, D., Tickle, T. L., Morgan, X. C., Garrett, W. S., and Huttenhower, C. (2013). Computational meta'omics for microbial community studies. Molecular systems biology, 9(1) :666.

[Shigenobu et al., 2000] Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature, 407(6800) :81–86.

[Silva et al., 1998] Silva, F. J., Van Ham, R. C. H. J., Sabater, B., and Latorre, A. (1998). Structure and evolution of the leucine plasmids carried by the endosymbiont (Buchnera aphidicola) from aphids of the family Aphididae. FEMS Microbiology Letters, 168(1) :43–49.

[Simon et al., 2011] Simon, J. C., Boutin, S., Tsuchida, T., Koga, R., Gallic, J. F., Frantz, A., Outreman, Y., and Fukatsu, T. (2011). Facultative symbiont infections affect aphid reproduction. PLoS ONE, 6(7) :e21831.

[Simpson et al., 2009] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS : A parallel assembler for short read sequence data. Genome Research, 19(6) :1117–1123.

[Sommer and Bäckhed, 2013] Sommer, F. and Bäckhed, F. (2013). The gut microbiota-masters of host development and physiology. Nature Reviews Microbiology, 11(4) :227–238.

[Soueidan and Nikolski, 2015] Soueidan, H. and Nikolski, M. (2015). Machine learning for metagenomics : methods and tools.

[Tamas et al., 2002] Tamas, I., Klasson, L., Canbäck, B., Näslund, A. K., Eriksson, A. S., Wernegreen, J. J., Sandström, J. P., Moran, N. A., and Andersson, S. G. (2002). 50 Million years of genomic stasis in endosymbiotic bacteria. Science, 296(5577) :2376–2379.

[Teeling et al., 2004] Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., and Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. Environmental Microbiology, 6(9) :938–947.

[Thao et al., 1998] Thao, M. L., Baumann, L., Baumann, P., and Moran, N. A. (1998). Endosymbionts (Buchnera) from the aphids Schizaphis graminum and Diuraphis noxia have different copy numbers of the plasmid containing the leucine biosynthetic genes. Current Microbiology, 36(4) :238–240.

[Truong et al., 2015] Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C., and Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling.

[Truong et al., 2017] Truong, D. T., Tett, A., Pasolli, E., Huttenhower, C., and Segata, N. (2017). Microbial strain-level population structure & genetic diversity from metagenomes. Genome Research, 27(4) :626–638.

[Tsuchida et al., 2004] Tsuchida, T., Koga, R., and Fukatsu, T. (2004). Host Plant Specialization Governed by Facultative Symbiont. Science, 303(5666) :1989.

[Tsuchida et al., 2010] Tsuchida, T., Koga, R., Horikawa, M., Tsunoda, T., Maoka, T., Matsumoto, S., Simon, J. C., and Fukatsu, T. (2010). Symbiotic bacterium modifies aphid body color. Science, 330(6007) :1102–1104.

[Turnbaugh et al., 2007] Turnbaugh, P. J., Ruth, E., Ley, M. H., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. Nature, 449.

[Venter et al., 2001] Venter, J. C., Adams, M. D. M., Myers, E. E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. L. G., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-r., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-h., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S. C., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., Mccawley, S., Mcintosh, T., Mcmullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y.-h., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-stine, J., Caulk, P., Chiang, Y.-h., Coyne, M., Dahlke, C., Mays, A. D., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., Mcdaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001). The

sequence of the human genome.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H., and Smith, H. O. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science, 304(5667) :66–74.

[Vogel and Moran, 2011] Vogel, K. J. and Moran, N. A. (2011). Effect of host genotype on symbiont titer in the aphid-Buchnera symbiosis. Insects, 2(3) :423–434.

[Walker et al., 2014] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon : An integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE, 9(11) :e112963.

[Wang and Wu, 2013] Wang, Z. and Wu, M. (2013). A phylum-level bacterial phylogenetic marker database. Molecular Biology and Evolution, 30(6) :1258–1262.

[Ward, 1883] Ward, H. M. (1883). Ueber die Entwickelung der Chlorophyllk{ö}rner und Farbk{ö}rper Ueber Chlorophyllk{ö}rner St{ä}rkebildner und Farbk{ö}rper. Nature, 28(716) :267.

[Welch and Huse, 2011] Welch, D. B. and Huse, S. M. (2011). Microbial Diversity in the Deep Sea and the Underexplored "Rare Biosphere". Handbook of Molecular Microbial Ecology II : Metagenomics in Different Habitats, 103(32) :243–252.

[Wernegreen, 2005] Wernegreen, J. J. (2005). For better or worse : Genomic consequences of intracellular mutualism and parasitism. Current Opinion in Genetics and Development, 15(6) :572–583.

[Wick et al., 2015] Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage : Interactive visualization of de novo genome assemblies. Bioinformatics, 31(20) :3350–3352.

[Wood and Salzberg, 2014] Wood, D. E. and Salzberg, S. L. (2014). Kraken : ultrafast metagenomic sequence classification using exact alignments. Genome biology, 15(3) :R46.

[Wu et al., 2014] Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014). MaxBin : an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome, 2(1) :26.

[Xia et al., 2011] Xia, L. C., Cram, J. A., Chen, T., Fuhrman, J. A., and Sun, F. (2011). Accurate genome relative abundance estimation based on shotgun metagenomic reads. PloS one, 6(12) :e27992.

[Zhang and Zhao, 2016] Zhang, C. and Zhao, L. (2016). Strain-level dissection of the contribution of the gut microbiome to human metabolic disease. Genome Medicine, 8(1) :41.

[Zhao, 2010] Zhao, L. (2010). Genomics : The tale of our other genome. Nature, 465(7300) :879–880.

[Zhu et al., 2010] Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Research, 38(12).

[Zientz et al., 2004] Zientz, E., Dandekar, T., and Gross, R. (2004). Metabolic Interdependence of Obligate Intracellular Bacteria and Their Insect Hosts. Microbiology and Molecular Biology Reviews, 68(4) :745–770.

[Łukasik et al., 2013] Łukasik, P., van Asch, M., Guo, H., Ferrari, J., and Charles, H. (2013). Unrelated facultative endosymbionts protect aphids against a fungal pathogen. Ecology Letters, 16(2) :214–218.

UNIVERSITE ECOLOGIE BRETAGNE GEOSCIENCES LOIRE AGRONOMIE ALIMENTATION

UNIVERSITÉ DE RENNES 1

Titre : Développement et applications d'outils d'analyse métagénomique des communautés microbiennes associées aux insectes

Mots clés : puceron, métagénomique, symbiose, interactions hôte-microbiote, assemblage-guidé

Résumé : Le développement de la métagénomique donne des opportunités nouvelles d'explorer les interactions hôte-microbiote, mais pose également des défis méthodologiques. Cette thèse présente des outils nouveaux pour la caractérisation de la diversité métagénomique, et leur application à l'holobionte du puceron du pois. Nous présentons d'abord une approche qui repose sur l'assignation taxonomiques des lectures par alignement à des génomes de référence ou préalablement assemblés, puis sur la recherche de variants génomiques. L'étude de génotypes complets de symbiotes permet de retracer l'histoire évolutive des relations hôte-microbiote avec une résolution élevée. Chez le puceron du pois, nous mettons en évidence des structures de diversité génomique différents selon les symbiotes, que nous proposons d'expliquer par les modalités de transmission ou l'histoire évolutive propre à chacun des partenaires microbiens. Dans un second temps, nous présentons une méthode d'assemblage guidé par référence en métagénomique. Cette méthode se déroule en deux temps : l'assemblage de lectures s'alignant sur un génome de référence distant, puis l'assemblage ciblé des régions manquantes par une version étendue du logiciel *MindTheGap*. Par rapport à un assembleur métagénomique, cette méthode permet l'assemblage d'un seul génome en un temps réduit, et permet de détecter d'éventuels variants structuraux sur le génome ciblé. Appliqué au puceron du pois, *MindTheGap* a réalisé l'assemblage du symbiote obligatoire *Buchnera* en un seul contig, et a permis d'identifier différents variants structuraux du bactériophage APSE. Ces travaux ouvrent la voie à la fois à une caractérisation plus précise des relations hôte-microbiote chez le puceron du pois, ainsi qu'à l'application des outils présentés à des systèmes plus complexes.

Title : Bioinformatic tools and applications for metagenomics of microbial communities associated to insects

Keywords : aphid, metagenomics, symbiosis, host-microbiota interactions, guided-assembly

Abstract : The rise of metagenomics gives new opportunities to explore host-microbiota interactions, but raises new methodological challenges. Here, we present new tools for the characterization of metagenomics diversity, and their application to the pea aphid holobiont. We first present an approach based on the taxonomic assignation of reads by mapping to reference genomes, and the detection of short variants. Comparing whole genome genotypes makes it possible to sketch the evolutionary histories of host-microbiota interactions at a high resolution. For the pea aphid, we highlight different scales of genomic diversity for the different symbionts, that we link to the evolutionary history of each microbial partner and the modalities of their association with the host. In a second time, we present a method for reference-guided genome assembly from metagenomic data. It is based on two steps. First, reads are recruited by mapping to a distant genome, and assembled into backbone contigs. Then, missiong regions are assembled by an enhanced version of the software *MindTheGap*. It outperforms metagenomic assembly, and enables an efficient assembly of a single genome from metagenomic data. In the pea aphid complex, we applied it to assemble the primary symbiont *Buchnera* in a single contig, and to detect structural variations of the bacteriophage APSE. Overall, this work paves the way to a better understanding of the aphid-microbiota relationships, and the application of the presented approaches to more complex systems.