Museum National d'Histoire Naturelle

École doctorale Science de la nature et de l'homme : écologie et évolution

Mécanismes Adaptatifs et Évolution - UMR7179

Équipe Function et Evolution

Institut de Biologie Paris Seine - UMR 7138

Équipe Adaptation, Intégration, Réticulation et Evolution

Étude de l'impact d'un changement de régime alimentaire sur le microbiome intestinal *de Podarcis sicula*

Par **Chloé VIGLIOTTI**

Thèse de doctorat en Sciences de la vie et de la santé

Dirigée par Eric BAPTESTE et Anthony HERREL

Présentée et soutenue publiquement le 20 novembre 2017 devant :

Dr Eric BAPTESTE (DR CNRS, Université Pierre et Marie Curie)	Encadrant
Dr Samuel CHAFFRON (CR CNRS, Université de Nantes)	Examinateur
Dr Philippe GERARD (DR INRA, Jouy en Josas)	Examinateur
Dr Anthony HERREL (DR CNRS, MNHN)	Encadrant
Pr Philippe LOPEZ (PR, Université Pierre et Marie Curie)	Encadrant
Dr Nicolas POLLET (DR CNRS, Université Paris-Sud)	Rapporteur
Pr Marc-André SELOSSE (PR, MNHN)	Examinateur
Pr Jesse SHAPIRO (PR, Université de Montréal)	Rapporteur

Remerciements

Je souhaite remercier Eric Bapteste, pour son encadrement, sa disponibilité, sa patience, et son investissement. J'ai beaucoup apprécié toutes nos conversations, qui m'ont beaucoup apporté. La formation que j'ai reçue auprès de toi, me sera très utile pour la suite.

Je souhaite également remercier Philippe Lopez, pour son encadrement, mais aussi pour toutes les fois où je lui ai demandé de l'aide dans l'urgence, et où il me l'a apportée. Je souhaite également le remercier de s'être autant investi dans mes soucis administratifs.

Je souhaite également remercier Anthony Herrel de m'avoir encadrée, et emmenée avec lui sur ces îles paradisiaques en Croatie, me permettant ainsi de faire de la biologie de terrain.

Je remercie chaleureusement Catherine Ozouf, d'avoir pris le temps de m'apprendre à réaliser des caryotypes de lézards, de m'avoir donné le savoir-faire, les outils, et les conseils pour mener à bien ma mission sur les îles.
Je remercie également l'équipe avec qui j'ai été sur le terrain, d'avoir été si bienveillante, et agréable, plus particulièrement Nina, et Anne-Claire, que j'ai eu le bonheur de rencontrer lors de cette mission.

Enfin, je souhaite remercier l'ensemble des doctorants de l'IBPS, plus particulièrement Arnaud et Anne-Sophie, qui m'ont aidée durant ma thèse, mais aussi Juliette, pour sa gentillesse et sa bienveillance.

Je souhaite également remercier les membres de l'équipe AIRE. Tout d'abord, Eduardo, pour son encadrement sur le chapitre de livre que nous avons écrit ensembles, mais aussi pour sa bonne humeur quotidienne. Je remercie Raphaël, pour sa bonne humeur, tous les bons souvenirs de pauses café, les bières, et son sens de l'humour qui a mis une ambiance vraiment agréable dans le labo. Je remercie Guillaume Andrew et Romain, pour leur gentillesse, leur disponibilité et leur bonne humeur. Merci Guillaume de ton aide et ton soutien pour ces derniers jours de rédaction difficiles. Enfin, je remercie Jananan, avec qui j'ai passé mes 3 ans de thèse. Je ne connais pas de personne plus dévouée, merci pour toute l'aide que tu m'as apportée pendant 3 ans. Merci pour toutes les discussions que nous avons eues. Merci pour ton soutien quand il y en avait besoin. Merci pour le chocolat.

Je souhaite remercier les personnes avec qui j'ai collaboré : Michel Habib, Léo Planche, Finn Völkel. Je souhaite également remercier Lucie Bittner, François-Joseph Lapointe et Maïté Ribère pour leur aide.

Enfin, je souhaite remercier mes proches, de leur soutien précieux pendant ces trois ans. Je souhaite également remercier plus particulièrement, Anthony Nina Diogo, pour son soutien et pour nos midis à Jussieu. Enfin, je souhaite remercier tous ceux qui m'ont soutenu la semaine précédant l'envoi de mon manuscrit.

TABLE DES MATIERES

Liste des figures

1. INTRODUCTION

1.1 Contexte de l'étude des microbiomes de *Podarcis sicula*

Podarcis sicula est une espèce de lézard présente dans différents pays, dont la Croatie. Au début des années 1970, Nevo *et al.* (1972)(Nevo et al. 1972) ont choisi d'étudier la compétitivité interspécifique entre des *Podarcis sicula* et des *Podarcis melisellensis* sur plusieurs îles croates. Ils ont alors introduit 10 *Podarcis sicula* insectivores de l'île de Pod Kopište sur l'île de Pod Mrčaru et 10 *Podarcis melisellensis* de l'île de Pod Mrčaru sur l'île de Pod Kopište. Ces îles mesurent environ 1km carré et sont présentées Figure 1.

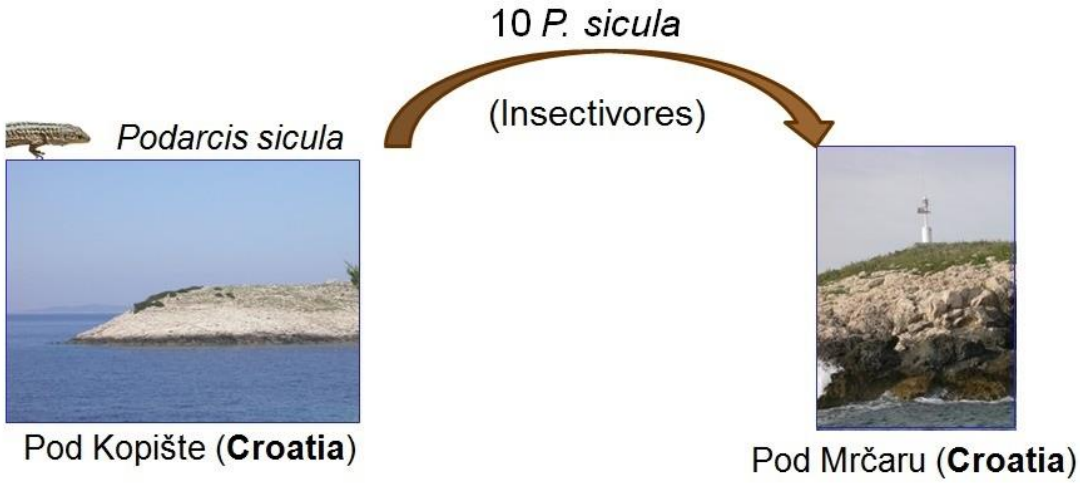


Figure 1 : Contexte de l'étude des microbiomes intestinaux des Podarcis sicula.

35 ans plus tard, une équipe de scientifiques, incluant le Dr Anthony Herrel, est revenue sur les îles. Ils ont observé que les *Podarcis melisellensis* avaient disparu et que les *Podarcis sicula* sur l'île de Pod Mrčaru étaient devenus omnivores (avec un régime alimentaire à 80% herbivore)(Herrel 2007; Herrel et al. 2008; Herrel, Vanhooydonck, and Van Damme 2004). L'herbivorie chez le lézard est un phénomène assez rare (0,8% des espèces décrites ont un régime alimentaire constitué à 90% de plantes). L'omnivorie (le lézard a un régime alimentaire contenant de 10% à 90% de plantes), bien que plus courante, demeure peu répandue (Cooper Jr and Vitt 2002). De nombreux changements morphologiques sont corrélés avec ce changement de régime alimentaire. En effet, les lézards sont plus larges, et on peut noter l'apparition spécifique d'une valve cæcale (Figure 2) dans l'intestin (Dearing 1993; Herrel et al. 2008; Herrel, Vanhooydonck, and Van Damme 2004; Iverson 1980).

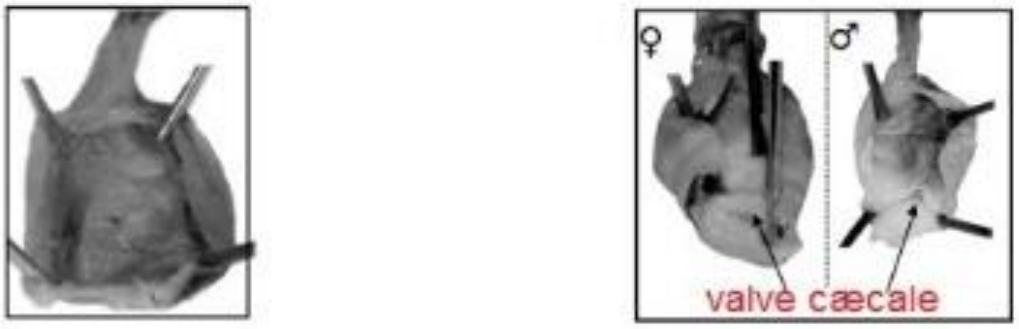


Figure 2: Apparition d'une valve cæcale dans l'intestin des Podarcis sicula omnivores (photos issues de l'article (Herrel et al. 2008)*).*

A gauche, exemple d'un Intestin de Podarcis sicula insectivore

A droite, Intestin de Podarcis sicula omnivore

Compte-tenu de l'apparition de ce nouvel organe, l'étude du microbiote (contenu en micro-organismes) et du microbiome (contenu en gènes microbiens) intestinaux associés à ce changement de régime alimentaire semblait incontournable. Au-delà des modifications morphologiques, l'acquisition d'un régime omnivore à 80 % herbivore implique que les lézards sont capables de digérer des plantes. Or, la digestion de plantes par des vertébrés dépend du microbiote intestinal et nécessite la présence de bactéries capables de digérer la cellulose (Ley et al. 2008; Zhu et al. 2011).

L'herbivorie a été étudiée chez l'iguane (animal appartenant au même ordre que les *Podarcis sicula*), notamment le lien entre le contenu microbien de l'intestin de l'iguane et son herbivorie (Hong et al. 2011). Pour cela, des études utilisant la séquence de l'ARN ribosomique 16S, un constituant de la petite sous-unité des ribosomes des procaryotes) comme marqueur phylogénétique ont été effectuées, démontrant que les 2 phyla majoritaires du microbiote intestinal des iguanes herbivores des îles des Galápagos sont les Firmicutes et les Bacteroides. Des phyla majoritaires ont aussi été trouvé dans les microbiotes intestinaux d'organismes modèles tels que l'homme (dans le cadre des impressionnants travaux du Human Microbiome Project) et la souris. Ainsi, chez la plupart des Mammifères, quatre phyla semblent majoritairement présents dans l'intestin : Actinobactéries, Bactéroidetes, Firmicutes et Protéobactéries (Kinross, Darzi, and Nicholson 2011; Ley, Knight, and

Gordon 2007).Néanmoins, les études de microbiomes sur des organismes non modèles sont encore peu développées. En se concentrant sur l'espèce *Podarcis sicula*, cette thèse se place donc dans le cadre des analyses de microbiome d'hôtes non modèles comme l'illustre la Figure 3.

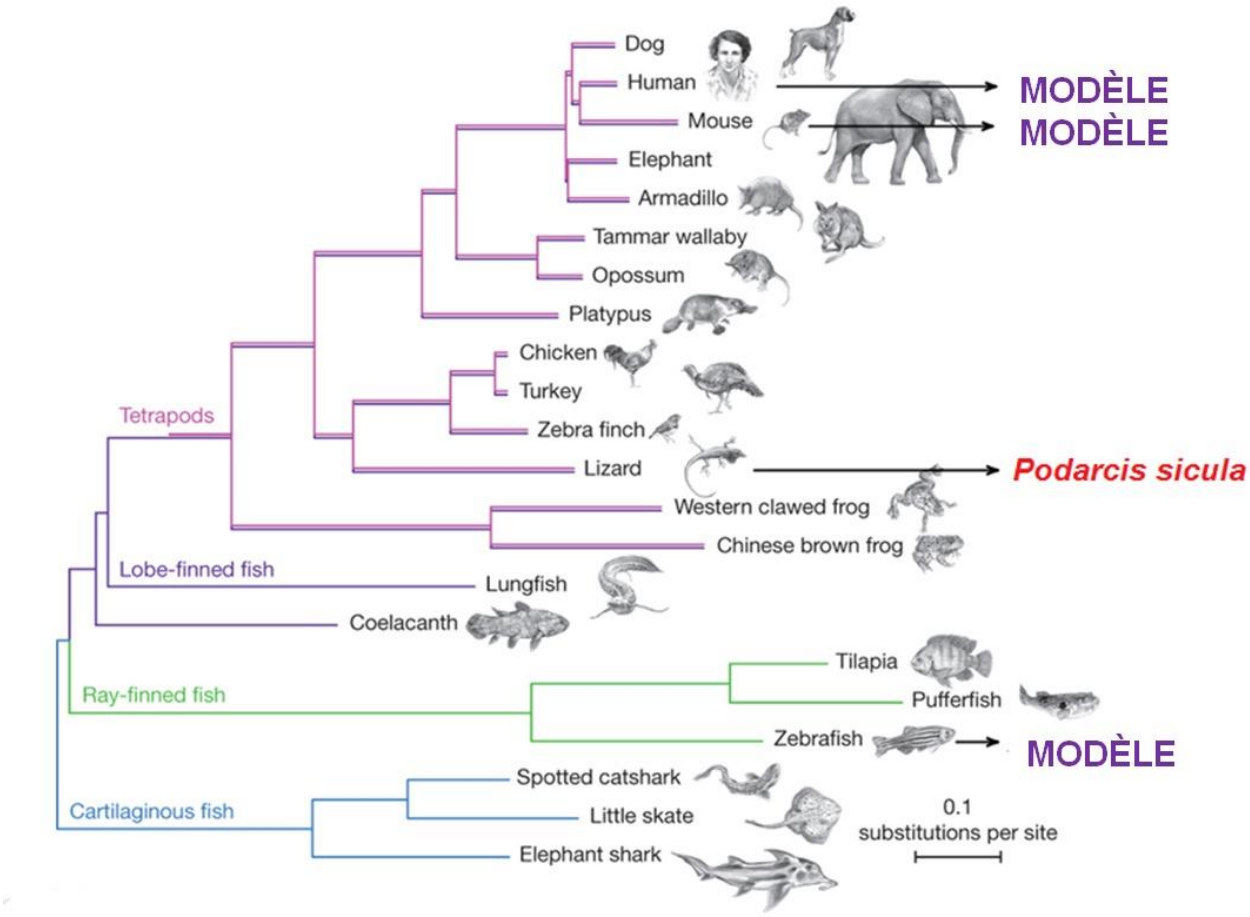


Figure 3 : Position phylogénétique de Podarcis sicula.

Adaptée de Amemiya & al. Nature 2013

L'une des questions soulevées dans notre étude est : peut-on identifier des phyla majoritaires dans le microbiote des *Podarcis sicula* ? Ces phyla sont-ils identiques pour les *Podarcis sicula* omnivores et insectivores ?

1.2 De l'individu à l'holobionte

L'étude du changement de régime alimentaire des *Podarcis sicula* et de ses adaptations morphologiques est complétée par des analyses de microbiotes et microbiomes intestinaux afin d'appréhender l'ensemble du système hôte (*Podarcis sicula*) - microorganismes. Si on peut dater la découverte de la communauté microbienne orale à 1676 avec les travaux de Leeuwenhoek (Escobar-zepeda, León, and Sanchez-flores 2015), depuis la découverte de l'existence d'une communauté microbienne intestinale chez les animaux (et plus particulièrement chez l'homme), la conception biologique et philosophique des animaux n'a cessé d'évoluer.

En 1991, le mot holobionte a été créé par Lynn Margulis (Guerrero, Margulis, and Berlanga 2013; Margulis and Fester 1991) pour désigner un hôte eucaryote et ses microbiotes. La création de ce mot permet ainsi de définir explicitement une nouvelle unité de base, qui n'est plus l'organisme eucaryote, mais l'holobionte (soit l'eucaryote et ses microorganismes). Ce changement de conception des individus hôtes a permis de voir émerger des théories sur la co-dépendance des hôtes et de leurs microorganismes. La plupart des animaux, longtemps considérés comme des entités eucaryotes indépendantes et autonomes, sont désormais souvent considérés comme des eucaryotes avec leurs différentes communautés microbiennes (orales, intestinales, cutanées,...). Les études actuelles montrent que ces différents microbiotes sont essentiels à la survie de leurs hôtes. En effet, d'une part, les microorganismes bénéficient d'un environnement sélectif, qui les alimente régulièrement en nutriments ; d'autre part, l'hôte bénéficie d'une activité microbienne qui complète ses voies digestives (ce qui permet une meilleure absorption des aliments par l'organisme), mais aussi d'une activité microbienne renforçant l'immunité de l'hôte (destruction des xénobiotiques, régulation de l'homéostasie) (Belkaid and Hand 2014; Eberl 2010; Selosse, Bessis, and Pozo 2014; Wu and Wu 2012). Par exemple, certaines bactéries symbiotiques sont impliquées dans l'immunité innée de l'hôte en occupant des sites potentiels d'adhésion et en produisant des antibiotiques (Ritchie 2006; Zilber-Rosenberg and Rosenberg 2008). Il est également possible d'illustrer à quel point ce microbiote est indispensable à la digestion : sans microorganismes, un vertébré est dans l'incapacité de digérer des plantes, dans la

mesure où les vertébrés sont incapables de digérer la cellulose (Ley et al. 2008; Zhu et al. 2011).

D'un point de vue évolutionniste, la notion d'holobionte permet de proposer une nouvelle unité de sélection : l'hologénome (ensemble des gènes microbiens et eucaryotes contenus par l'holobionte) (Bordenstein and Theis 2015; Margulis and Fester 1991; Zilber-Rosenberg and Rosenberg 2008, 2013). Dans ce cadre, des recherches visent à comprendre comment l'eucaryote et ses microorganismes co-évoluent (Bäckhed et al. 2005; NEISH 2009; Thursby and Juge 2017). En effet, il existe des pressions au sein de l'holobionte limitant le nombre de souches bactériennes pouvant s'implanter dans le microbiote (liées en grande partie à l'immunité de l'hôte) (Zilber-Rosenberg and Rosenberg 2008). Enfin, d'un point de vue écologique, chaque hôte et ses bactéries peut être considéré comme un écosystème (Yeoman et al. 2011).

Dans la mesure où cette thèse a pour objet d'étude une espèce animale, le lézard *Podarcis sicula*, nous ne détaillerons pas le cas des espèces végétales, cependant, il est important de noter que les plantes sont elles aussi associées à des microorganismes et peuvent elles aussi être considérées comme des holobiontes (Vandenkoornhuyse et al. 2015). Afin de définir de façon plus détaillée notre système d'étude, il est important de s'intéresser aux communautés microbiennes présentes sur l'hôte, que l'on appelle le microbiote.

1.2.1 Présentation du microbiote

Un microbiote est l'ensemble des taxa présents dans une communauté microbienne associée à un hôte (Turnbaugh et al. 2007; Ursell et al. 2012). Les animaux abritent plusieurs microbiotes. Par exemple, chez de nombreux animaux, on trouve : un microbiote oral (Avila, Ojcius, and Yilmaz 2009; Schueller et al. 2017), un microbiote intestinal (Robles Alonso and Guarner 2013), un microbiote stomacal (Wu, Yang, and Peng 2014), un microbiote cutané (Rosenthal et al. 2011), un microbiote vaginal (pour les femelles) (Kim et al. 2009; Ma, Forney, and Ravel 2012), un microbiote du rumen chez les ruminants (Abecia et al. 2013; Carberry et al. 2012;

McCann, Wickersham, and Loor 2014; Morgavi et al. 2013; Yáñez-Ruiz, Abecia, and Newbold 2015).

Comme nous l'avons évoqué plus haut, le microbiote peut être étudié soit en utilisant une ou plusieurs régions d'un marqueur de l'ARN ribosomique, en général l'ARNr 16S, soit en utilisant des « shotgun reads » (courtes séquences d'ADN séquencées aléatoirement dans le microbiote) (Figure 4). Dans le cadre de mon doctorat, nous avons utilisé la région V4 de l'ARNr 16S.

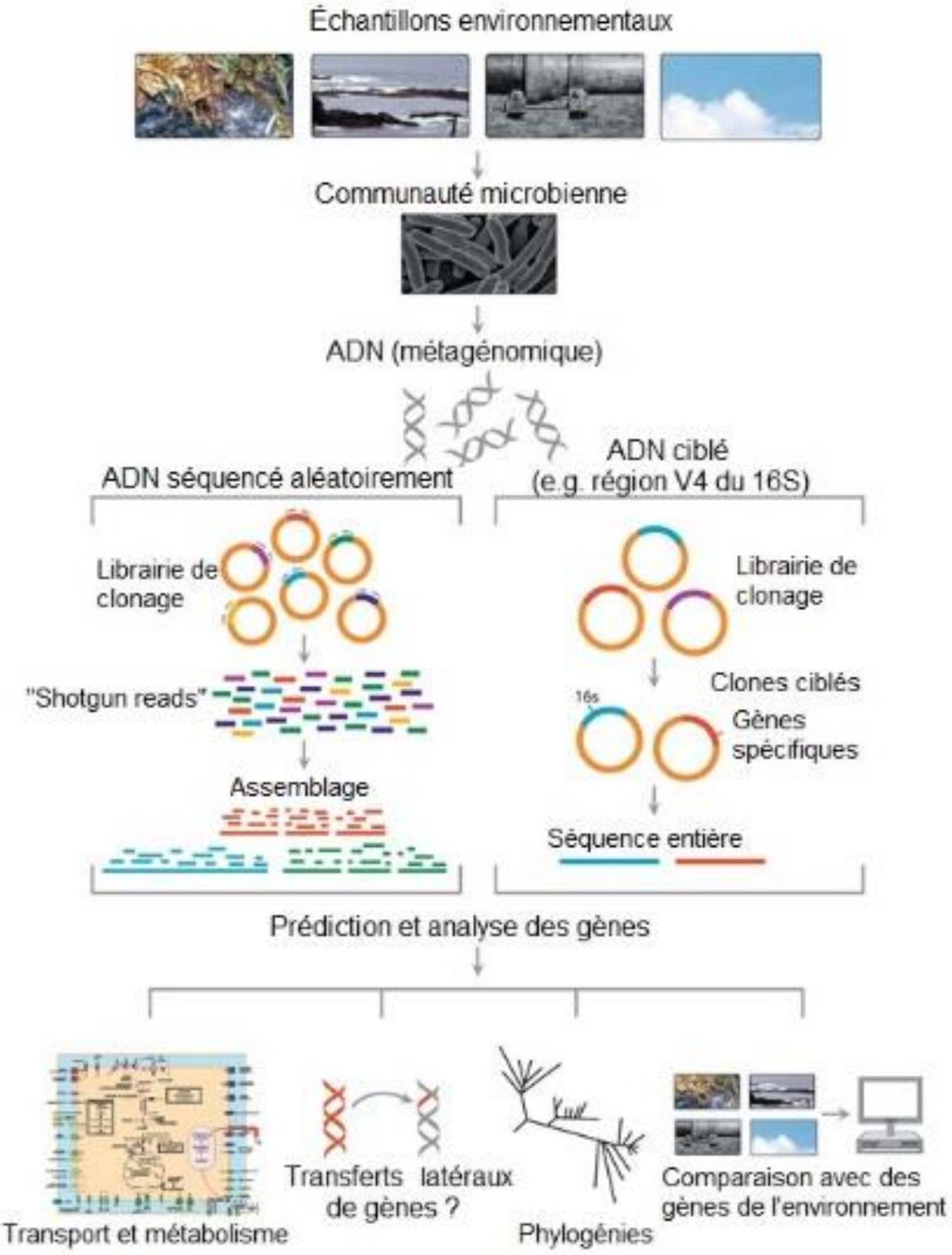


Figure 4 : Production de données de métagénomique (shotgun reads) et de données de « métabarcoding ».

(dans le cadre de cette étude : reads provenant de la région V4 de l'ARNr 16S). Adapté de http://squgenetics.pbworks.com/w/page/38604801/Introduction%2C%20definition%20and%20process%20of%20metagenomics

Grâce à ce marqueur, nous avons étudié le microbiote intestinal des *Podarcis sicula* (pris dans un sens large puisque les microbiotes intestinaux (oral, microbiote de l'iléon, du colon, de l'intestin distal, et des fèces (Yeoman et al. 2011) car le long du tube digestif les communautés ont des compositions différentes) . En effet, dans la mesure où il est difficile de ne récupérer qu'une seule communauté, soit parce que l'organisme étudié est petit (cas des lézards, pucerons, termites (Su et al. 2016)), soit parce qu'on récupère le microbiote dans les fèces pour éviter de sacrifier l'animal (humains, espèces protégées telles que le panda, mais aussi dans le cas où l'on souhaite étudier l'évolution du microbiote dans le temps), il est délibérément choisi ici (et dans de nombreuses études) de considérer l'ensemble de ces microbiotes intestinaux comme étant un seul microbiote intestinal (Sender, Fuchs, and Milo 2016).

1.2.2 Le microbiote intestinal et le régime alimentaire

Le microbiote intestinal a déjà été étudié dans un grand nombre d'espèces de vertébrés et d'invertébrés, principalement l'homme, notamment pour établir des liens entre le régime alimentaire, le microbiote et les différences morphologiques, comprendre comment cette communauté microbienne se structure. Les études de microbiote chez l'humain à partir de l'ARNr 16S, ont entre autre permis de définir la notion d'entérotypes (Arumugam et al. 2011; Knights et al. 2014), et de proposer un lien entre les entérotypes et le régime alimentaire (Lim et al. 2014; Wu et al. 2011). Les entérotypes sont des groupes d'hôtes construits à l'aide d'une méthode de « clustering » supervisée en se basant sur la composition taxonomique des microbiotes (Arumugam et al. 2011). La méthode de constitution des entérotypes est détaillée dans le chapitre 3. On retrouve d'ailleurs cette notion d'entérotypes chez d'autres lignées d'hôtes, telles que chez le bourdon, espèce présentant 2 entérotypes (Li et al. 2017), ou encore chez le chimpanzé, espèce présentant 3 entérotypes (Moeller et al. 2012). Cependant, cette notion d'entérotypes est assez controversée.

Si certains auteurs parlent de 3 entérotypes chez l'homme (Wu et al. 2011), d'autres n'en trouvent que deux (Wu et al. 2011) et enfin, certains contestent la notion même d'entérotypes, et proposent plutôt que différents hôtes d'une même espèce hébergent des communautés microbiennes dont la diversité n'est pas facilement partitionnable en types (Ian B. Jeffery et al. 2012). D'autres études soulignent également que les méthodes utilisées peuvent conduire à des erreurs dans le « clustering » (i.e. algorithme permettant de créer des groupes d'échantillons), notamment quand le nombre de genres trouvés dans les données est plus important que le nombre d'échantillons (Knights et al. 2017). Dans la continuité de ces études, nous avons recherché des entérotypes chez *Podarcis sicula*, afin de vérifier si ces groupes (dans l'hypothèse où on les détecterait) soient en lien (ou pas) avec le régime alimentaire.

Une autre façon d'étudier la structure des communautés microbiennes consiste à quantifier et comparer leur diversité. Pour cela, deux types d'études sont couramment utilisées : la diversité alpha, qui correspond à la diversité d'un écosystème, c'est à dire combien d'espèces (ou phyla, genres,...) différentes sont présentes au sein de la communauté microbienne étudiée, et la bêta diversité, qui correspond à la comparaison de la diversité de deux écosystèmes, ou à la comparaison de la diversité d'un même écosystème à deux moments différents (Hamady, Lozupone, and Knight 2010). L'étude comparative du microbiote entre une population d'humains obèses et d'humains sains a ainsi permis de montrer que des différences morphologiques peuvent être corrélées à des différences d'abondances faibles (Everard et al. 2013) (moins de 5% pour les *Akkermansia muciniphila* par exemple). Un changement morphologique peut donc être associé à des variations chez certaines espèces bactériennes, sans que pour autant cela débouche sur des entérotypes. Nous avons donc également recherché des espèces bactériennes dont l'abondance pourrait être différente entre les individus insectivores et omnivores, indépendamment de la notion d'entérotypes.

Enfin, deux types de résultats sont communément observés lorsque l'on étudie le lien entre le régime alimentaire et le microbiote ciblé. Un première ensemble d'études a pour résultats de grandes différences dans le microbiote associés à un changement de régime alimentaire, touchant des genres majoritaires (David et al. 2014; Sonnenburg et al. 2016). Cependant, les résultats de comparaison d'humains à régimes alimentaires très différents en provenance de continents différents sont à

nuancer, car ces modifications peuvent aussi découler de contraintes imposées par la génétique de l'hôte, qui ne sont pas les mêmes d'une population à une autre, de différences en terme d'hygiène de vie (notamment avec l'utilisation très importante d'antibactériens dans les pays occidentaux), en termes de parasites, et en termes d'environnement. Dans l'idéal, il faudrait pouvoir comparer une population contenant des individus avec des régimes alimentaires différents. Ce type de comparaison est effectué chez les souris, et se traduit effectivement par des changements importants du microbiote associé au changement de régime alimentaire (Hildebrandt et al. 2009; Zhang et al. 2012).

Un autre ensemble d'études indique néanmoins que deux régimes alimentaires différents ne sont associés qu'à de petites différences au niveau du microbiote, et que la variabilité interindividuelle a plus d'impact que le régime alimentaire sur le microbiote. C'est un résultat trouvé par exemple chez l'humain par Lozupone (Lozupone et al. 2012), mais aussi chez le panda (Y. Li et al. 2015; Wei, Hu, et al. 2015; Xue et al. 2015).

Dans le cadre de cette thèse la question suivante se pose : le changement de régime alimentaire des *Podarcis sicula* est-il associé à peu ou à beaucoup de différences au sein du microbiote ?

1.2.3 Le microbiome : fonctions de la communauté microbienne

Si le microbiote constitue la moitié d'un holobionte chez l'humain en nombre de cellules (soit environ $3.9x10^{13}$ bactéries (Sender, Fuchs, and Milo 2016)), le nombre de gènes microbiens par rapport au nombre de gènes humains est considérablement plus important. Regarder « qui est là » au sein d'une communauté bactérienne est très informatif et indispensable à la compréhension de son fonctionnement. Pour cela il faut analyser notamment les fonctions des gènes de la communauté bactérienne. Précisément, le microbiome correspond au contenu génétique d'un microbiote (Yeoman et al. 2011). Des études préliminaires du microbiome ont été effectuées sur plusieurs microbiomes dont les microbiomes intestinaux, vaginaux, oraux, aussi bien chez les mammifères dont l'humain (Avila, Ojcius, and Yilmaz 2009; Dewhirst et al.

2010; Kim et al. 2009; Ma, Forney, and Ravel 2012; Medina-Colorado et al. n.d.; Schueller et al. 2017), chez les reptiles (Costello et al. 2010), ou encore chez les insectes, dont les hyménoptères (Li et al. 2017; Suen et al. 2010).

Pour les raisons évoquées dans la partie précédente, nous nous intéressons plus spécifiquement au microbiome intestinal. Le microbiome intestinal correspond au contenu génétique provenant de micro-organismes présents dans l'intestin, c'est-à-dire, aux gènes microbiens présents dans l'intestin. Des études ont montré des différences au niveau des fonctions de gènes pour différents régimes alimentaires chez l'être humain (David et al. 2014). Dans la mesure où nous nous intéressons à l'impact du changement de régime alimentaire des *Podarcis sicula* sur leur microbiome intestinal, l'une des questions que l'on se pose est la suivante : les gènes présents dans le microbiome intestinal des lézards insectivores sont-ils les mêmes que ceux présents dans le microbiome intestinal des lézards omnivores ?

Répondre à ces questions nécessite d'avoir des « reads » obtenus par séquençage non ciblé (données de métagénomique) (Figure 4). L'une des études couramment menée se base sur la quantité de reads par catégorie COG (Clusters of Orthologous Groups) (Gill et al. 2006). Il existe en tout 25 catégories COGs regroupées dans les différentes grandes classes de fonctions : une classe de catégories COGs relative aux processus cellulaires et de signalisation, une classe relative au stockage d'informations et aux processus informationnels, une classe relative au métabolisme, et une classe de catégories COGs relatives aux fonctions peu connues. Les catégories COGs sont détaillées dans le chapitre 4.

Dans cette thèse nous avons plus précisément étudié les abondances des différentes catégories COG en fonction du régime alimentaire, ainsi que les différences d'abondances d'expression des enzymes présentes dans les microbiomes en fonction du régime alimentaire des lézards.

Il semble communément accepté qu'une différence de régime alimentaire se traduise par des changements spécifiques concernant l'abondance de certaines enzymes impliquées dans des voies métaboliques spécifiques (David et al. 2014; Wei, Hu, et al. 2015; Zhu et al. 2011). En fonction des études, ces différences sont plus ou moins importantes. Dans un premier temps, nous nous sommes basés sur la littérature, en étudiant l'abondance d'enzymes réputées importantes pour la digestion

des végétaux (notamment de la cellulose) et pour la digestion des insectes (notamment de la chitine). Puis nous avons étudié de façon moins ciblée les différences d'abondances en enzymes des différentes voies métaboliques entre les microbiomes de lézards insectivores et omnivores.

1.3 Objectifs de la thèse

Le premier objectif de mon travail de thèse a été la caractérisation du microbiote intestinal de *Podarcis sicula*, en termes de diversité taxonomique, de présence et d'abondance des espèces microbiennes présentes dans les microbiotes de ces lézards (chapitre 3). L'objectif suivant a été d'étudier l'impact de certaines caractéristiques (régime alimentaire, genre, année et saison d'échantillonnage, insularité, localisation) de ce lézard sur son microbiote et microbiome intestinal (chapitre 3). Cela consiste en la comparaison de la diversité microbienne et génétique des microbiomes de lézards, mais aussi en l'identification des taxa et des gènes présents dans le microbiome (chapitre 4). Au-delà de réponses aux questions biologiques soulevées précédemment, ces objectifs ont permis de proposer de nouvelles méthodes (réseaux de reads, et graphes bipartis) relevant de la théorie des graphes afin d'étudier la diversité des microbiomes et les modes de transmission des microbes (chapitre 5). En effet, il est important de noter que la représentation de l'histoire évolutive des organismes par un unique arbre des espèces n'est absolument pas représentative de la réalité dans la mesure où la construction de cet arbre ne se base que sur une trentaine de gènes (soit environ 1% de génome procaryote) (Ciccarelli et al. 2006). L'étude de l'évolution à l'aide de réseaux (Figure 5) permet de se placer à l'échelle des holobiontes (Corel et al. 2016) et non plus uniquement à celle de l'organisme.

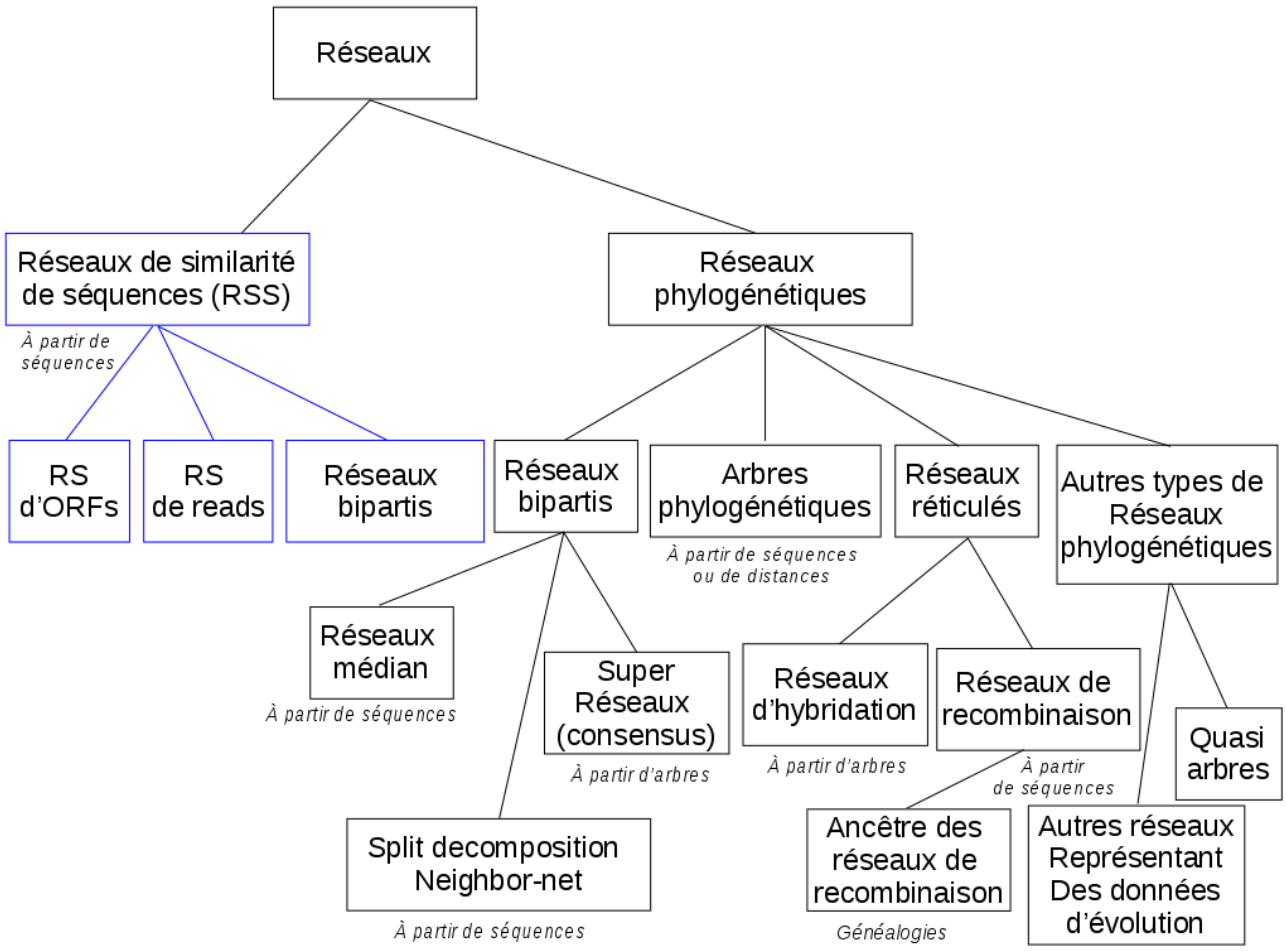


Figure 5 : Description de différents types de réseaux en biologie.

adaptation de la figure d'Huson et Kloepper (2005)(Huson and Bryant 2006; Huson and Kloepper 2005).

Sous la dénomination "réseaux", sont ici regroupés les réseaux phylogénétiques et les réseaux de similarité de séquences (RSS). Le terme de réseau phylogénétique regroupe des concepts différents, dont les arbres phylogénétiques, les graphes bipartis, les réseaux réticulés, concept qui contient les réseaux de type hybridation et recombinaison, ainsi que d'autres types de réseaux tels que les quasi-arbres aussi appelés arbres augmentés. Les réseaux de recombinaison sont très proches des réseaux de recombinaison d'ancêtres utilisés dans les études de population. Le concept de RSS fait partie des développements de l'équipe AIRE, et sont en partie appliqués dans cette thèse.

2. De la diversité des méthodes à la standardisation des analyses

Ce chapitre est né des réflexions menées au sujet des études de microbiomes et de microbiotes : existe-t-il des analyses standards ? Que font les autres équipes de recherche ? Pourrait-on enrichir le jeu de données des microbiomes intestinaux de lézards avec des jeux de données de microbiomes intestinaux de différents hôtes, séquencés par différentes équipes, afin de trouver des familles de gènes spécifiques de la digestion de plantes ? Ces questions de comparaison d'études entre elles nous ont montré qu'il n'était pas si simple en métagénomique, de réunir plusieurs jeux de données en un seul, et l'objet de ce chapitre est de détailler cette opinion.

2.1 De la diversité des méthodes en métagénomique

La métagénomique est une discipline récente (Escobar-zepeda, León, and Sanchez-flores 2015; Thomas, Gilbert, and Meyer 2012) dont la dénomination a été proposée en 1998 par Handelsman (Escobar-zepeda et al. 2015). Cette science résulte des progrès en matière de séquençage (de l'invention de la technique de séquençage de l'ADN par Sanger en 1977 aux premières analyses de communautés microbiennes se basant sur de l'ARNr 16S en 1990) (Escobar-zepeda et al. 2015) et des progrès en matière de traitement des données massives, appelées « Big Data » (terme apparu en 1997). En effet, le nombre de publications dans le domaine des « Big data » a explosé à partir de(Liu et al. 2016; Mokane Bouzeghoub 2017).

2.1.1 Qu'est-ce que la métagénomique ?

La métagénomique est une discipline appartenant au champ de la collecte de données (« data gathering ») (Krohs 2012). Les données collectées sont des fragments d'ADN recueillis dans un environnement donné (Thomas et al. 2012). Il peut s'agir aussi bien de microbiomes animaux (Gomez et al. 2015; Hong et al. 2011; Kohl et al. 2013; Martinson et al. 2011; McCann, Wickersham, and Loor 2014; Moeller et al. 2012, 2015; Su et al. 2016; Wei, Wang, and Wu 2015; Xu et al. 2016; Yáñez-Ruiz, Abecia, and Newbold 2015; Zeng et al. 2015; Zheng et al. 2016; Zhu et al. 2011) (intestinaux, oraux, vaginaux, etc) (Avila, Ojcius, and Yilmaz 2009; Le Chatelier et al. 2013; Gill et al. 2006; Huttenhower and Human Microbiome Project Consortium 2012; Kim et al. 2009; Ma, Forney, and Ravel 2012; Medina-Colorado et al. n.d.; Prado-Irwin et al. 2017; Schueller et al. 2017; Turnbaugh, Ridaura, et al. 2009; Turnbaugh, Hamady,

et al. 2009; Walter and Ley 2011; Yatsunenko et al. 2012) que de microbiomes végétaux (Hartman et al. 2017; Turner, James, and Poole 2013; Vandenkoornhuyse et al. 2015), ou du microbiome d'un environnement (exemple du microbiome marin avec TARA (Pesant et al. 2015; Sunagawa, Coelho, Chaffron, Kultima, Labadie, Salazar, Djahanschiri, Zeller, Mende, Alberti, Cornejo-Castillo, Costea, Cruaud, D'Ovidio, et al. 2015; de Vargas et al. 2015)). Suite à cette collecte, la seconde étape de la métagénomique est l'analyse de ces jeux de données (Escobar-zepeda et al. 2015).

2.1.2 Difficultés engendrées par la diversité des méthodes en métagénomique

Avant l'apparition d'un nouveau champ disciplinaire, il est attendu de voir naître une myriade de nouvelles méthodes et approches différentes, dont certaines deviendront éventuellement les questions posées par la discipline naissante (Krohs 2012; Sydow, Schreyögg, and Koch 2005). Si cette créativité est nécessaire pour l'émergence d'un nouveau champ et de nouvelles idées, l'absence de standardisation des méthodes rend les résultats obtenus par les scientifiques difficilement comparables (voire incomparables) d'une étude à l'autre.

2.1.3 La production des données métagénomiques

L'une des premières difficultés pour comparer des métagénomes est le fait qu'il existe différentes méthodes d'acquisition des données (cf. les différentes méthodes de séquençage et d'extraction de l'ADN) (Burke, Kjelleberg, and Thomas 2009; Delmont et al. 2011; Liu et al. 2012; Venter et al. 2004) dont les résultats ne sont comparables ni en termes de nombre et de longueur de reads, ni en termes de qualité (cf. table 1 de l'article (Escobar-zepeda et al. 2015)). Dans le chapitre « Analyse de la diversité microbienne : de la difficulté (paradoxale) de voir large en métagénomique » du livre *Biodiversité et Evolution* présenté en partie 2.5, la Figure 1 illustre que deux séquençages sur une même plateforme utilisant deux kits de séquençage proches (illumina 2x300paires de bases en 2014 et illumina 2x250 paires de bases en 2015) ne donnent pas des résultats comparables entre eux. L'absence de standardisation des méthodes d'acquisition des données est donc un frein à la comparaison des

études entre elles.

Un avantage d'une démarche pluraliste en termes de méthodes est la mise en évidence du biais de chaque méthode grâce à la comparaison des résultats des méthodes les uns avec les autres. Ainsi, concernant l'assemblage, les performances de chaque assembleur sont différentes. Certains sont plus précis, d'autres plus rapides, d'autres moins coûteux en mémoire (i.e. Meta-IDBA requiert moins de mémoire que MetaVelvet), certains permettent de prédire plus de contigs (concaténation de reads en plus longue séquence, en se basant sur la similarité entre les reads) que d'autres (i.e. MetaVelvet permet de prédire 66 241 gènes alors que Meta-IDBA n'en prédit que 62 833 sur le même jeu de données) (Namiki et al. 2012). Selon les critères d'évaluation utilisés pour déterminer quel est le meilleur assembleur, le résultat ne sera pas le même (Bradnam et al. 2013). D'un jeu de données à l'autre, l'assembleur le plus précis n'est pas toujours le même (Treangen et al. 2013). Ainsi selon les ressources computationnelles disponibles et selon le jeu de données à étudier, la méthode choisie ne sera pas systématiquement la même.

Du fait de l'évolution des techniques, cette pluralité des méthodes semble inévitable. Tout d'abord, une partie de la métagénomique est externalisée. En effet, le séquençage est souvent réalisé par des laboratoires spécifiques, qui ne sont pas les mêmes que ceux d'analyse. Cela a donc un coût, et la "kittification" du séquençage ainsi que son évolution dépend du marché économique du séquençage.

Par ailleurs, l'évolution des modèles et des champs disciplinaires dépend aussi de cette pluralité des méthodes. En effet, un modèle est une simplification de la réalité. Une des questions que l'on se pose souvent est donc : à quel point le modèle simplifie la réalité ? Peut-on créer un modèle plus proche de la réalité que le précédent ? Par exemple, pour répondre à la question : « existe-t-il une structure dans le microbiote permettant de trouver des groupes d'individus ? » (que l'on peut résumer par « existe-t-il des entérotypes ? ») deux méthodes d'analyses sont utilisées. Un premier modèle se base sur une discrétisation des données, et l'on cherche à trouver les meilleurs groupes possibles en choisissant le nombre optimal de groupes (Arumugam et al. 2011). Un autre modèle consiste à regarder les abondances relatives des espèces présentes dans les microbiomes, et à appliquer un clustering non supervisé, tel qu'un clustering hiérarchique, afin de voir si l'on trouve des groupes. Cette seconde méthode permet de trouver que la structure de la population ne peut se regrouper de façon discrétisée, mais plutôt que la structure des communautés microbiennes étudiées est

représentable sous forme d'un gradient (Ian B Jeffery et al. 2012). On a donc pour une même question, deux modèles différents : un plus synthétique (les entérotypes) et un plus précis (les gradients) grâce à cette diversité des méthodes.

2.1.4 Absence de standardisation des méthodes en métagénomique

Une seconde difficulté rencontrée est l'absence de standardisation dans l'analyse des métagénomes obtenus. En effet, si en métagénomique des pipelines d'analyse commencent à émerger, le choix des outils pour réaliser chacune des étapes reste encore très vaste (Figure 6).

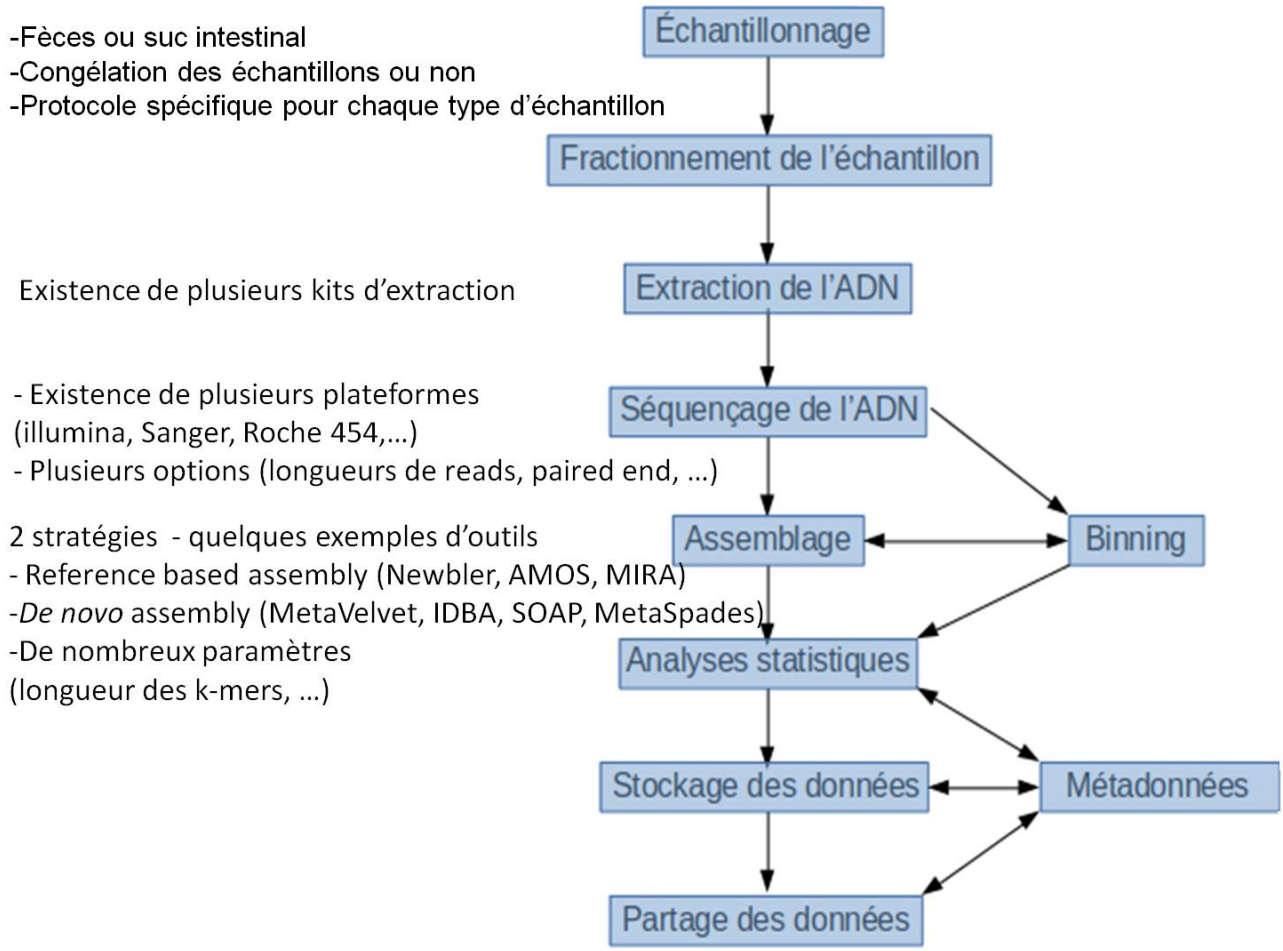


Figure 6 : Pipeline d'analyse en métagénomique et diversité des méthodes.

Certaines étapes, comme le binning (classement des objets à analyser dans diverses catégories) sont optionnelles. Le binning peut être placé avant ou après l'assemblage. Les analyses statistiques, quant à elles, sont variables et décrites dans

la partie 2.3.

Chaque étape présente donc une multitude d'outils, pour lesquels les paramètres diffèrent, ce qui illustre bien l'absence de standardisation des méthodes.

2.2 De la diversité des données en métagénomique et en analyse de données microbiennes

Cette nouvelle discipline qu'est la métagénomique a donné naissance à de nouveaux jeux de données, les métagénomes. Il existe deux types de métagénomes (cf. Figure 4). En premier lieu, les métagénomes dits « ciblés » (ou encore « amplicon sequencing ») (1) séquencent dans l'environnement un marqueur précis, en général une partie de l'ARN ribosomique. Suite à cela, les microbes présents dans l'environnement sont prédits et des gènes sont prédits pour chaque microbe. Le terme métagénomique employé ici est considéré comme abusif dans plusieurs publications, et il est plutôt recommandé d'utiliser le terme « metaprofiling » (Escobar-zepeda et al. 2015) ou « metabarcoding » (Nagaraj et al. 2017).

L'autre type de données produites sont les métagénomes « non ciblés » (aussi appelés « shotgun sequencing »). Ce type de jeu de données est constitué en séquençant aléatoirement des fragments d'ADN (appelés « reads »). Les reads sont ensuite assemblés en « contigs », dans le but de reconstruire au maximum les génomes des microbes présents dans l'environnement.

Ces deux types de jeux de données sont traités à l'aide d'outils d'analyses différents, visant à répondre aux questions « quels sont les microbes présents ? » et « quels sont les gènes présents ?». Si les méthodes d'analyse des données ciblées tendent à converger vers un sentier de dépendance (*cf* 2.5), les méthodes d'analyse des données non ciblées sont, elles, encore très diverses.

2.3 Etude des microbiomes intestinaux de *Podarcis sicula* et sentier de dépendance

Nous avons adapté une définition de sentier de dépendance fournie par Sydöw pour appliquer cette notion en sciences. Selon cette nouvelle définition, un sentier de dépendance correspond à une convergence des méthodes d'analyses relatives à un

type de données (par exemple, les métagénomes), dans le but de répondre à des questions bien identifiées. L'étude du microbiote en utilisant le marqueur 16S est engagé sur un sentier de dépendance. En effet, il existe des pipelines d'étude du microbiote qui sont standards. Par exemple, on utilise des Operational Taxonomic Units (OTUs) (Sneath and Sokal 1962) pour regrouper les séquences d'ARNr 16S. Les séquences sont regroupées ainsi : si une séquence présente 97 % de similitude avec au moins une autre séquence d'une OTU, alors cette séquence appartient elle aussi à l'OTU. Le seuil de similitude peut varier, cependant le chemin de dépendance emprunté par les études de microbiote le choisit préférentiellement. Cela permet de considérer qu'une OTU correspond à une espèce. En effet, une des définitions de l'espèce est que deux espèces sont distinctes l'une de l'autre si leurs ARNr 16S présentent moins de 97 % d'identité (Konstantinidis and Tiedje 2005). Bien que cette définition soit controversée, l'utilisation de ce seuil nous permet de nous inscrire dans un modèle biologique pré-existant. Il existe différents outils pour construire les OTUs, dont QIIME (Caporaso et al. 2010; Kuczynski, Stombaugh, Walters, González, J. Gregory Caporaso, et al. 2012) et MOTHUR (Schloss et al. 2009).

A partir de ces OTUs, la diversité des métagénomes est étudiée à l'aide d'outils tels que les indices de Shannon, Simpson, et Chao1 (Shannon 1948; Whittaker 1960, 1972). Les études de diversité utilisent un seul ou plusieurs de ces indices, qui sont détaillés dans le chapitre 3. Ensuite, on s'intéresse à la beta diversité. Là encore, plusieurs outils d'analyses sont disponibles (PcoA, NMDS, …)(Borcard, Gillet, and Legendre 2011) et sont détaillés dans le chapitre 3. Enfin, les OTUs sont annotées afin de pouvoir déterminer quelles sont les espèces présentes dans les métagénomes (à l'aide de BLAST (BLAST n.d.; Camacho et al. 2009; Johnson et al. 2008), de BLAT (Kent 2002), de CD-HIT (Fu et al. 2012),…).

En revanche, les études de microbiome ne convergent pas autant vers un sentier de dépendance. En effet, il existe de multiples façons de les étudier. Si l'on se réfère à la Figure 1 par exemple, on peut constater que la plupart du temps, l'assemblage des métagénomes en contigs (concaténation de reads en plus longue séquence, en se basant sur la similarité entre les reads) est une étape indispensable à l'analyse. Après l'assemblage en contigs, des ORFs (Open Reading Frames) sont prédites à l'aide d'outils tels que MetaGeneMark (Ismail, Ye, and Tang 2014; Zhu, Lomsadze, and Borodovsky 2010) et MetaGeneAnnotator (Noguchi, Taniguchi, and

Itoh 2008). Suite à ces prédictions d'ORFs, il est possible d'annoter taxonomiquement et/ou fonctionnellement les ORFs, mais aussi de construire des Réseaux de Similarités de Séquences (RSS), comme présenté dans le chapitre 5. Ce second type d'analyses permet de choisir d'étudier des processus très différents, affectables au microbiome. Par exemple, il est possible d'étudier la structure de la communauté et de faire apparaître différentes figures de transmission, dont les transferts latéraux de gènes (Doolittle 1999; Zhaxybayeva and Doolittle 2011) (Roberts and Mullany 2010) (Figure 7).

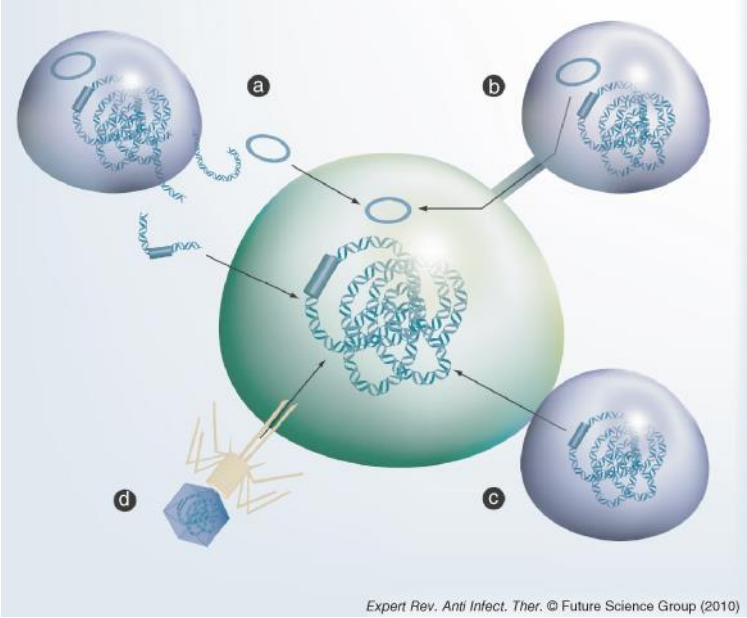


Figure 7 : Transfert latéral de gènes et d'ADN.

Les cellules bactériennes sont représentées par les ovales verts et violets. L'ADN est représenté par l'hélice. Les transposons et plasmides sont des rectangles et cercles bleus. Les flèches montrent la direction du transfert d'ADN. (a) Transformation : la cellule donneuse (en haut à gauche) a subi une lyse et l'ADN a été libéré dans l'environnement. Cet ADN peut être réceptionné par une bactérie et incorporé dans son génome. (b) Conjugaison de plasmides. (c) Conjugaison de transposons via un pore d'accouplement. (d) transduction par l'intermédiaire d'une injection d'ADN par un bactériophage (Roberts and Mullany 2010).

Cependant, il est aussi possible d'étudier les microbiomes directement à partir des reads, en les annotant taxonomiquement et/ou fonctionnellement, mais aussi en construisant des réseaux de similarités de reads afin d'étudier la diversité présente dans les microbiomes (*cf* chapitre 5).

Il existe donc au moins 2 stratégies pour l'étude de microbiomes (assembler ou

ne pas assembler les reads), comprenant au moins 3 types d'analyses chacune, avec plusieurs méthodes pour chaque type d'analyse. Il n'y a donc pas encore de sentier de dépendance pour l'étude de microbiomes.

2.4 Analyse de la diversité microbienne : de la difficulté (paradoxale) de voir large en métagénomique (chapitre de livre n°1).

Nous présentons ici le chapitre que nous avons rédigé pour l'ouvrage 'Evolution et Biodiversité' édité par Philippe Grandcolas et Marie-Christine Maurel aux éditions ISTE. L'ouvrage étant sous presse à l'heure actuelle, le formatage n'est pas définitif.

Analyse de la diversité microbienne : de la difficulté (paradoxale) de voir large en métagénomique

Vigliotti, Lopez, Bapteste

Introduction

La richesse des communautés biologiques est prodigieuse. La complexité de l'organisation de ces communautés comme la diversité de leur composition (en gènes, en fonctions, en organismes, en taxa) posent des défis de taille aux scientifiques qui cherchent à développer des méthodes visant à produire des descriptions et des connaissances aussi réalistes que possible du monde vivant. Pour comprendre les causes de la biodiversité, il parait notamment essentiel de pouvoir voir large, c'est-à-dire d'être capable d'énumérer les principaux acteurs (à défaut de tous les acteurs), les principales relations entre eux (à défaut de toutes leurs relations) et les principaux processus (à défaut de tous les processus) qui structurent les communautés biologiques, puis de comparer ces organisations entre elles dans l'espace et dans le temps. Cet objectif ambitieux, d'inspiration holiste, peut-il être tenu ? Une chose est sûre : les données moléculaires n'ont jamais été aussi abondantes, ni aisées à obtenir qu'aujourd'hui. Pour ce qui concerne les communautés microbiennes, la métagénomique fournit une mine de données colossale: les séquences des gènes, des ARNS, voire des génomes des procaryotes, des protistes et de leurs éléments mobiles (virus, plasmides) présents dans un environnement. Par exemple, le Human Microbiome Project (HMP) a réalisé 1265 projets de métagénomique entre 2011 et 2014, et produit ainsi plus de 80 millions de séquences de gènes. (cf http://hmpdacc.org/catalog/grid.php?dataset=metagenomic). Le projet Tara a étudié 210 écosystèmes océaniques provenant de 20 zones biogéographiques [PES 15], et fournit plus de 111 millions de séquences de gènes [SUN 15]. Et ces projets emblématiques sont loin d'être les seuls, l'EBI recense

ainsi 67 375 échantillons provenant 1073 projets de métagénomique (cf https://www.ebi.ac.uk/metagenomics/search).

Ces données sont obtenues en plusieurs étapes. Premièrement, il y a une phase d'échantillonnage (de l'eau, de la terre, des fèces, etc. sont récoltés). Ensuite, ces échantillons sont filtrés, notamment par fraction de taille, ce qui permet de séparer les micro-organismes de tailles différentes et les virus, etc. Puis, l'ADN est isolé, amplifié et séquencé. Enfin, les données moléculaires sont analysées par des pipelines bioinformatiques.

La métagénomique s'inscrit donc, par sa démarche, dans un mode de science émergent : la collecte de données (« data gathering ») au moyen de méthodes déjà développées en dehors des laboratoires qui les utilisent [KRO 12]. Par exemple, les premières étapes de l'étude des microbiomes intestinaux de lézards consistent, une fois les lézards capturés et l'ADN contenu dans leurs estomacs isolé, à choisir une méthode de séquençage et à identifier un prestataire de service qui la mettra en œuvre une fois les échantillons reçus. Le philosophe des sciences U. Krohs qualifie ce type de procédure de production des données, à la fois massive, simple et relativement prémâchée, de « science confortable » (« convenience science ») ou « science automatisée » [KRO 12]. Dans la mesure où le séquencage a un coût abordable (l'obtention de 104 millions de paires de reads par 2*300bp illumina revenait à 17 500$ en 2015), de très nombreux laboratoires dans le monde collectent ainsi leurs données, et la taille les bases de données de métagénomique, publiquement disponibles, comme MGRAST (http://metagenomics.anl.gov), explose. On pourrait donc naïvement supposer que la métagénomique offre une vue très large des communautés microbiennes. Autrement dit, on pourrait espérer qu'il soit possible de produire des généralisations, des modèles, voire des théories sur la composition, l'organisation, la dynamique et l'évolution des communautés microbiennes, en comparant toutes ces données, bien qu'aujourd'hui aucune base de données ne centralise la majorité des séquences obtenues. Pourtant, répondre par l'affirmative reviendrait à minorer un paradoxe au cœur de ces approches volontiers inclusives : leur étendue et leur profondeur de vue sont limitées par des facteurs épistémiques et pratiques, que ce chapitre exposera, en s'appuyant largement sur notre expérience de l'analyse de microbiomes (i.e. des gènes des communautés microbiennes), et de microbiotes (i.e. des taxa des communautés microbiennes) intestinaux de lézards de l'espèce *Podarcis sicula*, ayant pour particularité d'avoir subi des variations de régime alimentaires. En effet, en 35 ans, une population de ces lézards originellement insectivore, est devenue omnivore en intégrant à son régime alimentaire près de 80 % de végétaux [HER 08].

La comparaison des jeux de données métagénomique est difficile

Les jeux de données métagénomiques (ou microbiomes) sont à la fois nombreux et très divers, ce qui rend leur comparaison difficile. Cette diversité des jeux de données existe sur plusieurs plans. Tout d'abord, il faut distinguer les métagénomes environnementaux (e.g. TARA oceans [PES 15] [SUN 15] [VAR 15]), de ceux associés à des hôtes. Si l'on s'intéresse plus spécifiquement aux métagénomes associés à des hôtes, la littérature témoigne d'une diversité d'études portant sur des microbiomes associés à une grande variété d'hôtes : microbiomes associés à l'homme [WAL 11] [HMPC 12], à la souris [XU 16] [ZHE 16], au gorille [GOM 15] [MOE 15], au chimpanzé [MOE 12], à la grenouille [KOH 13], au panda [WEI 15] [ZHU 11], aux termites [SU 16], à la vache [MCC 14] [YAN 15], à l'iguane [HON 11], au lapin [ZEN 15], aux abeilles [MAR 11], etc. L'association de métagénomes à des hôtes génère là aussi une diversité de jeux de données dépendant du microbiome séquencé [HMPC 12] : buccal [SCH 17], cutané [PRA 17], intestinal [WAL 11] [LEC 13][TUR 09a] [YAT 12] [GIL 06] [TUR 09b], vaginal [MA 12] [MED 17],.... On pourrait même soutenir, qu'en partie, beaucoup de ces microbiomes sont incommensurables. Cela tient tout d'abord au fait que les modes d'acquisition des séquences sont très hétérogènes. Même au sein de la « science confortable », il existe une diversité de méthodes de séquencage. Parmi les principales technologies de séquençage employées, on compte le SOLiD, l'Ion Torrent PGM (Life Sciences), le HiSeq 2000, MiSeq (Illumina), et le 454 (Roche) [LIU 12]. Ces différentes technologies ont des caractéristiques différentes, notamment en termes de longueur de reads produits (par exemple, en 2012, le 454 GS LFX produisait des reads de 700 paires de bases alors que le HiSeq 2000 produisait des reads de 50 à 100 paires de bases), en termes de précision (99.9% pour le 454 contre 98% pour le Hiseq), en termes de coût temporel et monétaire [LIU 12], et en terme de biais de séquençage (propre à chaque technologie), ce qui engendre des jeux de données difficilement comparables entre eux. En effet, ceci se traduit par la constitution de jeux de données présentant des quantités, qualités, et longueurs de séquences variables, qui capturent d'ailleurs des proportions différentes de l'ADN environnemental (ce qui se mesure par la couverture de séquençage des microbiomes et des courbes de saturations), et ont donc des représentativités inégales de ces communautés. Ces biais expérimentaux limitent les comparaisons de microbiomes (ainsi que celles de microbiotes), même pour des systèmes proches. L'effet de l'année de séquençage était par exemple particulièrement visible dans nos études. En 2014 et en 2015, nous avions obtenus séparément environ 1 million de séquences de la région V4 d'un marqueur de l'ADN (l'ARNr 16S) en 2014 et 1.7 million de séquences de la région V4 en 2015, dans le but de caractériser la composition des communautés microbiennes de la valve caecale des *P. sicula*. La première année, ces données avaient été produites par une première méthode de séquençage (de l'illumina 2*300 paires de bases), la deuxième année, une méthode différente avait été utilisée (de

l'illumina 2*250 paires de bases). L'annotation taxonomique de ces séquences permet de distinguer deux grands types de communautés microbiennes au sein de ces lézards. Après vérification, nous avons pu conclure que ce résultat fascinant ne reflétait pas un aspect essentiel de la biologie des lézards, mais un biais de séquençage. Les microbiomes des lézards séquencés par la même méthode se ressemblaient plus entre eux qu'ils ne ressemblaient aux microbiomes des autres lézards (*cf* Figure 1).

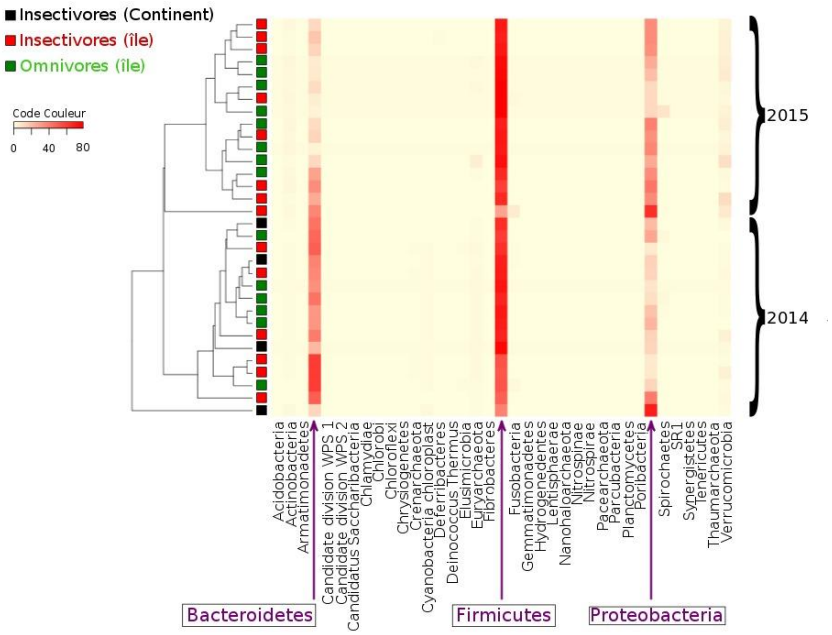


Figure 1 : Impact de la méthode de séquençage sur la composition taxonomique du microbiote de lézards.

Autre biais notable, le nombre d'individus séquencés peut avoir un impact sur les résultats obtenus. Les analyses d'entérotypes humains (i.e. communautés microbiennes singulières dont la composition taxonomique est proche) ont ainsi révélé des résultats différents selon le nombre d'individus échantillonnés. L'étude de 39 échantillons provenant de différents pays européens, d'Amérique et du Japon [ARU 11] avait conclu à l'existence de 3 entérotypes, là où une étude ultérieure, menée sur 98 échantillons, n'en proposait plus que 2 [WU 11]. Cette hétérogénéité complique l'interprétation des analyses de métagénomique comparative, car celle-ci supposerait de pouvoir distinguer les signaux qui reflètent des variations des méthodes de séquençages et ceux qui reflètent des propriétés biologiques des

microbiomes. Autrement dit, l'abondance de jeux de données n'offre pas immédiatement aux métagénomiciens les éléments qui permettraient de réaliser facilement des synthèses.

En outre, la nature des molécules séquencées réduit significativement le périmètre de la biodiversité sur lequel la métagénomique porte effectivement son regard. Un très grand nombre de jeux de données ciblent en effet des régions particulières d'un marqueur particulier, l'ARNr 16S (pour les procaryotes) et l'ARNr 18S (pour les eucaryotes). Le choix de ces régions et de ces molécules permet de construire des amorces (sur la base des séquences connues) pour aller identifier leurs homologues dans les environnements. Le choix de ce marqueur s'explique par le fait que les chercheurs aimeraient pouvoir associer les formes de 16S (ou de 18S) qui divergent par plus de 3% entre elles à des espèces distinctes. Dans ce contexte, les collections de ces molécules peuvent notamment servir à inférer, en première approximation, le microbiote, c'est-à-dire la composition en taxa du microbiome. Cette démarche laisse cependant de côté de nombreux organismes et ce de façon de plus en plus notoire. Ainsi, les bactéries CPR, ultra-petites et dotées d'une biologie fascinante, bien qu'ubiquitaires dans l'environnement, n'avaient pas été découvertes car leur 16S était atypique [BRO 15]. De même, cette approche laisse de côté l'immense diversité virale ; et les très nombreux microbes rares sont faiblement détectés du seul fait de leur faible abondance individuelle, même si collectivement ils représentent des portions considérables des communautés. En effet, au sein d'une communauté microbienne co-existent une minorité d'espèces abondamment présentes, et une grande majorité d'espèces rares [JOU 17] [NEM 11]. Ces microbes rares constituent ce qu'on appelle la " biosphère rare ". Plus précisément, la biosphère rare correspond à l'ensemble des espèces microbiennes dont l'abondance dans l'écosystème considéré n'excède pas 0.1% de la communauté microbienne totale [WAN 17]. Si l'on considère un litre d'eau, la biosphère rare présente dans cet écosystème se compte en centaines, voire milliers, d'espèces de bactéries et d'archées [WAN 17]. Les métagénomiciens savent donc que non seulement ils n'analysent pas toutes les séquences d'un environnement mais aussi qu'ils ne les assignent pas toutes à des taxa connus. Les séquences non annotées (sur le plan taxonomique, mais aussi sur le plan fonctionnel) constituent même une matière microbiologique noire [LOP 15] : autant d'acteurs des communautés microbiennes dont on a bien du mal à inférer le rôle et l'histoire. De plus, nous verrons plus bas que connaître la diversité des molécules de 16S n'est nullement synonyme de connaître la diversité des autres gènes présents dans les micro-organismes.

Par conséquent, il existe un paradoxe : la métagénomique offre une vue qu'on pourrait qualifier à la fois de large et d'étroite. D'un côté, les séquences sont plus abondantes que jamais, d'un autre, elles ne sont pas forcément aisément

comparables ni nécessairement représentatives. Il est donc tentant d'identifier des moyens d'élargir la vision de la diversité microbienne, ce qui pourrait de prime abord être entrepris en suivant deux directions de recherche très distinctes. Nous allons présenter ces directions et vérifier si elles peuvent parvenir à l'objectif affiché: enrichir les connaissances au sujet des microbiomes en en fournissant une vue plus large. Mais pour cela, un détour et une réflexion préalables s'imposent afin de vérifier si la métagénomique, telle qu'elle est aujourd'hui pratiquée, est ou non engagée sur des sentiers de dépendance.

Sentier de dépendance et production de connaissances

Plusieurs philosophes des sciences se sont demandé dans quelle mesure les pratiques et les technologies de production de données affectent la production des connaissances scientifiques [LEO 12]. Ils s'accordent sur le fait que l'accroissement des données, et les développements de technologies pour les obtenir et les traiter, induisent des changements conceptuels en sciences et dans la manière dont les scientifiques évaluent, comparent, interprètent et réutilisent les jeux de données disponibles. Ainsi, U. Krohs [KRO 12], s'appuyant notamment sur les travaux remarquables de S. Sydow sur les sentiers de dépendances [SYD 05], a proposé qu'à l'instar de l'évolution des organisations sociales, la production de connaissances, opère par phases, et que la « science confortable » tend à entraîner les scientifiques sur des sentiers de dépendances conformistes. Cette conclusion s'appliquait néanmoins à la biologie des systèmes, et afin de préciser la situation de la métagénomique, nous devons présenter en quelques lignes ce que nous entendons par la notion de sentiers de dépendances épistémiques.

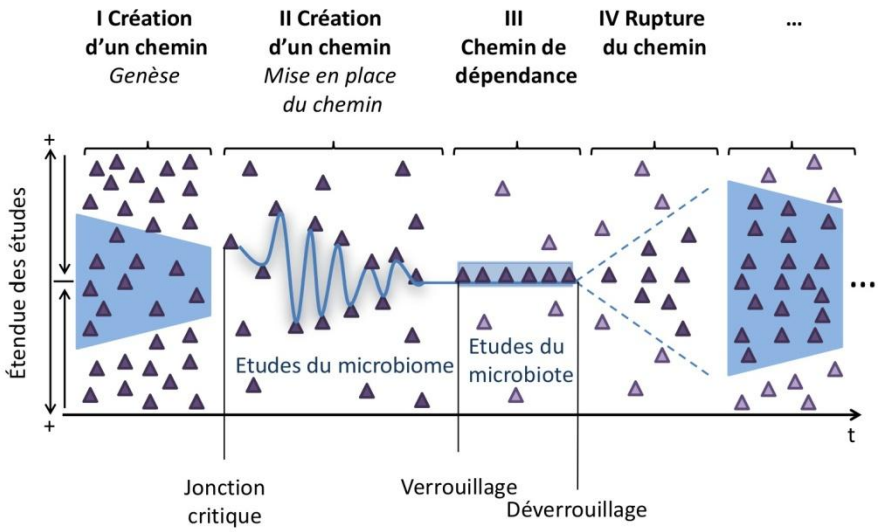


Figure 2 : Déroulé hypothétique des étapes permettant d'acquérir des connaissances dans une discipline scientifique, adapté de Sydow et al. 2009.

Nous nous appuierons pour cela sur le célèbre schéma de S. Sydow [SYD 05] (*cf* Fig 2), que nous réinterpréterons dans le contexte de l'étude des communautés microbiennes. Selon ce schéma, la première étape de la production de connaissances (la phase I) est caractérisée par un foisonnement de questions, engendrées par l'application d'une diversité de méthodes d'analyses et de production de données. Les connaissances obtenues de manières hétérogènes sont donc peu, voire pas, comparables, et les *explananda* fleurissent, formant un large catalogue d'observations. Les méthodes employées dans cette première phase ne sont pas complètement arbitraires, ni aléatoires : elles dépendent de l'histoire des disciplines qui ont précédé la métagénomique, mais elles sont très peu contraintes et seulement faiblement sélectionnées. Une progression vers une seconde phase (phase II) survient lorsque les choix des méthodes d'analyses commencent à converger, en particulier parce que l'emploi de certaines approches (plutôt que d'autres) se traduit par des récompenses. Au-delà de ce point critique, un feedback positif trie parmi les méthodes et donc parmi les *explananda* qui seront l'objet d'études de la discipline. Ceci se comprend très intuitivement puisque si certaines méthodes produisent des publications qui engendrent des financements, ces méthodes peuvent être privilégiées pour des études futures par un effet boule de neige. Lorsque les méthodes de plusieurs scientifiques commencent ainsi à converger, des généralisations au sujet des phénomènes étudiés deviennent en principe possibles,

puisqu'on peut comparer des travaux comparables entre eux. Néanmoins, dans cette seconde phase, toute la connaissance n'est pas exclusivement produite par un nombre restreint d'approches. De nombreux travaux demeurent pratiquement et conceptuellement hétérogènes. La discipline n'est donc pas encore sur un sentier de dépendance : elle commence simplement à trouver son chemin. La situation peut néanmoins évoluer vers une troisième étape (phase III) lorsqu'un verrouillage survient : une ou quelques approches s'imposent à tous les nouveaux venus. La « science confortable » favorise ce verrouillage. C'est la fameuse « kitification de la biologie », puisque beaucoup de procédures expérimentales sont standardisées et basées sur des kits fournis par l'industrie. Cette troisième phase permet donc de produire des *explananda* standardisés, ce qui permet d'approfondir les connaissances en autorisant un maximum de comparaisons entre les études, mais limite paradoxalement leur portée puisqu'elles se concentrent toutes sur le même aspect précis de la diversité biologique. Ceci se traduit même d'ailleurs par la promotion d'une ontologie spécifique de telles études, à l'image de l'usage des Operational Taxonomic Units (OTUs). Une OTU [SNE 62] est définie comme l'unité de base permettant de regrouper les individus proches phylogénétiquement. Au sein d'un échantillon, la construction d'OTUs nécessite donc de comparer toutes les séquences de l'échantillon deux à deux et de considérer que deux séquences appartiennent à la même OTU si la similarité entre ces séquences est supérieure à une valeur seuil choisie. La valeur seuil choisie lorsqu'il s'agit d'ARNr 16S est souvent 97% de similarité entre les deux séquences, parce qu'il est communément admis qu'une OTU définie à 97% de similarité correspond à une espèce [KON 05]. La "kitification" de la biologie se traduit ici par l'utilisation de QIIME [CAP 10] [SHA 14] dans de nombreuses études pour construire les OTUs, OTUs qui seront ensuite utilisées par ce même logiciel pour effectuer une suite d'analyses (annotations taxonomiques, analyses de diversité, ...). Du choix de ces méthodes naîtront des théories. Néanmoins, les types de résultats obtenus en phase III se caractérisent aussi par des angles morts, des *explananda* invisibles. Une telle standardisation risque potentiellement de couper les scientifiques de découvertes encore plus intéressantes. C'est pourquoi S. Sydow [SYD 05] souligne la possibilité d'une dernière phase de production de connaissance (phase IV) : la sortie d'un sentier de dépendance. Cette démarche doit être particulièrement active pour parvenir à déverrouiller les pratiques caractéristiques d'une discipline, même quand ces pratiques ne donnent pas de résultats optimaux. Par analogie, il est très difficile de changer les claviers des ordinateurs bien que l'organisation des touches ne se justifie plus de la même manière que lorsque les claviers QWERTY ont été inventés avec les machines à écrire.

Par rapport à ces phases, où se situent les études de la diversité des communautés microbiennes par la métagénomique ? Celles-ci nous paraissent être dans une situation hybride, différentes selon qu'il s'agit des études du microbiome ou des

études du microbiote. Incontestablement, au-delà de l'échantillonnage (qui peut présenter de nombreux points communs entre différentes études), l'acquisition des données emprunte plusieurs approches parallèles, qui se distinguent par des détails techniques mais qui conceptuellement ambitionnent à produire le même type de données moléculaires. Il y a donc un corridor de méthodes de séquençage relativement semblables plutôt que disparates, ce qui se traduit notamment par la production de nombreux reads de 16S et de 18S dans le cas des analyses de métagénomique ciblée sur un marqueur, et par la production de nombreux reads provenant de régions aléatoires des génomes dans le cas de la métagénomique non-ciblée. La différence entre ces séquences tient principalement dans la longueur des reads et le type d'erreurs de séquençage associées à chaque méthode. Cette faible diversité des méthodes de production de données cadre parfaitement avec l'existence d'une « science confortable », puisque les développeurs des méthodes de séquençage ont pour but de faire passer le plus rapidement possible ces technologies de recherche du statut de nouveautés au statut de standard.

En revanche, la phase d'analyse suivante (le « data mining ») pour comprendre ces données métagénomiques est pour le moment moins contrainte, même si les grandes questions à traiter sont déjà posées. Plus précisément, au sujet des gènes comme des taxa, les métagénomiciens se demandent surtout: « qui est là ? », « qui fait quoi ? » et « quelles communautés sont les plus diverses ? ». Autrement dit, il y a une forte sélection sur les méthodes, non pas en raison de leur implémentation qui peuvent être distinctes, mais en fonction des questions auxquelles elles permettront de répondre, qui sont généralement les mêmes. Ces observations placent la métagénomique au minimum en phase II dans le schéma de Sydow, c'est-à-dire à proximité d'un sentier de dépendance. De fait, nous pensons qu'en raison de la diversité des pipelines bioinformatiques mis en œuvre lors de l'étape de « data mining », les études du microbiome sont probablement assignables à cette phase II. En effet, il existe plusieurs manières de répondre à la question « quels sont les taxons présents dans le microbiome étudié ? ». D'une part, l'étude peut se concentrer sur certains gènes particuliers détectés dans le microbiome. Ces marqueurs phylogénétiques sont ainsi comparés à leurs homologues des bases de données de réference (recherches d'homologues), afin de recevoir une assignation taxonomique puis d'être inclus sur un arbre phylogénétique [SHA 14]. Cette méthode se base sur des algorithmes de classification et sur la similarité des reads par rapport à des séquences de marqueurs de gènes. Plusieurs outils peuvent être utilisés pour répondre à la question posée avec cette méthode : MetaPhyler [LIU 11], MetaPhIAn [SEG 12], AMPHORA [WU 08] [WU 12]. D'autre part, diverses approches, dites de binning, peuvent être employées pour distinguer les différents taxa présents dans un métagénome [SHA 14]. Les reads du microbiome peuvent aussi être assemblés en contigs (indépendamment ou suite au binning). Cette concaténation de petites fractions d'ADN de taille variant entre 35 paires de base et

plusieurs centaines de paires de base (selon la méthode de séquençage) facilite l'annotation taxonomique et fonctionnelle du contenu du microbiome, tout en fournissant des informations sur le contexte génomique dans lequel les gènes microbiens se trouvent. Il existe plusieurs outils pour réaliser ces assemblages: IDBA [PEN 12], MetaVelvet [NAM 12], metaSPAdes (spades.bioinf.spbau.ru/release3.10.0/manual.html), Ray Meta [BOI 12], MetAmos [TRE 13] [TRE 11]... En outre, toutes les analyses du microbiome ne portent pas forcément sur les mêmes questions. Là où certains se demandent, quels taxa sont présents dans la communauté, d'autres se demandent plutôt : quels gènes sont présents dans la communauté ? Comment interagissent les microbes au sein de la communauté ? Quelle est la diversité génétique d'un microbiome ? Dans quels contextes génomiques un gène donné se retrouve-t-il ? En résumé, les études du microbiome impliquent différentes méthodes, qui exploitent chacune plusieurs outils, dont l'usage nécessite de choisir un nombre important de paramètres. Cette succession de méthodes et d'heuristiques plongent l'analyste face à de nombreux dilemmes.

Néanmoins, pour les études de microbiotes (probablement parce que les données de départ sont encore plus homogènes), les pipelines et les « bonnes pratiques » sont nettement plus définies et restreintes. Les analyses de la diversité microbienne (fondées sur du 16S ou du 18S) semblent être entrées en phase III, c'est-à-dire s'être embarquées sur un sentier de dépendance. Si l'on prend l'exemple de la question "Qui est là ?" pour l'étude d'un microbiote, dont les données à analyser sont de l'ARNr 16S, il existe un sentier de dépendance couramment emprunté : tout d'abord, on définit des OTUs à 97% d'identité. Ensuite ces OTUs sont assignés taxonomiquement en comparant les séquences centroïdes des OTUs à une base de données de référence. Ces assignations taxonomiques permettent ensuite de calculer la diversité au sein de chaque échantillon (alpha diversité), puis de comparer la diversité des échantillons deux à deux (bêta diversité). Enfin, on utilise les assignations taxonomiques pour calculer les abondances relatives de chaque taxa par échantillon. Si, à notre connaissance, cette façon de répondre à la question "Qui est là ?" est standard, il demeure cependant une diversité d'outils et de mesures pour chaque étape : l'outil QIIME est le plus couramment utilisé pour la construction d'OTUs, mais il dispose de plusieurs algorithmes pour effectuer cette tâche (pick_de_novo_otus.py, pick_closed_reference_otus.py, pick_open_reference_otus.py). Il existe cependant d'autres outils permettant de construire des OTUs, tels que MOTHUR [SCH 09], et il existe une diversité de mesures de distance pour réaliser le clustering des séquences en OTUs [NGU 16]. Pour l'alpha diversité, c'est à dire la quantification de la diversité au sein de chaque échantillon, il existe de nombreux indices utilisables dont ceux de Shannon, Simpson et Chao1 mais aussi d'autres mesures se basant sur la phylogénie telles que la Balance-Weighted Phylogenetic Diversity (BWPD) afin de caractériser la

diversité au sein des échantillons. De la même façon, il existe différentes mesures de distance pour la bêta diversité (Jaccard, euclidienne, Bray-Curtis pour n'en citer que trois) et différentes représentation (PCoA, NMDS,...). Il existe donc bel et bien un sentier de dépendance pour les analyses de microbiote, cependant, si les étapes et les analyses sont systématiquement les mêmes, il existe plusieurs façons de réaliser ces étapes. Ce sentier de dépendance est balisé non seulement par les réponses standard, mais aussi par les questions que l'on se pose, qui sont déterminées et identiques dans la plupart des études (qui est là? En quelles proportions ? Quelle est la diversité taxonomique ?).

Que tirer de cette catégorisation ? Si la phase II permet certaines généralisations, elle ne permet pas de maximiser les comparaisons standard. Par conséquent, pour voir plus loin, c'est-à-dire pour étendre les études comparatives, une piste pourrait être de faire en sorte que l'étude des microbiomes passe de la phase II à la phase III. Pour sa part, la phase III impose certaines œillères analytiques. Aussi, pour voir plus loin, c'est-à-dire pour voir autre chose et autrement, une piste supplémentaire pourrait également être appliquée, qui viserait à faire passer les études des microbiotes de la phase III à la phase IV. Dans la mesure où sur un plan épistémique la phase IV n'est probablement pas très différente de la phase II, voire de la phase I (la différence entre ces phases étant simplement leur ordre d'apparition chronologique dans l'histoire des sciences) [KRO 12], ces deux inspirations pour renforcer les aptitudes holistiques de la métagénomique apparaissent de prime abord philosophiquement contradictoires. Nous allons discuter rapidement des avantages et des limites de chacune de ces stratégies afin de vérifier si elles pourraient réellement jouer le rôle espéré : accroître la profondeur et la portée des explications de la métagénomique.

Standardiser la métagénomique

Standardiser la production et l'analyse des jeux de données métagénomiques semblent en principe offrir un moyen de voir au-delà d'un jeu de donnée métagénomique particulier. En effet, différentes études deviendraient alors comparables. Ce type d'approche est entrepris notamment grâce à la mise en commun de scripts bioinformatiques. Ceci est illustré par le cas du script permettant de trouver des entérotypes, qui est disponible avec un tutoriel (*cf* enterotype.embl.de). Ce tutoriel permet d'avoir une méthode standardisée pour chercher des entérotypes et donc de rendre les études un peu plus comparables entre elles (tout en gardant à l'esprit que la méthode de séquençage, le type de reads, ... peuvent avoir un impact sur le résultat obtenu). Des scientifiques ont pu utiliser ce tutoriel pour comparer leurs résultats [MOE 15] [MOE 12] [LIM 14] [LI 17] avec

ceux obtenus dans l'article original [ARU 11]. Cela permet de comparer des groupes réalisés avec la même méthode de clustering (ici la méthode PAM du package R "cluster"). Le résultat d'une telle analyse est remarquable, mais ambivalent. D'une part, on peut tester l'existence de communautés semblables dans différentes lignées d'hôtes, et donner ainsi une interprétation phylogénétique des ressemblances des microbiomes associés à des espèces apparentées. Par exemple, les primates partageraient 3 entérotypes. Parce que ces entérotypes se retrouvent chez les chimpanzés, les gorilles et l'homme, il semble légitime de proposer que ces 3 types de communautés aient pu évoluer dans leur dernier ancêtre commun (et peut-être même antérieurement). Néanmoins, et ceci illustre bien la difficulté, voire l'impossibilité de standardiser réellement les études de métagénomique, deux individus peuvent à la fois présenter le même entérotype et néanmoins abriter des microbes ayant des contenus en gènes très différents, donc pour le dire très simplement les mêmes entérotypes peuvent correspondre à des microbes très différents. Il y a là un paradoxe remarquable, qui s'explique d'une part par les processus biologiques affectant les génomes microbiens et d'autre part, par la réduction (un peu essentialiste) imposée par la standardisation des études des communautés microbiennes. En effet, on dit que deux individus partagent le même entérotype s'ils présentent les mêmes combinaisons (et selon les mêmes abondances relatives) d'OTUs de 16S et de 18S. Mais les génomes des microbes dont le 16S appartient au même OTU peuvent déjà différer dans des proportions considérables, les souches procaryotes gagnant et perdant des gènes de façon très rapide [DOO 10]. Il y a donc dans une large mesure semblant (voire illusion) de comparabilité entre des jeux de données métagénomiques standardisés : les séquences révélant la présence des mêmes OTUs dans deux jeux de données peuvent signifier qu'on a affaire dans ces deux jeux de données aux « mêmes espèces, sauf leurs gènes ! » (ce qui interdit de s'en remettre d'ailleurs aux seules études de 16S/18S pour comprendre vraiment la diversité et le fonctionnement de ces communautés).

D'autre part, si la production et l'analyse bioinformatique des données sont standardisables, la nature même des objets étudiés en métagénomique et l'objectif scientifique de cette approche, à savoir comprendre les causes de la diversité environnementale, signifie que les scientifiques vont s'intéresser à des environnements différents, et cela interdit par définition d'uniformiser les contextes divers desquels proviennent les échantillons. C'est une des différences majeures entre les études en laboratoire, dans lesquels les différents paramètres génétiques et environnementaux peuvent être contrôlés pour donner lieu à des expériences reproductibles, et les études de métagénomique. Les microbiomes disponibles auront beau être obtenus et analysés de façon semblable, ils n'en proviendront pas moins de conditions différentes, et les comparer reviendra à comparer des échantillons obtenus toutes choses n'étant pas égales par ailleurs. De plus, il sera très difficile de faire la part de l'impact des différents paramètres environnementaux sur la diversité

des communautés microbiennes, dans la mesure où il est généralement compliqué de maintenir constants certains de ces paramètres tout en en faisant varier d'autres. Par conséquent, il semble réaliste d'anticiper la découverte de corrélations entre différents facteurs environnementaux et la diversité microbienne plutôt que la découverte de causes de cette diversité, même au terme de comparaisons extrêmement standardisées entre microbiomes. Ainsi, plusieurs études ont eu pour but de comparer l'impact d'un régime alimentaire humain occidental avec un régime alimentaire différent (par exemple régime alimentaire du Venezuela, du Malawi [YAT 12], du Groenland, ou du Burkina Faso [DEF 10],...). Le problème, est qu'il est difficile de savoir si la différence observée entre ces microbiomes est due à une différence de régime alimentaire, ou de contrainte génétique de l'hôte (du fait que l'on travaille sur des populations différentes), ou encore due à l'environnement (par exemple les pathogènes en présence ne sont pas les mêmes) ou à une combinaison de ces variables. On pourrait imaginer y voir plus clair en séquençant également le métagénome de l'environnement des ces organismes, si cela permettait de vérifier quelles bactéries et quels gènes sont présents (et se retrouvent éventuellement dans leurs microbiomes intestinaux). Mais un tel complément d'études aurait finalement un coût élevé, et nécessiterait un espace de stockage considérable. De même, l'étude comparative de la génétique des populations dont le microbiome est étudié pourrait améliorer les interprétations. Aussi, il n'est pas certain que standardiser la recherche en métagénomique remplisse *in fine* les objectifs espérés par les scientifiques qui s'engageraient dans cette entreprise de création de nouveaux sentiers de dépendance. Et cette piste aurait probablement l'inconvénient de canaliser la recherche sur la diversité.

Déverrouiller la métagénomique

Une autre stratégie pour essayer de voir plus large en métagénomique pourrait adopter l'attitude inverse, et jouer à fond le jeu d'une science dirigée par ses données [LEO 12]. Il s'agirait alors de multiplier les jeux de données ainsi que les *explananda* (les descriptions de ces données et des phénomènes qu'on peut y déceler), en particulier en déverrouillant les méthodes de « data-mining » pour étudier les microbiotes (et les microbiomes). Une manière triviale de commencer à déverrouiller les études du microbiote consiste à ne pas s'appuyer uniquement sur des molécules de 16S et de 18S (ou sur les mêmes régions de ces marqueurs), mais de considérer aussi d'autres régions ou d'autres marqueurs, telles que les protéines ribosomales, obtenues lors du séquençage aléatoire du microbiome afin de comparer les prédictions (souvent contradictoires) [STO 10] faites par ces types de marqueur différents au sujet des taxa présents dans la communauté. Un déverrouillage plus

effectif encore est envisageable. Il supposerait de favoriser des méthodes qui suivent des buts plus exploratoires que pragmatiques, c'est-à-dire des méthodes de production et d'analyse de données moins sélectionnées en raison de questions établies auxquelles elles permettent de répondre qu'en raison des nouvelles questions qu'elles pourraient permettre de poser.

Le « data mining » purement exploratoire peut notamment prendre la forme de création de réseaux de similarité de séquences et d'approches de « graph-mining ». Il est en effet possible d'analyser la diversité des communautés microbiennes sous de multiples facettes, en analysant celles de leurs reads, de leurs contigs, de leurs gènes, de leurs génomes, de leurs taxa, et des co-occurrences et des interactions entre ces taxa, etc. Par exemple, il est possible de construire des réseaux de similarités entre reads dont la structure du graphe est informative sur la biologie de la communauté microbienne (*cf* figure 3).

Dans le cadre de l'étude de l'impact d'un changement de régime alimentaire sur le microbiome intestinal de lézards, nous avons construit un tel réseau de similarités entre reads pour chaque lézard. Pour cela, l'ensemble des reads constituant chaque microbiome ont d'abord été comparés entre eux, deux à deux à l'aide d'un BLAST [NCBI] [ALT 90] tout contre tout. La sortie de ce BLAST a été filtrée afin de ne conserver que les alignements présentant au moins 90% d'identité (i.e. 90% des paires de base de l'alignement entre deux reads sont identiques) et au moins 80% de couverture (i.e. au moins un des deux reads impliqués dans l'alignement doit avoir 80% de la longueur de sa séquence d'impliquée dans l'alignement) et une E-value inférieure à 1e-5. Le fichier de sortie est donc un réseau enregistré sous forme d'une liste d'arêtes, qui représentent les similarités entre deux reads (qui sont les nœuds du réseau). Ce réseau peut ensuite être visualisé à l'aide d'outils tels que Gephi [BAS 09] ou Cytoscape [CHR 05]. Dans le cas idéal où l'on aurait réussi à appréhender toute la diversité d'un métagénome, on s'attendrait à ce que ce genre de réseaux reproduise les génomes circulaires des bactéries. Dans la mesure où l'on sait que les métagénomes sont des échantillons de l'environnement, et que l'on ne parvient pas à en séquencer toute la diversité, on s'attend plutôt, dans les jeux de données réelles, à obtenir des contigs, c'est à dire des chaînes de reads plus ou moins longues formant chacune un sous-réseau, que l'on nomme également composante connexe. C'est ce que l'on obtient majoritairement dans nos réseaux, observation qui nous a d'ailleurs permis de caractériser une nouvelle classe de graphe appelée k-laminaire [VOL 16]. Un k-laminaire (voir figure 3.a) est une composante connexe dont chaque nœud se situe à une distance inférieure ou égale à k arêtes du chemin diamétral, c'est-à-dire du plus long des plus courts chemins entre deux nœuds dans le graphe). Cette distance k au chemin diamétral est intéressante, car plus elle est grande, plus il y a de reads variants qui se rattachent à une même région d'ADN, et donc, plus il y a de

diversité génétique. La valeur de k est une nouvelle manière de quantifier la diversité génétique dans un métagénome.

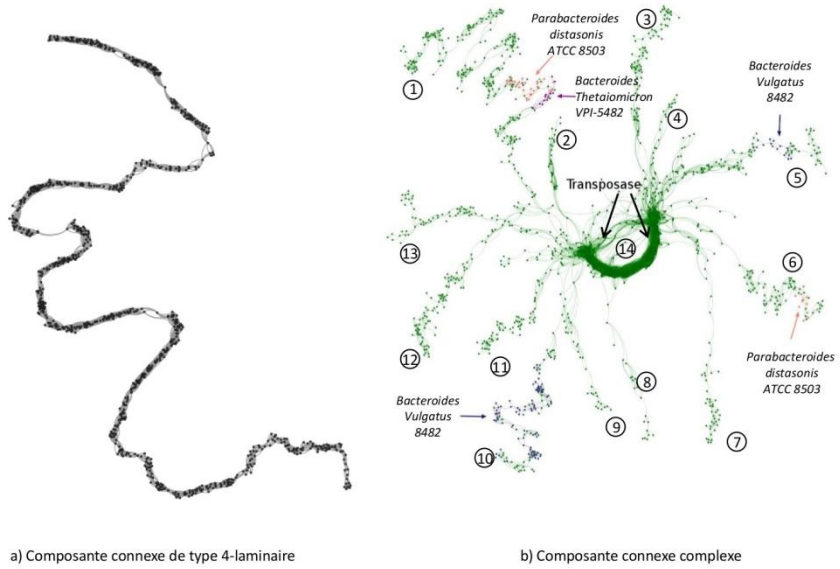


Figure 3 : Composantes connexes provenant d'un réseau de similarités entre reads d'un microbiome intestinal de lézard.

Cependant, certaines composantes connexes de nos réseaux de reads adoptent des topologies plus complexes qu'un k-laminaire, comme le montre la figure 3.b, qui présente un ensemble ressemblant à 13 laminaires (numérotés de 1 à 13 sur la figure), unis par une boucle centrale (Figure 3, 14), qui semble créer une jonction entre ces laminaires. La boucle au centre du réseau contient des reads dont l'annotation fonctionnelle correspond à une tranposase. Les différentes parties de la composante connexe reliées à cette transposase sont constituées de différentes souches bactériennes d'après l'annotation taxonomique. Ce graphe fournit des informations sur la diversité génétique dans la communauté microbienne d'un lézard, et permet de montrer comment certains gènes bien conservés sur la plan de la séquence, ici la transposase, se retrouvent dans différents contextes génomiques. A ce stade, il est cependant tout à fait aléatoire d'espérer un retour positif à l'issue du développement de telles méthodes. Précisément parce qu'elles sont exploratoires, leur utilité n'est pas prévisible, ni donc la vraisemblance ou la possibilité de les publier dans des journaux à fort facteur d'impact. La science qui sort des sentiers battus a par définition tous les risques de s'égarer… mais elle a aussi la possibilité

d'établir des connaissances inattendues. Dans le cas particulier des réseaux de reads, cette approche alternative de l'étude de la diversité génétique des communautés microbiennes a mis en évidence une nouvelle classe de réseau [VOL 16]. Cette découverte semble pour le moment plus importante pour les théoriciens des graphes que pour les biologistes. Mais qui peut prédire à l'avenir si ce lien supplémentaire et neuf entre ces disciplines ne sera pas autrement fructueux ?

Fondamentalement, la justification du déverrouillage de la métagénomique est qu'une connaissance plus large de la diversité des communautés microbiennes est envisageable, si l'on considère chaque jeu de données comme une pièce d'un puzzle plus grand dans une démarche volontairement intégrative, c'est-à-dire une démarche qui combine une gamme d'approches afin d'explorer la diversité biologique [OMA 12]. Dans ce cadre, chaque jeu de données et chaque analyse constitue une pièce d'un puzzle dont la vue d'ensemble se révélera plus tard. La perspective intégrative paraît particulièrement naturelle en métagénomique puisque les biologistes étudiant les microbiomes des animaux et des plantes sont manifestement en quête de modèles plus systémiques, qui décrivent la dynamique des communautés de microbes dans des communautés d'hôtes, voire dans les écosystèmes. Le déverrouillage des études métagénomiques se justifie également dans une démarche perspectiviste [CAL 12] (moins ambitieuse), mais qui vise à multiplier les points de vue théoriques, dans la mesure où chaque jeu de données capture des aspects de la diversité microbienne, phénomène pour le moins complexe et multi-échelles. Les données de métagénomique peuvent parfaitement jouer ce rôle. Ainsi, la découverte dans les grands fonds sous-marins de séquences d'un nouveau groupe d'archées, les Asgard, initialement représentés par des métagénomes de Loki qui ne sont pas forcément comparables à d'autres métagénomes, n'en a pas moins permis de consolider les théories sur l'origine symbiogénétique des eucaryotes [SAW 15] [SPA 15]. Ici, l'information d'un métagénome a servi à éclairer une théorie indépendante de la métagénomique. Par ailleurs, des méthodes nouvelles sont indispensables puisque les approches en place peinent à annoter fonctionnellement et taxinomiquement un grand nombre de séquences (ce qui illustre bien que, seules, elles sont insuffisantes) [SUN 13].

Si la piste du déverrouillage semble offrir une vue plus large, il s'agit néanmoins d'un pari optimiste, car rien ne garantit que les pièces de puzzles s'assembleront en une image générale, plutôt qu'elles n'enrichiront un catalogue de descriptions naturalistes, locales, du monde microbien. De plus, le déverrouillage amplifiera immanquablement la quantité d'*explananada* pour lesquels des comparaisons standardisées sont impossibles, ce qui risque de désespérer les chercheurs qui espèrent surnager dans le flot des connaissances métagénomiques plutôt que de s'y noyer.

Conclusion

Se lancer dans des études de métagénomique est certainement fascinant mais aussi source de frustrations multiples, tant les difficultés sont grandes pour valoriser des découvertes locales ou pour tenter de tirer des conclusions plus générales. Les données sont très abondantes, parce qu'elles peuvent être obtenues de façon relativement automatique. Ce type de « science confortable » garantissant un ensemble d'analyses conformistes dispose néanmoins d'une portée théorique limitée. Analyser les microbiomes pour espérer en extraire les connaissances les plus profondes et les plus générales possibles constitue d'ailleurs aujourd'hui un défi considérable. Même s'il existe des bases de données centralisées comme celles du NCBI [COO 16], HMP (*cf* http://hmpdacc.org/), EBI [MIT 16], TARA [PES 15] [SUN 15] [VAR 15], MGRAST [MEY 08], en l'état, il est compliqué de comparer les données et les conclusions des études existantes entre elles. Tous les microbiomes ne sont pas comparables entre eux, en raison de la diversité de leurs hôtes ou de leurs environnements d'origine, de la taille et de la nature de ces jeux de données, des limites d'interprétation des ontologies standardisées associées. Tester des hypothèses au sujet de la diversité microbienne et en identifier les causes s'avère difficile à une échelle locale, et très difficile à une échelle plus générale. La métagénomique est donc encore largement une science naturaliste, qui produit des catalogues. Nous prévoyons que ce statut a de bonnes chances de perdurer au vu de la taille et de la complexité réelles des microbiomes. Cet état (épistémique) de l'art que nous avons brossé invite, selon nous, à mener de front deux stratégies contradictoires : la standardisation et le déverrouillage de la métagénomique, dans l'espoir que celles-ci s'avèreront complémentaires. Evidemment, cette proposition pluraliste se heurte à un dernier problème pratique incontournable dans les études de la diversité : quelles ressources devraient et pourraient être respectivement attribuées à ces deux stratégies et au stockage de données résultantes pour permettre leur bonne cohabitation?

Remerciements

Eric Bapteste est financé par l'ERC (European Research Council FP7/2007-2013 Grant Agreement 615274) Et Chloé Vigliotti par le LabexBCDIV.

Nous remercions l'équipe qui a récolté les lézards en Croatie, plus particulièrement Beck Wehrle et Anthony Herrel, ainsi que l'entreprise qui a séquencé nos microbiotes et microbiomes : MrDNA (Scot Dowd).

Enfin, nous souhaitons remercier les théoriciens des graphes avec qui nous avons collaboré pour l'étude de réseaux de similarités dans les microbiomes de lézards : Michel Habib, Léo Planche et Finn Völkel.

Légendes des figures

Figure 1 : Impact de la méthode de séquençage sur la composition taxonomique du microbiote de lézards.

Cette figure est une matrice lézards x phyla bactériens représentant l'abondance de chaque phylum au sein de chacun des microbiomes de lézards, estimée en quantifiant le nombre de reads de la région V4 de l'ARN16S assignés à ce phylum. A chaque couple (lézard X, phylum Y) de la matrice, a été associée une couleur tenant compte de l'abondance du phylum Y dans le microbiote du lézard X. Plus la couleur est foncée, plus le phylum Y est abondant dans le microbiote du lézard X. Les trois phyla encadrés correspondent aux trois phyla les plus abondants (phyla majoritaires).

Un clustering basé sur une distance de Bray-Curtis regroupant les lézards dont les abondances en phyla sont proches a été appliqué sur cette matrice. Lorsque l'on s'intéresse aux caractéristiques génétiques et écologiques des lézards au sein des groupes, on se rend compte qu'ils sont regroupés par année de séquençage. Le régime alimentaire des lézards a aussi été indiqué sur cette matrice, et une distinction entre les lézards insulaires et continentaux a été prise en compte. La méthode de séquençage (contrainte technologique) a un impact plus important sur la structure du microbiote que la différence de régime alimentaire (qui est une réalité biologique) ou qu'un effet potentiel de l'insularité.

Figure 2 : Déroulé hypothétique des étapes permettant d'acquérir des connaissances dans une discipline scientifique, adapté de Sydow *et al.* 2009.

L'axe x représente l'axe temporel, depuis les débuts d'une méthode (à gauche), jusqu'à ses développements ultérieurs (vers la droite). Les triangles représentent les différentes approches/méthodes employées dans la discipline. Plus la couleur du triangle est claire, moins la méthode/approche qu'il représente est utilisée. La dispersion des triangles représentent la différence entre méthodes : des méthodes

très similaires occupent un espace proche. Les étapes clefs de l'évolution de la discipline sont indiquées en haut et en bas du schéma.

Figure 3 : Composantes connexes provenant d'un réseau de similarités entre reads d'un microbiome intestinal de lézard.

Le réseau est visualisé à l'aide de l'outil Gephi**.** Les nœuds de chacune des composantes connexes du réseau sont des reads individuels, et deux reads sont reliés par une arête s'ils ont un pourcentage de couverture (couverture non mutuelle) supérieur à 80% et un pourcentage d'identité supérieur à 90% et une Evalue < 1^{e}-5 selon BLAST.

a)Composante connexe de type 4-laminaire (i.e. le nœud le plus éloigné du chemin diamétral est à 4 arêtes de ce chemin) contenant 2136 nœuds et 13 991 arêtes. *b) Composante connexe plus complexe* contenant 1843 nœuds et 25022 arêtes. La numérotation de 1 à 14 correspond aux différentes parties de la composante connexe. Les treize premiers numéros sont des régions de la composante connexe des laminaires annotés taxonomiquement et reliés entre eux par une région (boucle, numero 14) au centre du réseau contient des reads dont l'annotation fonctionnelle correspond à une tranposase. La structure de cette composante connexe (des laminaires joints par une boucle) est informative sur la biologie de la communauté microbienne et permet de visualiser comment certains gènes (ici, la transposase) peuvent être présents dans différents contextes génomiques (ici dans des populations de *bacteroides vulgatus* et *thetaiomicron*, ainsi que de *parabacteroides distasonis*).

Bibliographie

[ALT 90] ALTSCHUL S., GISH W., MILLER W., et al., « Basic local alignment search tool », J Mol Biol, n° 215, p. 403-410, 1990.

[ARU 11] ARUMUGAM M., RAES J., PELLETIER E., *et al*., « Enterotypes of the human gut microbiome », *Nature*, n° 473, p. 174–180, 2011.

[BAS 09] BASTIAN M., HEYMANN S., JACOMY M., « Gephi: An Open Source Software for Exploring and Manipulating Networks », *International AAAI Conference on Weblogs and Social Media*, San Jose, USA, 2009.

[BOI 12] BOISVERT S., RAYMOND F., GODZARIDIS É, et al., « Ray Meta: scalable de novo metagenome assembly and profiling », *Genome Biol*, n° 13, p. R122, 2012.

[BRO 15] BROWN CT., HUG LA., THOMAS BC., *et al.*, « Unusual biology across a group comprising more than 15% of domain Bacteria », *Nature*, n° 523, p. 208–211, 2015.

[CAL 12] CALLEBAUT W. 2012. « Scientific perspectivism: A philosopher of science's response to the challenge of big data biology », *Stud Hist Philos Sci Part C*, n° 43, p. 69–80, 2012.

[CAP 10] CAPORASO JG., KUCZYNSKI J., STOMBAUGH J, *et al.*, « QIIME allows analysis of high-throughput community sequencing data », *Nat Methods*, n° 7, p. 335–336, 2010.

[CHR 05] CHRISTMAS R., AVILA-CAMPILLO I., BOLOURI H., *et al.*, « Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks », *Genome Research*, n° 13(11), p. 2498-2504, 2005.

[COO 16] COORDINATORS NR., « Database resources of the National Center for Biotechnology Information », *Nucleic Acids Res*, n° 44, p. D7–D19, 2016.

[DEF 10] De FILIPPO C., CAVALIERI D., Di PAOLA M., *et al.*, « Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa », *Proc Natl Acad Sci U S A*, n° 107, p. 14691–14696, 2010.

[DOO 10] DOOLITTLE WF., ZHAXYBAYEVA O., « Metagenomics and the Units of Biological Organization », *Bioscience*, n° 60, p. 102–112, 2010.

[GIL 06] GILL SR., POP M., DEBOY RT., *et al.*, 2006. « Metagenomic analysis of the human distal gut microbiome », *Science*, n° 312, p. 1355–1359, 2006.

[GOM 15] GOMEZ A., PETRZELKOVA K., YEOMAN CJ., *et al.*, « Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (Gorilla gorilla gorilla) reflect host ecology. », *Mol Ecol*, n° 24, p. 2551–2565, 2015.

[HER 08] HERREL A., HUYGHE K., VANHOOYDONCK B., *et al.*, « Rapid large-scale evolutionary divergence in morphology and performance associated with exploitation of a different dietary resource », *Proc Natl Acad Sci*, n° 105, p.4792–4795, 2008.

[HMPC 12] HUMAN MICROBIOME PROJECT CONSORTIUM ,« Structure, function and diversity of the healthy human microbiome », *Nature*, n° 486, p.207–214, 2012.

[HON 11] HONG P-Y., WHEELER E., CANN IKO., *et al.*, « Phylogenetic analysis of the fecal microbial community in herbivorous land and marine iguanas of the Galápagos Islands using 16S rRNA-based pyrosequencing », *ISME J*, n° 5, p. 1461–1470, 2011.

[JOU 17] JOUSSET A., BIENHOLD C., CHATZINOTAS A., *et al.*, « Where less may be more: how the rare biosphere pulls ecosystems strings », *ISME J*, n° 11, p. 853–862, 2017.

[KOH 13] KOHL KD., CARY TL., KARASOV WH., *et al.*, « Restructuring of the amphibian gut microbiota through metamorphosis », *Environ Microbiol Rep*, p. 5, n° 899–903, 2013.

[KON 05] KONSTANTINIDIS KT., TIEDJE JM., « Genomic insights that advance the species definition for prokaryotes », *Proc Natl Acad Sci U S A*, n° 102, p. 2567–2572, 2005.

[KRO 12] KROHS U., « Convenience experimentation », *Stud Hist Philos Sci Part C* 43, p.52–57,2012.

[KUC 12] KUCZYNSKI J., STOMBAUGH J., WALTERS WA., *et al*., « Using QIIME to analyze 16s rRNA gene sequences from microbial communities », *Curr Protoc Microbiol*., n° 36, p. 10.7:10.7.1–10.7.20, 2012.

[LEC 13] LE CHATELIER E., NIELSEN T., QIN J., *et al.*, « Richness of human gut microbiome correlates with metabolic markers », *Nature*, n° 500, p. 541–546, 2013.

[LEO 12] LEONELLI S., « Introduction: Making sense of data-driven research in the biological and biomedical sciences », *Stud Hist Philos Sci Part C*, n° 43, p. 1–3, 2012.

[LI 17] LI J., POWELL JE., GUO J., *et al*., « Two gut community enterotypes recur in diverse bumblebee species », *Curr Biol*, n° 25, p. R652–R653, 2017.

[LIM 14] LIM MY., RHO M., SONG Y-M., *et al*., « Stability of gut enterotypes in Korean monozygotic twins and their association with biomarkers and diet », *Sci Rep*, N° 4, p. 7348, 2014.

[LIU 11] LIU B., GIBBONS T., GHODSI M., *et al*., « Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences », *BMC Genomics* n° 12, S4, 2011.

[LIU 12] LIU L., LI Y., LI S., *et al.*, « Comparison of Next-Generation Sequencing Systems », *J Biomed Biotechnol* vol. 2012, 251364, 11 pages, 2012.

[LOP 15] LOPEZ P., HALARY S., BAPTESTE E. 2015. « Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life », *Biol Direct*, n° 10, p. 64, 2015.

[MA 12] MA B., FORNEY LJ., RAVEL J., « Vaginal microbiome: rethinking health and disease », *Annu Rev Microbiol*, n° 66, p. 371–89, 2012.

[MAR 11] MARTINSON VG., DANFORTH BN., MINCKLEY RL., *et al*., « A simple and distinctive microbiota associated with honey bees and bumble bees », *Mol Ecol*, n° 20, p. 619–628, 2011.

[MCC 14] MCCANN JC., WICKERSHAM TA., LOOR JJ., « High-throughput methods redefine the rumen microbiome and its relationship with nutrition and metabolism », *Bioinform Biol Insights*, n° 8 p. 109–125, 2014.

[MED 17] MEDINA-COLORADO AA., VINCENT KL., MILLER AL., *et al.*, « Vaginal ecosystem modeling of growth patterns of anaerobic bacteria in microaerophilic conditions », Anaerobe, 2017.

[MOE 12] MOELLER AH., DEGNAN PH., PUSEY AE., *et al.*, « Chimpanzees and Humans Harbor Compositionally Similar Gut Enterotypes », *Nat Commun*, n° 3, p. 1179, 2012.

[MOE 15] MOELLER AH., PEETERS M., AYOUBA A., *et al.*, « Stability of the gorilla microbiome despite simian immunodeficiency virus infection », *Mol Ecol*, n° 24, p. 690–697, 2015.

[MEY 08] MEYER F., PAARMANN D., D'SOUZA M., *et al.*, « The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes », *BMC Bioinformatics*, n° 9, p. 386, 2008.

[MIT 16] MITCHELL A., BUCCHINI F., COCHRANE G., *et al.*, « EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data », *Nucleic Acids Res*, n° 44, p. D595, 2016.

[NAM 12] NAMIKI T., HACHIYA T., TANAKA H., *et al.*, « MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads », *Nucleic Acids Res*, n° 40, p. e155–e155, 2012.

[NCBI] NCBI BLAST web site[http://blast.ncbi.nlm.nih.gov/Blast.cgi].

[NEM 11] NEMERGUT DR., COSTELLO EK., HAMADY M., *et al.*, « Global patterns in the biogeography of bacterial taxa », *Environ Microbiol*, n° 13, p. 135–144, 2011.

[NGU 16] NGUYEN N-P., WARNOW T., POP M., *et al.*,. 2016. « A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity », *Npj Biofilms Microbiomes*, n° 2, p. 16004, 2016.

[OMA 12] O'MALLEY MA., SOYER OS., « The roles of integration in molecular systems biology », *Stud Hist Philos Sci Part C*, n° 43, p. 58–68, 2012.

[PEN 12] PENG Y., LEUNG HCM., YIU SM., *et al.*, « IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth », *Bioinformatics*, n° 28, p. 1420, 2012.

[PES 15] PESANT S., NOT F., PICHERAL M., *et al.*, « Open science resources for the discovery and analysis of Tara Oceans data. », *Sci data*, 2:150023, 2015.

[PRA 17] PRADO-IRWIN SR., BIRD AK., ZINK AG., *et al.*, « Intraspecific Variation in the Skin-Associated Microbiome of a Terrestrial Salamander », *Microb Ecol*, p. 1–12, 2017.

[SAW 15] SAW JH., SPANG A., ZAREMBA-NIEDZWIEDZKA K., *et al.*, « Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes », *Nature*, n° 499, p. 431-437, 2015.

[SCH 09] SCHLOSS PD., WESTCOTT SL., RYABIN T., *et al.*, « Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities », *Appl Environ Microbiol*, n° 75, p. 7537–7541, 2009.

[SCH 17] SCHUELLER K., RIVA A., PFEIFFER S., *et al.*, « Members of the Oral Microbiota Are Associated with IL-8 Release by Gingival Epithelial Cells in Healthy Individuals », *Front Microbiol*, n° 8, p.416, 2017.

[SEG 12] SEGATA N., WALDRON L., BALLARINI A., *et al.*, « Metagenomic microbial community profiling using unique clade-specific marker genes », *Nat Meth* n° 9, p.811–814, 2012.

[SHA 14] SHARPTON TJ., « An introduction to the analysis of shotgun metagenomic data », *Front Plant Sci*, n° 5, p. 209, 2014.

[SNE 62] SNEATH PH., SOKAL RR., « Numerical taxonomy » *Nature*, n° 193, p. 855–860, 1962.

[SPA 15] SPANG A., SAW JH., JORGENSEN SL., *et al.*, « Complex archaea that bridge the gap between prokaryotes and eukaryotes », *Nature*, n° 521, p. 173–179, 2015.

[STO 10] STOECK T., BASS D., NEBEL M., *et al.*, « Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water », *Mol Ecol*, n° 19, p. 21–31, 2010.

[SU 16] SU L., YANG L., HUANG S., *et al.*, « Comparative Gut Microbiomes of Four Species Representing the Higher and the Lower Termites », *J Insect Sci*, n° 16, p. 97, 2016.

[SUN 13] SUNAGAWA S., MENDE DR., ZELLER G., *et al.*, « Metagenomic species profiling using universal phylogenetic marker genes », *Nat Methods*, n° 10, p. 1196–1199, 2013.

[SUN 15] SUNAGAWA S., COELHO LP., CHAFFRON S., *et al.*, « Structure and function of the global ocean microbiome », *Science*, 348:1261359, 2015.

[SYD 05] SYDOW J., SCHREYÖGG G., KOCH J., « Organizational Paths: Path Dependency and Beyond », 2005.

[TRE 11] TREANGEN TJ., KOREN S., ASTROVSKAYA I., *et al.*, « MetAMOS: a metagenomic assembly and analysis pipeline for AMOS », *Genome Biol*, n° 12, p. 25, 2011.

[TRE 13] TREANGEN TJ., KOREN S., SOMMER DD., *et al.*, « MetAMOS: a modular and open source metagenomic assembly and analysis pipeline », *Genome Biol*, n° 14, p. R2–R2, 2013.

[TUR 09a] TURNBAUGH PJ., RIDAURA VK., FAITH JJ., *et al.*, « The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice », *Sci Transl Med*, n° 1, 6ra14, 2009.

[TUR 09b] TURNBAUGH PJ., HAMADY M., YATSUNENKO T., *et al.*, « A core gut microbiome in obese and lean twins », *Nature*, n° 457, p. 480–484, 2009.

[VAR 15] DE VARGAS C., AUDIC S., HENRY N., *et al.*, « Eukaryotic plankton diversity in the sunlit ocean », *Science*, n° 348, p. 1261605–1261605, 2015.

[VOL 16] VÖLKEL F., BAPTESTE E., HABIB M., LOPEZ P., VIGLIOTTI C., « Read networks and k-laminar graphs », *arXiv*, p. 1–14, 2016.

[WAL 11] WALTER J., LEY R., « The Human Gut Microbiome: Ecology and Recent Evolutionary Changes », *Annual Reviews of Microbiology*, n° 65, p. 411–429,2011.

[WAN 17] WANG Y., HATT JK., TSEMENTZI D., *et al.*, « Quantifying the importance of the rare biosphere for microbial community response to organic pollutants in a freshwater ecosystem », *Appl Environ Microbiol.*, n° 83(8), p. e03321-16, 2017.

[WEI 15] WEI F., WANG X., WU Q., « The giant panda gut microbiome », *Trends Microbiol.*, n° 23:p. 450-452, 2015.

[WU 08] WU M., EISEN JA., « A simple, fast, and accurate method of phylogenomic inference », *Genome Biol*, n° 9, p. R151–R151, 2008.

[WU 11] WU GD., CHEN J., HOFFMANN C., BITTINGER K., *et al.*, « Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes », *Science*, n° 334, p.105–108, 2011.

[WU 12] WU M., SCOTT AJ., « Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2 », *Bioinformatics*, n° 28, p.1033, 2012.

[XU 16] XU J., GALLEY JD., BAILEY MT., *et al.*, « The impact of dietary energy intake early in life on the colonic microbiota of adult mice », *Sci Rep*, n° 6, p. 19083, 2016.

[YAN 15] YÁÑEZ-RUIZ DR., ABECIA L., NEWBOLD CJ., « Manipulating rumen microbiome and fermentation through interventions during early life: A review », *Front Microbiol.*,n° 6, p. 1133, 2015.

[YAT 12] YATSUNENKO T., REY FE., MANARY MJ., *et al.*, « Human gut microbiome viewed across age and geography », *Nature*, n° 486, p. 222–227, 2012.

[ZEN 15] ZENG B., HAN S., WANG P., *et al.*, « The bacterial communities associated with fecal types and body weight of rex rabbits », *Sci Rep*, n° 5, p. 9342, 2015.

[ZHE 16] ZHENG J., XIAO X., ZHANG Q., *et al.*, « The programming effects of nutrition-induced catch-up growth on gut microbiota and metabolic diseases in adult mice », *Microbiologyopen*, n° 5, p. 296–306, 2016.

[ZHU 11] ZHU L., WU Q., DAI J., *et al.*, « Evidence of cellulose metabolism by the giant panda gut microbiome », *Proc Natl Acad Sci U S A*, n° 108, p. 17714–17719, 2011.

3. Le changement de régime alimentaire des *Podarcis sicula* est associé à des changements ciblés dans le microbiote

3.1 Études du microbiote : description des données

Pour analyser les microbiotes des *Podarcis sicula*, nous avons utilisé un marqueur de l'ARN ribosomique, la région V4 de l'ARNr 16S. Cette région mesure environ 250 paires de base (300 paires de bases chez *E. Coli*) (Yarza et al. 2014) . Bien que ce ne soit pas la seule région de l'ARN utilisée comme marqueur, l'utilisation de cette région est assez classique (Walters et al. 2016; Yarza et al. 2014). Cela permet, à partir du marqueur séquencé, d'associer à chaque read une assignation taxonomique de l'organisme auquel correspond le marqueur (Figure 8).

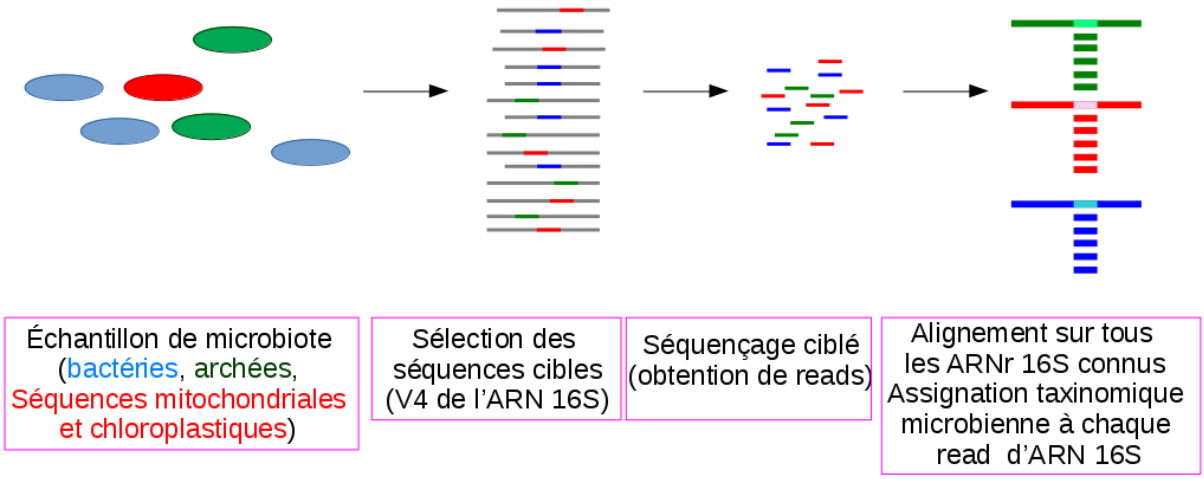


Figure 8 : Description de la méthode d'obtention de la région V4 de l'ARNr 16S.

En bleu sont représentées les bactéries et leur matériel génétique, en vert sont représentées les archées et leur matériel génétique, en rouge sont représentés les organelles des eucaryotes (chloroplastes et mitochondries) et leur matériel génétique.

Le jeu de données est constitué de 62 lézards. On s'intéresse aux caractéristiques suivantes pour chacun des lézards (Figure 9) :

- Le régime alimentaire (insectivore et omnivore), qui est la principale caractéristique que l'on souhaite étudier au cours de cette thèse.

- l’année d’échantillonnage (2014, 2015 ou 2016) qui permettra de détecter une éventuelle évolution du microbiote dans le temps.

- la saison d’échantillonnage (été ou printemps) afin de tester si la saisonnalité est corrélée avec un changement dans le microbiote.

- le genre du lézard. Nous souhaitons tester si les mâles et les femelles ont le même microbiote.

- L’insularité. Les lézards sur les îles ont une disponibilité en ressources alimentaires plus faible que celle des lézards du continent. Par ailleurs les conditions d’insularité dans le cas présent pourraient aussi avoir un impact sur la variabilité génétique des lézards. En effet, les îles Pod Kopište et Pod Mrčaru ayant une superficie d’environ 1km², il s’agit de petites populations de *Podarcis sicula*, contrairement aux populations présentes sur le continent.

- La localisation des lézards : cela correspond à l’origine des lézards (les îles Pod Kopište, Pod Mcraru, le continent : Split et Zagreb).

Variables	Groupes	Nombre de lézards	Nombre total de lézards
Année d'échantillonnage	2014	25	62
	2015	19	
	2016	18	
Saison déchantillonnage	Printemps	22	62
	Été	40	
Genre	Mâle	38	62
	Femelle	24	
Insularité	Île	29	62
	Continent	33	
Localisation	Pod Kopiste (île)	14	62
	Pod Mrcaru (île)	15	
	Continent	33	
Régime alimentaire	Insectivore	47	62
	Omnivore	15	

Figure 9 : Description du jeu de données sur lesquelles sont effectuées les analyses microbiote. Au sein de chaque caractéristique, le nombre de lézards n'est pas toujours équilibré.

Le jeu de données 16S pour ces 62 lézards est constitué de 5 493 157 reads (courtes séquences de quelques dizaines à quelques centaines de paires de bases), ce qui représente en moyenne environ 89 000 reads par échantillon. Ces reads sont des « paired end » (reads appariés) de 250 paires de bases.

Les données ont été filtrées et démultiplexées en utilisant QIIME (Caporaso et al. 2010; Kuczynski, Stombaugh, Walters, González, J Gregory Caporaso, et al. 2012) (avec le script split_libraries.py, paramètres par défaut). Ensuite, les chimères (séquences dues à l'amplification par PCR de plusieurs modèles de séquences ou séquences parentes) ont été enlevées (utilisation du script identify_chimeric_seq.py, qui est un script additionnel de contrôle de la qualité recommandé dans le tutoriel de QIIME (http://qiime.org/tutorials/). Ce script s'appuie sur usearch61 (Edgar 2010) et

sur la base de données greengenes v13.8 (DeSantis et al. 2006)). Suite à ces traitements, il reste 4 789 443 reads soit en moyenne environ 77 000 reads par échantillon. Par conséquent, 87,2% des reads ont été conservés. Enfin, après ces traitements, les reads ont été regroupés en OTUs avec un seuil de 97% .Les OTUs sont construites à l'aide du script QIIME pick_open_reference.py (http://qiime.org/scripts/pick_open_reference_otus.html), qui se base sur la méthode de clustering uclust (Edgar 2010). La première étape consiste à comparer les reads des microbiotes à la base de données de référence greengenes (DeSantis et al. 2006). Deux reads appartiennent à la même OTU s'ils présentent 97% d'identité avec une séquence de référence (pick closed reference). Les reads ne se retrouvant dans aucune OTU sont traités au cours d'une seconde étape. Au cours de cette seconde étape, un clustering *de novo* est appliqué sur les reads restants, et le read centroïde de chaque cluster est considéré comme "séquence de référence" et une nouvelle base de données de référence est construite. Ensuite, lors de la troisème étape, le clustering de l'étape 1 (pick closed reference) est appliqué sur les reads qui n'avaient pas été regroupés au sein d'OTUs à la fin de l'étape 1, en prenant comme base de données de référence, la base de données créée à l'étape 2. Enfin une quatrième étape construit des OTUs à l'aide de la méthode de clustering *de novo* implémentée dans QIIME. Les OTUs obtenues aux étapes 1, 2, 3 et 4 sont concaténées afin d'avoir un fichier complet contenant toutes les OTUs.

La fabrication des OTUs peut donc être simplement résumée par le schéma suivant (Figure 10) :

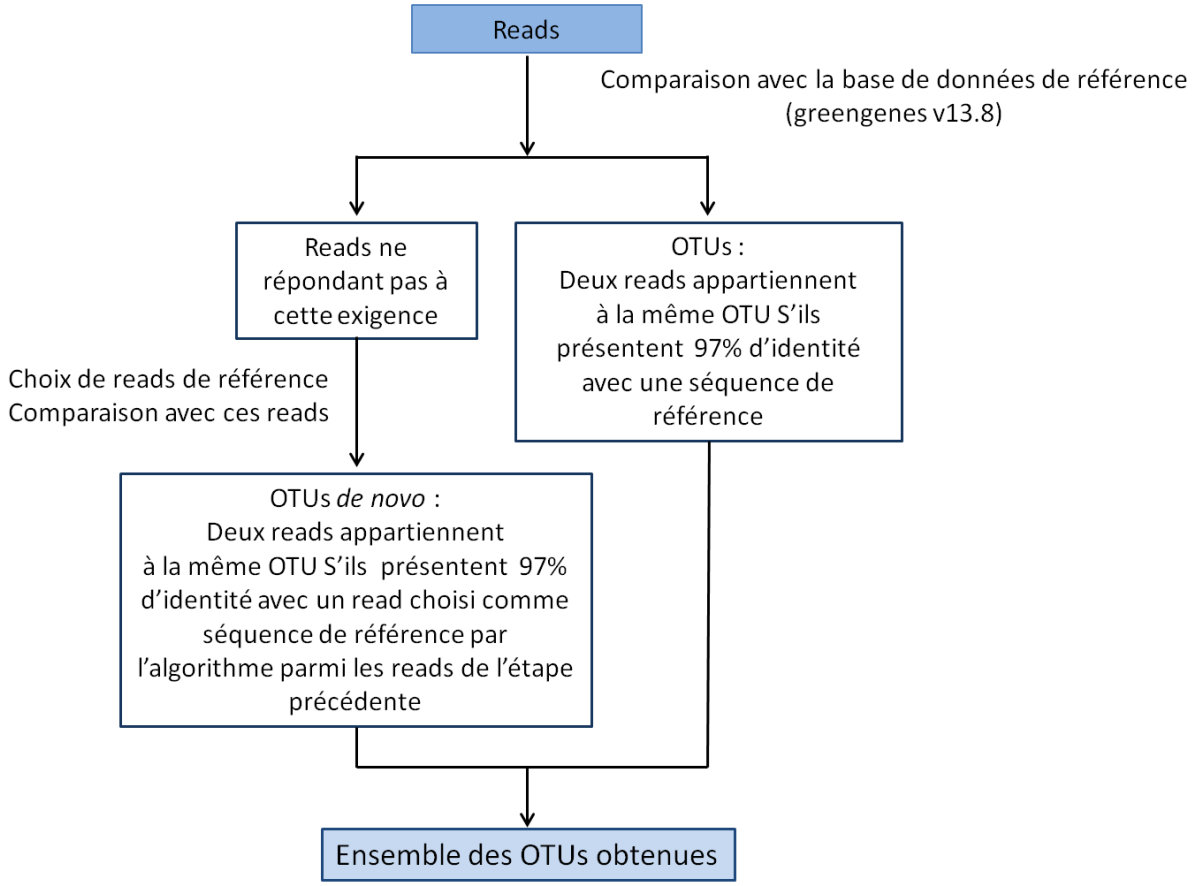


Figure 10 : Construction des OTUs à partir des reads à l'aide du script pick_open_reference.py.

3.2 Etudes du microbiote : une discipline engagée sur la phase II du sentier de dépendance (début de standardisation)

L'analyse des données ciblées correspond à une science engagée sur un sentier de dépendance. En effet, les analyses que l'on présentera par la suite sont assez standards.

Par exemple la diversité alpha, qui est une mesure classique de biodiversité à laquelle nous nous intéressons dans ce chapitre, peut se calculer à l'aide de différents indices. Les indices que nous présenterons ci-dessous (Shannon, Simpson, et Chao1) sont utilisés dans les études de diversité du microbiote (Bennett et al. 2013; Blasco et al. 2017; Yang et al. 2017). La méthode de calcul de la diversité bêta (distance de Bray-Curtis) est elle aussi standard, et sa visualisation à l'aide de la NMDS ("Non-metric MultiDimensional Scaling", appelée en français analyse multidimensionnelle

non métrique, méthode statistique permettant de visualiser les similarités entre les individus d'un jeu de données) est classique (Borcard, Gillet, and Legendre 2011).

3.3 Choix des analyses et des méthodes utilisées

3.3.1 Analyse de la diversité

La première analyse que l'on a souhaité réaliser est une analyse de la diversité taxonomique dans les microbiotes. Cela fait suite à l'article de Ruth Ley (Ley et al. 2008), dans lequel il a été démontré que le microbiote intestinal des mammifères herbivores présentait plus de diversité taxonomique que le microbiote intestinal des Mammifères omnivores, lui-même plus diversifié que le microbiote intestinal des carnivores. Il est donc intéressant d'étudier si chez *Podarcis sicula*, qui est un vertébré non mammifère, nous retrouvons ce type de résultat, à savoir une différence de diversité associée à une différence de régime alimentaire.

La diversité taxonomique est souvent mesurée à l'aide du concept de richesse spécifique du milieu donné (Whittaker 1960, 1972). Trois mesures de richesses spécifiques existent :

- la diversité alpha, qui correspond au nombre d'espèces (ou d'OTUs) qui coexistent dans un milieu donné (ici, l'intestin du lézard) (Whittaker 1960, 1972),

- la diversité bêta, qui correspond à la différence de diversité des espèces entre plusieurs milieux (Whittaker 1960) (par exemple, l'intestin des lézards insectivores et celui des lézards omnivores). Il s'agit donc de comparer le nombre de taxons (phyla, ou genres, ...) qui sont exclusifs et partagés entre des milieux que l'on compare.

- la diversité gamma, qui correspond au taux d'addition d'espèces lorsque l'on échantillonne le même milieu à différents endroits.

Dans le cas de notre étude, nous avons donc choisi de regarder la diversité alpha, qui permet de connaître la richesse spécifique d'un microbiote et la diversité bêta, afin de comparer la diversité entre nos échantillons, à l'échelle du genre et du phylum, pour les 62 individus.

La diversité alpha et la diversité bêta de chaque microbiote sont calculées à partir de la table d'abondances d'OTUs produites par QIIME. Dans ce cas, les variables en entrée sont les 32 850 OTUs.

3.3.1.1 Mesures de diversité alpha

L'indice de Shannon a été défini indépendamment par Claude Shannon (Shannon 1948) et par Norbert Wiener en 1948 (Wiener 1948). Cet indice représente à la fois le nombre d'espèces d'un milieu mais aussi la répartition des effectifs individuels au sein des espèces présentes. L'indice de Shannon se calcule ainsi (Shannon 1948):

$$H = -\sum_{i=1}^{n} P_i \ln P_i$$

où *Pi* représente l'abondance relative du taxon *i*.

Plus l'indice de Shannon est grand, plus la diversité taxonomique du microbiote est importante. Si cet indice est informatif sur la richesse en espèces et sur la répartition des effectifs individuels entre les espèces, il reste cependant des cas ambigus. En particulier, il est compliqué de différencier les cas de figures suivants, en se basant uniquement sur l'indice de Shannon :

- le milieu contient un grand nombre d'espèces à faibles effectifs,

- le milieu contient un petit nombre d'espèces très abondantes.

Le second indice utilisé est l'indice de Simpson. Cet indice a été défini par E.H. Simpson en 1949 (Simpson 1949). Il s'agit une mesure de régularité, cela signifie qu'il mesure la probabilité que deux individus pris au hasard appartiennent à la même espèce. La formule de l'indice de Simpson est la suivante :

$$D = \sum_{i}^{n} N_i \ (N_i - 1)/(N - 1)$$

avec *Ni* le nombre d'individus de l'espèce *i*, et *N* le nombre total d'individus

Enfin, le troisième indice de diversité alpha que l'on a souhaité utiliser est l'indice Chao1, parce qu'il tient davantage compte des espèces peu abondantes. Il est utilisé depuis 1984. Cet indice estime le nombre d'espèces non observées à partir de celles qui n'ont été observées qu'une ou deux fois. Cet indice de diversité est un estimateur minimum. Pour qu'il soit adapté au jeu de données, il est nécessaire que les singletons et les doublons représentent une part importante de l'information (i.e. que les taxa peu abondants soient nombreux). La formule permettant de calculer l'indice Chao1, S_1^* est le suivant (Colwell , Robert K., Coddington 1994) :

$$S_1^* = S_{obs} + (\frac{a^2}{2b})$$

où S_{obs} est le nombre de taxa observés dans l'échantillon, *a* le nombre d'espèces représentées par un seul individu dans l'échantillon, et *b* le nombre d'espèces représentées par exactement deux individus dans l'échantillon.

D'autres estimateurs de diversité alpha existent (Chao2, ACE, ICE,..) (Gotelli and Colwell n.d.), cependant les trois indices présentés ci-dessus nous permettent d'appréhender la diversité alpha avec suffisamment de précision.

3.3.1.2 Mesures de diversité bêta

Il existe plusieurs façons de calculer la diversité bêta. D'un point de vue général, et quelle que soit la manière choisie pour calculer cet indice, il est important de noter que la diversité bêta est une mesure de dissimilarité entre deux échantillons.

La mesure de diversité bêta la plus utilisée est la formule suivante, proposée par Whittaker en 1960 (Whittaker 1960) :

$$\beta = \frac{S}{\alpha - 1}$$

où *S* est le nombre total d'espèces, et *α* le nombre moyen d'espèces par échantillon.

Le problème de cet indice est que la diversité bêta augmente avec la taille de l'échantillon. Une solution consiste à utiliser la distance de Sorensen décrite par Watson en 1966 (WATSON, WILLIAMS, and LANCE 1966), qui considère *a*, le nombre d'espèces partagées par les deux sites et *b* et *c*, le nombre d'espèces uniques (non partagées). Alors S = a + b + c et α= (2a + b + c)/2. La formule de la diversité bêta devient alors :

$$\beta = (b + c)/(2a + b + c)$$

La formule de diversité bêta que l'on a choisie est la formule de Bray-Curtis, parce qu'elle est couramment utilisé en écologie microbienne (Borcard et al. 2011). La distance de Bray-Curtis (définie par Bray & Curtis en 1957)(Bray and Curtis 1957) entre un premier microbiote x_1 et un deuxième microbiote x_2 est

$$\beta(x_1, x_2) = \frac{\sum_{j=1}^{p} |y_{1j} - y_{2j}|}{\sum_{j=1}^{p} (y_{1j} + y_{2j})}$$

où y_{1j} est l'abondance de l'espèce j dans le microbiote x_1 et y_{2j} est l'abondance de l'espèce j dans le microbiote x_2. P est le nombre d'espèces total. La distance de Bray-Curtis est en réalité un indice de dissemblance et non une mesure de distance, du fait qu'elle ne respecte pas l'inégalité triangulaire. Cet indice est compris entre 0 (les deux microbiotes x_1 et x_2 ont la même composition), et 1 (les deux microbiotes x_1 et x_2 sont totalement dissemblables). L'utilisation de cet indice nécessite que les microbiotes soient de même taille dans la mesure où il se calcule sur des abondances absolues et non relatives (Borcard, Gillet, and Legendre 2011).

Ensuite, nous avons souhaité représenter cette diversité bêta, et pour cela nous avons choisi la NMDS (pour « Non Metric MultiDimensional Scaling » soit analyse multidimensionnelle non métrique en français), qui est une méthode proposée par Kruskal en 1964 (Kruskal 1964). La NMDS consiste à représenter sur un plan (2D) des observations faites dans un espace de dimensions plus grandes, de telle façon que les distances entre paires d'observations soient aussi proches que possible sur le plan

que dans l'espace d'origine. Cette méthode consiste donc à réduire au maximum le niveau de stress, c'est-à-dire la différence entre les distances 2D et les distances d'origine, exprimé par la formule suivante :

$$Stress = \sqrt{\frac{\sum_{h,i}(d_{hi} - \hat{d}_{hi})^2}{\sum_{h,i} d_{hi}^2}}$$

où d_{hi} est la distance entre les échantillons h et i, et $\hat{d}_{hi}$ est la distance prédite par la régression.

Une NMDS adaptée au jeu de données, renvoie une valeur de stress faible. On considère que la valeur de stress doit être inférieure ou égale à 0.2 pour que la NMDS soit de qualité suffisante (Borcard, Gillet, and Legendre 2011).

Ensuite, afin d'analyser si la NMDS sépare bien les groupes en fonction du régime alimentaire (ou des 5 autres variables étudiées présentées Figure 9), nous avons appliqué une anosim sur la matrice de dissimilarité obtenue avec la NMDS, parce que cela permet d'avoir un test de la significativité des résultats obtenus par la NMDS (Buttigieg and Ramette 2014). L'anosim est une analyse des similarités. Il s'agit d'un test non paramétrique. Ce test a été proposé pour la première fois en 1993 par K.R. Clarke (CLARKE 1993). L'anosim fournit deux informations importantes : la statistique R et la Pvalue. La statistique R est calculée de la façon suivante :

$$R = \frac{r_B - r_W}{M/2}$$

où r_B correspond à la moyenne des rangs des similarités des paires d'échantillons et r_W correspond à la moyenne des rangs des similarités des paires d'échantillons au sein d'un groupe (par exemple le groupe des lézards insectivores), et M = n(n-1)/2, où n est le nombre d'échantillons.

La statistique R est comprise entre -1 et 1. Elle permet de savoir si la distance entre les groupes (par exemple microbiotes de lézards insectivores et microbiotes de lézards omnivores) calculée lors de la NMDS est plus importante que la distance intra-groupe (calculée lors de la NMDS). Plus R est proche de 1, plus les deux groupes sont distants l'un de l'autre. Si R est proche de zéro, alors les deux groupes sont assez similaires, et peu distinguables. Enfin, si R est proche de -1, cela signifie que les

différences entre individus intra-groupes, sont plus importantes que les distances inter-groupes. Enfin, la Pvalue nous indique si l'on peut considérer que R est fiable ou non. Si la Pvalue < 0.05, on accepte les résultats proposés par la statistique R.

3.3.2 Analyse de la composition du microbiote

3.3.2.1 Présence d'un microbiote ubiquitaire chez Podarcis sicula

Le microbiote ubiquitaire est la fraction du microbiote présente en commun dans tous les individus. Elle permet de mesurer la part variable du microbiote et la part universelle du microbiote chez *Podarcis sicula*. On a choisi de déterminer le microbiote ubiquitaire à partir de la table d'OTUs à 97%, parce que cela permet, en première approximation, de se placer au niveau de l'espèce.

Au-delà de la répartition des OTUs dans la population, l'abondance des OTUs est une information intéressante. Combien d'OTUs sont très abondantes ? Combien d'OTUs sont rares ? On considère une OTU abondante dès lors qu'elle contient 10% des reads (ou plus) contenus dans un microbiote. La définition d'une OTU abondante n'est pas encore figée et varie d'une étude à une autre. Certaines études considèrent les 10 OTUs les plus abondantes (Ye 2011), d'autres considèrent qu'une OTU est abondante si elle contient au moins 1% des reads d'un microbiote (Kageyama et al. 2017), enfin, certaines études considèrent une OTU abondante dès lors que l'OTU contient au moins 0.1% des reads d'un échantillon (Lau et al. 2016). Les réponses aux questions relatives à l'abondance des OTUs orientent le choix du type d'analyse que l'on souhaite appliquer aux données. En effet, si les OTUs peu abondantes sont nombreuses, on va davantage s'intéresser à des méthodes statistiques permettant de les prendre en compte (telles que l'analyse linéaire discriminante par exemple), alors que si les OTUs peu abondantes sont peu nombreuses, on peut se permettre de ne pas ou peu les prendre en compte dans les analyses.

Enfin, en couplant les données de présence et d'abondance des OTUs précédemment recueillies, on peut calculer la prévalence. La prévalence d'une OTU correspond au nombre de microbiotes de lézards contenant cette OTU. Ce couplage d'informations permet de savoir si les OTUs du microbiote ubiquitaire sont abondantes

ou non, si les OTUs peu abondantes sont présentes chez beaucoup de lézards ou spécifiques de leur hôte.

3.3.2.2 Présence d'entérotypes chez le Podarcis sicula

Après avoir étudié la répartition des OTUs au sein des microbiotes de lézards, nous avons souhaité savoir si la population de lézards peut être structurée en fonction de sa composition taxonomique, en recherchant des entérotypes. Dans le but de pouvoir comparer nos résultats avec ceux obtenus par la communauté(Arumugam et al. 2011; De Filippo et al. 2010; Knights et al. 2017; J. Li et al. 2015; Lim et al. 2014; Moeller et al. 2012; Wu et al. 2011), nous avons suivi le pipeline proposé par Arumugam et Raes (Arumugam et al. 2011), adapté à l'environnement R et mis en ligne. Ainsi, nous avons bénéficié du fait que les études de microbiote ciblé sont engagées sur un sentier de dépendance.

La détection d'entérotypes a été réalisée au niveau du phylum et au niveau du genre. Historiquement, le concept d'entérotypes était développé au niveau du genre, cependant nous avons souhaité remonter au niveau du phylum afin de déterminer si l'on pouvait créer des groupes à un niveau taxonomique plus général. En entrée est fournie à R la table d'abondance relative des phyla (Figure 11) ou des genres en fonction de l'analyse que l'on souhaite réaliser.

	PSMF27DIGC	PSKF20MDI	PSS37DIGC	PSZ41MDI	PSSF18MDI	PSS32MDI	PSM52DIGC	PSS20MDI	PSMF33DIGC	PSSF8MDI	PSMF32DIGC	PSS21MDI	PSSF21MDI
Crenarchaeota	0,001363413	0,0002640264	0,0009681663	0,0011174994	0,0030645698	0,0010450631	0,0010570452	0,000101814	0,0010055249	0,000602752	5,33E-005	0,0002429937	7,91E-005
Euryarchaeota	0,0165511993	0,002310231	6,31E-005	9,72E-005	0,0001187818	4,18E-005	0,0034882492	0,000101814	0,1256685083	5,17E-005	0,0023144198	8,10E-005	7,91E-005
Acidobacteria	0	0	0	0	7,13E-005	0	0	0	0	0	0	0	0
Actinobacteria	0,019420708	0,0036138614	0,0504814522	0,1160093935	0,0931249109	0,0393988797	0,0160847046	0,006906382	0,0211160221	0,0076463396	0,0065059727	0,0169555592	0,0163721784
Bacteroidetes	0,2827338015	0,2986963696	0,3729860563	0,1696007774	0,1178790326	0,0648775186	0,2647193545	0,3356807113	0,0263425414	0,1675133897	0,0502133106	0,1377234192	0,2571934764
Chlamydiae	0	8,25E-005	0	0	0,0356466111	0,0028032499	0,0001045063	1,76E-005	3,39E-005	0	0,0001205504	0	0,0004589881	0,0001265482
Chloroflexi	0,0007609747	4,95E-005	0,0002946593	0,0002915216	0,000736447	0,0017139035	0,0004228181	0,000186659	0,0003646409	0,0001205504	0,0001493174	0,0003239916	0,0001740038
Cyanobacteria	0,0005390237	0,0014026403	0,0015048671	0,011353146	0,0006889343	0,0100953098	0,000211409	0,0005939149	0,0003756906	0,0001033289	0,0007145904	8,10E-005	0,0004429188
Deferribacteres	9,51E-005	8,25E-005	0,0002736122	0,0001133695	9,50E-005	0,0003344202	8,81E-005	5,09E-005	8,84E-005	0,0001205504	0,0002239761	0	0,0007592894
Elusimicrobia	0,0012365839	0,001039604	0,0027255985	0	9,50E-005	0	0,000669462	0	7,73E-005	1,72E-005	9,60E-005	2,70E-005	3,16E-005
Firmicutes	0,5051286523	0,4637293729	0,3584004209	0,5575512187	0,6214187295	0,6652871833	0,5254219372	0,4502893214	0,6522762431	0,6231250108	0,7555674061	0,3344943031	0,5506588418
Fusobacteria	0,0001426828	0,0003465347	0,0001368061	0,0004696737	0,0006176652	0,0005016303	0,0007223142	0,0005260389	6,63E-005	0,0016015982	0,0001493174	0,0004859874	0,0006485597
Gemmatimonadetes	0	0	0	0	0	0	0	0	1,10E-005	0	0	0	0
Lentisphaerae	3,17E-005	1,65E-005	0	0	0	0	1,76E-005	0	1,10E-005	0	0	0	0
Planctomycetes	0,0012682911	3,30E-005	0,0001999474	0,0001295651	0,0008552288	0,0008151492	0,0006166097	1,70E-005	0,0019668508	0,0001205504	0,0004372867	0,0001619958	1,58E-005
Proteobacteria	0,1113559618	0,098960396	0,1542751907	0,0522633412	0,1200646173	0,1963046568	0,0585955393	0,1182569446	0,1218232044	0,0350629445	0,1362734642	0,4553431611	0,073208156
Spirochaetes	0,0004121946	8,25E-005	0,0001894238	4,86E-005	7,13E-005	4,18E-005	0,0047214686	1,70E-005	0,0002651934	1,72E-005	0,0004799488	0,0041848912	3,16E-005
Synergistetes	0	0	0	0	0	0	0	0	0	0	0	0	0
TM7	0	0	0	0	0	2,09E-005	0,0001761742	0	0,0005745856	0	0	0	0,0002530965
Tenericutes	0,014632909	0,0160891089	0,0149750066	0,0214106405	0,033140115	0,0122690411	0,0175821853	0,0028847296	0,0159337017	0,0032031963	0,0140465017	0,0073168098	0,0132242909
Verrucomicrobia	0,0403633654	0,1106930693	0,0319073928	0,0308850919	0,0030645698	0,0020065212	0,0979704732	0,0802972968	0,0293922652	0,1518590594	0,0294581911	0,0397969653	0,0787288229

Figure 11 : Exemple de table d'abondances relative au niveau du phylum.

Les lignes correspondent aux phyla présents dans les microbiotes, et les colonnes correspondent aux microbiotes de lézards.

Les variables de l'analyse au niveau du phylum sont les 21 phyla (1ere colonne de la Figure 11) présents dans les microbiotes de lézards (Crenarchaeota, Euryarchaeota, Acidobacteria, Actinobacteria, Bacteroidetes, Chlamydiae, Chloroflexi, Cyanobacteria, Deferribacteres, Elusimicrobia, Firmicutes, Fusobacteria, Gemmatimonadetes, Lentisphaerae, Planctomycetes, Proteobacteria, Spirochaetes, Synergistetes, TM7, Tenericutes, Verrucomicrobia), les abondances relatives des 62 échantillons sont renseignées à partir de la deuxième colonne. Les variables au niveau du genre sont les 291 genres présents dans les microbiotes des lézards. Nous avons supprimé des analyses la fraction de reads non assignés. Elle représente 0.9% au niveau du phylum et 49% au niveau du genre.

L'objectif de l'analyse est de regrouper des échantillons en se basant sur leur composition taxonomique (donc sur les variables). Pour cela, on utilise des méthodes dites de « clustering », basées sur des calculs de distance.

Les échantillons sont regroupés en utilisant la méthode de clustering utilisée dans l'article décrivant les entérotypes pour la première fois (Arumugam et al. 2011). Cette méthode est basée sur la distance de JSD (Jensen-Shannon Divergence) (Endres and Schindelin 2003) et l'algorithme de Clustering PAM (Partitioning Around Medoids) (Kaufman and Rousseeuw 1987), détaillés ci-dessous.

La première étape avant d'appliquer l'algorithme de clustering, est de déterminer le nombre optimal de groupes (i.e. de « clusters »). Pour cela, on utilise l'indice de *Calinski-Harabasz (indice CH)* (Caliński and Harabasz 1974)*. Cet indice a pour formule :*

$$\mathrm{CH_k} = \frac{\frac{\mathrm{B_k}}{\mathrm{k-1}}}{\frac{\mathrm{W_k}}{\mathrm{n-k}}}$$

Dans cette formule, $\mathrm{B_k}$ correspond à la somme des carrés des distances entre les groupes (c'est à dire les distances au carré entre tous les points *i* et *j*, avec *i* et *j* n'appartenant pas au même cluster). $\mathrm{W_k}$ correspond à la somme des carrés des distances au sein des groupes (c'est à dire les distances au carré entre tous les points *i* et *j*, avec *i* et *j* appartenant au même cluster). Dans cette formule, *k* est le nombre de

clusters. Cette formule traduit donc le fait que le « clustering » est plus robuste lorsque la distance entre les groupes (B_k) est plus importante que la distance au sein des groupes (W_k). Le choix du nombre optimal de groupes *k* se porte sur le *k* associé au CH_k maximal. D'un point de vue graphique, le nombre de clusters (ou groupes) optimal est représenté par un pic d'indice *CH* (Figure 12).

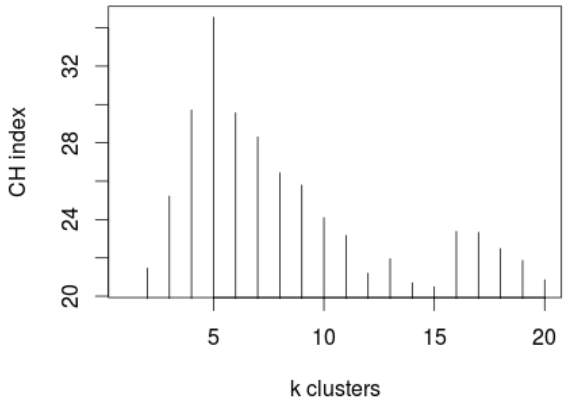


Figure 12 : Choix du nombre de clusters dans l'analyse des entérotypes à l'aide du graphique représentant l'indice CH en fonction du nombre de clusters.

Une fois le choix du nombre optimal de groupes determiné, on peut s'intéresser aux méthodes de « clustering » et au calcul de distance permettant de regrouper les microbiotes individuels par entérotype. Le calcul de distance contribuant à ces regroupements est une méthode se basant sur la *JSD*. La *JSD* est une mesure finie et positive qui doit satisfaire les deux conditions suivantes :

$$JSD(a,b) \geq 0$$

$$JSD(a,b) = 0 \; si \; et \; seulement \; si \; a = b$$

Cette divergence est symétrique, c'est à dire que *JSD(a,b)* = *JSD(b,a),* cependant, la *JSD* ne satisfait pas l'inégalité triangulaire, de ce fait, elle ne peut pas être considérée comme une réelle métrique. C'est pour cette raison que la métrique utilisée dans l'algorithme permettant de déterminer des entérotypes est la racine carrée de la *JSD* (métrique qui respecte l'inégalité triangulaire et est considérée comme une vraie mesure de distance) (Endres and Schindelin 2003).

La mesure de distance utilisée dans l'algorithme est donc la suivante :

$$D(a,b) = \sqrt{JSD(p_a, p_b)}$$

Dans cette formule, p_a et p_b sont les distributions d'abondances des échantillons *a* et *b*, et JSD(x,y) est la divergence de Jensen-Shannon entre deux distributions de probabilités x et y, définie telle que :

$$JSD(x,y) = \frac{1}{2} KLD(x,m) + \frac{1}{2} KLD(y,m)$$

Dans cette équation, $m = \frac{(x+y)}{2}$ et KLD est la divergence de *Kullback-Leibler* entre x et y. Cette mesure de divergence est définie de la façon suivante :

$$KLD(x,y) = \sum_i x_i log \frac{x_i}{y_i}$$

Une fois que le calcul de distance a été effectué, nous pouvons appliquer l'algorithme de « clustering » PAM (« Partitionning Around Medoids »). Un médoïde est le représentant le plus central d'une classe. Cet algorithme se base sur l'algorithme des « k-means », mais possède en plus l'avantage de pouvoir utiliser n'importe quelle mesure arbitraire de distance. Cet algorithme est plus robuste que l'algorithme des « k-means » notamment vis à vis des données aberrantes. Il s'agit d'une procédure supervisée dans la mesure où le nombre de « clusters » est un des paramètres de l'algorithme, que l'on fournit au préalable, afin de créer le nombre de groupes que l'on a choisi (à l'aide de l'indice CH).

Après la création des groupes d'entérotypes, une étape de validation des groupes est nécessaire, afin d'évaluer la qualité du « clustering ». La technique utilisée pour cette étape est la validation de silhouette. La largeur de la silhouette S(i) des points de données individuels *i* est calculée en utilisant la formule suivante :

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

où *a(i)* est la moyenne de la dissimilarité (ou de la distance) de l'échantillon *i* aux autres échantillons du même cluster, et *b(i)* est la moyenne de la dissimilarité (ou distance) de tous les objets dans le cluster le plus proche de celui auquel appartient *i*.

S(i) est comprise entre -1 et 1. Un échantillon plus proche des échantillons de son propre cluster que des autres clusters aura une valeur de *S(i)* élevée (proche de 1), alors qu'une valeur proche de zéro implique que l'échantillon *i* se situe plutôt entre deux groupes. Enfin, des valeurs de *S(i)* très négatives indiquent que l'échantillon a été assigné au mauvais cluster.

Enfin, les entérotypes peuvent être visualisés de deux façons différentes : soit à l'aide d'une PCoA (« Principal Coordinates Analysis »), soit à l'aide d'une BCA (« Between Class Analysis »). Nous avons choisi de visualiser nos entérotypes à l'aide de la BCA. La BCA se traduit en français par « l'analyse inter-classe ». Cette dernière analyse est un cas particulier d'Analyse en Composante Principale (ACP), qui insiste plus que l'ACP sur la différence entre les groupes. On pourrait décrire simplement la BCA comme étant une ACP appliquée sur les moyennes par groupe (ici les entérotypes) et par variable. Dans cette analyse, le critère optimisé est la variance inter-classe (*cf* https://pbil.univ-lyon1.fr/R/pdf/jacquet-prodon.pdf). Les résultats que nous avons obtenus sont présentés dans la partie 3.4.

3.3.2.3 Identification des taxa associés au changement de régime alimentaire

Après avoir étudié la structuration de la population de *Podarcis sicula* en fonction de sa composition taxonomique, nous avons voulu identifier s'il existait des taxa associés à chacun des deux régimes alimentaires. Afin de pouvoir prendre en compte les différences taxonomiques et d'abondances entre insectivores et omnivores, que les taxa présentent de faibles abondances ou des abondances importantes, on a choisi d'effectuer une LefSE (il s'agit d'une Analyse Linéaire Discriminante LDA qui prend en compte l'effet taille des populations microbiennes). LefSe a été développé par le groupe Huttenhower afin de pouvoir détecter des bio-marqueurs. Le choix de cette méthode permet de travailler sur des comparaisons de classes à haute dimension (Segata et al. 2011).

Le principe de la LDA est d'identifier les variables permettant de discriminer plusieurs groupes d'individus. Dans le cadre de notre étude, on souhaite discriminer les microbiotes de lézards insectivores et des microbiotes de lézards omnivores. Afin d'avoir une meilleure compréhension du jeu de données nous avons aussi décidé de discriminer les mâles des femelles, les lézards échantillonnés en 2014, 2015 et 2016, les lézards échantillonnés au printemps ou en été, les lézards de Pod Kopište, de ceux de Pod Mrčaru de ceux du continent. Nous avons choisi de nous placer au niveau du phylum et au niveau du genre

3.3.2.4 Identification des variables permettant de construire un modèle expliquant les tables d'abondance taxonomiques

Après avoir cherché à identifier quels phyla (et genres) permettent de discriminer les microbiotes de lézards en fonction de leur régime alimentaire (mais aussi en fonction des autres variables détaillées dans la Figure 9), nous avons souhaité traiter le problème inverse, et nous nous sommes demandé si à partir des variables, il était possible de créer un modèle permettant d'expliquer les tables d'abondances au niveau du phylum ou du genre. Notre choix d'analyse s'est porté sur l'Analyse de Redondance (RDA). La RDA est une analyse d'ordination classique en écologie. Elle a été développée en 1977 par Van Den Wollenberg (van den Wollenberg 1977). La RDA permet d'étudier les relations entre deux tableaux X et Y (ici, un tableau contenant pour chaque microbiote de lézard les variables suivantes : régime alimentaire, genre, saison et année d'échantillonnage, insularité et localisation, et un deuxième tableau contenant pour chaque microbiote de lézard les abondances par phylum ou par genre). Nous avons choisi cette analyse, parce qu'elle permet aussi de prendre en compte les variables de façon partielle (i.e. si deux variables ne sont pas indépendantes ou si l'on souhaite supprimer certaines variables de l'étude, la RDA partielle permettra de ne pas prendre en compte l'effet de ces variables. L'analyse de la redondance étant dissymétrique, l'effet d'une variable sur une autre n'est pas réciproque, il faut donc effectuer les tests dans les deux sens). Ces analyses n'ont pas permis de construire un modèle concluant (pouvoir prédictif des modèles élaborés trop faible).

3.4 Le régime alimentaire et la provenance géographique de populations sauvages de lézards impacte leur microbiote intestinal au niveau des taxa rares (article n°1).

Nous présentons ici l'article intitulé « Diet and geography impact the rare gut microbiota of natural population of lizards », que nous avons rédigé dans le but de le soumettre au journal Microbiome. Dans la mesure où cet article est encore en préparation, le formatage n'est pas définitif.

Diet and geography impact the rare gut microbiota of natural populations of lizards

Chloé Vigliotti[1, 2], Scott Dowd[3], François-Joseph Lapointe[4], Zoran Tadić[5], Beck Wehrle[6], Donovan German[6], Anthony Herrel[2], Philippe Lopez[1], Eric Bapteste[1]

[1] UMR 7138 Evolution Paris Seine, Université Pierre et Marie Curie, 75005 Paris, France – EB is supported by an ERC grant FP7/2007-2013 Grant Agreement # 615274

[2] UMR 7179, CNRS/MNHN, Département d'Ecologie et de Gestion de la Biodiversité, Paris Cedex, France

[3]MR DNA (Molecular Research LP), 503 Clovis Rd, Shallowater, TX 79363

[4]Département de Sciences Biologiques, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal (QC) H3C3J7 Canada

[5]Department of Animal Physiology, University of **Zagreb, Rooseveltov trg 6**, 10000 Zagreb, Croatia.

[6]Department of Ecology and Evolutionary Biology, 5309 McGaugh Hall (Lab), University of California, Irvine, CA 92697, USA.

Keywords: methanogens, enterotypes, biomarkers, microbiome, 16SrRNA

Abstract:

Background: Fourty-six years ago, ten lizards from the species *Podarcis sicula* were introduced on the island of Pod Mrčaru. Since the introduction, these populations underwent a significant dietary shift with their descendants having become omnivorous (eating up to 80% plant material at the end of summer). In addition these lizards present morphological variations associated with the consumption of plants including caecal valves. Potential variation in their gut microbiota has, however, not been investigated. To elucidate the possible impact on and of the gut microbiota in this recent adaptation we compared the microbiota (V4 regions of the 16S rRNA) of the gut microbial communities of 62 *Podarcis sicula*, sampled at different locations, from 2014 to 2016.

Results: *Podarcis* microbiota are host to numerous unassigned OTUs at the genus level, five of which, assigned to methanogenic phyla, possibly correspond to novel genera of methanogens. Alpha- and beta-diversity analyses of 4,789,443 cleaned reads unraveled that diet and geography have impacted rare phyla and genera of these lizards gut microbiota, but that there was no global shift in the prevalent and dominant species in the gut microbial

communities of *Podarcis sicula.* The general alpha-diversity of gut microbiota was, however, higher in omnivorous lizards than in insectivorous lizards, consistent with prior findings indicating that omnivorous hosts harbour a richer diversity than carnivorous hosts. Analyses aiming at detecting enterotypes cluster the gut microbiota of these lizards into five groups at the phylum level, but these clusters do not correspond to canonical enterotypes since the microbial communities within each group are heterogeneous and with different dominant phyla. Rather than conventional enterotypes, our study reports the existence of some biomarkers of omnivory, i.e. methanogenic Euryarchaeota, Spirochaetes, and Planctomycetes.

Conclusions: The dietary shift that occurred in the introduced population of *Podarcis sicula* has had a limited overall impact on the microbiota in the sense that it affected rare rather than abundant microbial taxa. These changes occurred in few generations, and enhanced the gut microbial diversity of omnivorous lizards with respect to their insectivorous relatives. These rare taxa, however, correspond to phyla previously described for their potential in plant and fiber digestion. This observation is compatible with the notion that targeted changes in the microbiota can be adaptive and enhance the diversity within plant consuming hosts. Our work also shows that the gut microbiota of non-model poikilotherm vertebrates remains largely underexplored.

Background

Fourty-six years ago, insectivorous lizards from the species *Podarcis sicula* were introduced on the island of Pod Mrčaru. Nevo and colleagues (1) designed a study to analyze the competition between *Podarcis sicula* and *Podarcis melisellensis* on islands. They introduced ten *Podarcis sicula* from the island of Pod Kopište to Pod Mrčaru, and ten *Podarcis melisellensis* from Pod Mrčaru to Pod Kopište. Since this experiment, *Podarcis melisellensis* has disappeared from Pod Mrčaru, while the descendants of the ten *Podarcis sicula* introduced on Pod Mrčaru underwent a significant dietary shift (2). These lizards have become largely omnivorous (up 80% plant matter in the diet in summer) and present significant morphological variations associated with the consumption of plants including the presence of caecal valves (2).

Diet, however, is known to affect more than host physiology and morphology. It also impacts, and is impacted by, the composition of the gut microbiota. The microbial taxa hosted in the gut change during development and aging of animals, including humans (3–6). The microbiota also change over larger time scales (7), in relation to diet. For example, phylogeny correlates with variations in the composition and abundance of the taxa of the gut microbiota in mammals (7–9) and termites (10). Despite these changes, several authors have proposed that within a species (and sometimes, within a higher-level lineage) the gut microbiota of various hosts can be rather similar. This led to the debated notion of enterotypes, which are groups of similar gut microbial communities, reported in humans (11, 12), primates (13, 14), and bees (15). However, this notion is criticized because the structuring of gut microbiota into

discrete clusters depends on the methodology and data investigated (16–18). These important questions regarding the potential structure of the gut microbial communities and their actual drivers are usually tackled through alpha- and beta-diversity analyses of microbiomes. Alpha-diversity analyses inform about taxonomic diversity, e.g. the number and abundance of species which coexist in an environment (here, the gut of *Podarcis sicula*) (19–23), whereas beta-diversity analyses characterize the difference in species diversity between two samples (20) (for example here, between the gut microbiota of insectivorous lizards, and those of omnivorous lizards). Linear discriminant analyses with size effect analyses are typically used to identify the specific taxa that distinguish between groups of samples (24–27), and redundancy analyses to identify which environmental or hosts conditions might affect the composition of the microbiota.

However, there are few studies on potential variations in the structure of microbial gut communities of lizards in general (28–31), and potential variations in the structure of microbial gut communities of *Podarcis sicula* in particular, have not been investigated, since these poikilotherm organisms fall outside the group of more commonly investigated model organisms (32). To elucidate the possible impact on the gut microbiota of the observed changes in diet in a natural lizard population, we compared the microbiota (V4 regions of the 16S rRNA) of the gut microbial communities of 62 *Podarcis sicula*, sampled at different locations, between 2014 and 2016. We report that the gut microbial communities of *Podarcis sicula* are poorly known, and weakly structured into non-conventional enterotypes. *Podarcis sicula* contain many unassigned OTUs at the genus level, including novel genera of methanogens, and the proportion of core and abundant OTUs (shared and prevalent in a majority of lizards) is low. There was no global shift in the prevalent and dominant species of gut microbial communities in these lizards. Nonetheless, diet and geography significantly impacted the gut microbiota. Overall, omnivorous lizards harboured more OTUs than insectivorous ones. Moreover, rare phyla and little abundant genera were affected by the host diet and location. Some of these taxa (i.e. methanogenic Euryarchaeota, Spirochaetes and Planctomycetes) possibly impact in turn their hosts, pointing towards potential biomarkers of dietary changes.

Material & Methods

Sample collection

Sixty-two lizards of the species *P. sicula* were collected between 2014 and 2016 (26 in 2014, 19 in 2015, 18 in 2016), providing us with 14 insectivorous lizards from Pod Kopište (8 males and 6 females), 15 omnivorous *lizards* from Pod Mrčaru (9 males and 6 females), and 33 lizards from the continent (21 from Split, 12 from Zagreb). Animals were collected in the field under permits of the Croatian government and euthanized using an intramuscular injection of pentoparbital. Gut microbial communities were sampled from the hindgut or distal intestine (DI). We identified this region of the gut as approximately 2 mm before the bulge to the cloaca. To sample it, we removed the distal intestine and squeezed the contents out onto a chilled, RNAse free surface with a flat tool (the back of a razor blade). Each sample was frozen immediately at −80°C.

Diet quantification

Diet was quantified for lizards from all populations by stomach flushing (see (33)) or by removing the gut content from specimens that were euthanized for analyses of the gut microbiome and physiology. Gut contents were stored in vials with 70% ethanol and analysed using a binocular scope (Leica). Contents were identified to the level of order or family when possible for insects and all plant matter was assigned to the following categories: fruits, seeds, leaves and woody material (2).

16S rRNA gene sequencing and processing

Samples were kept at −80°C until DNA extraction. We used PCR primers 515/806 with barcode on the forward primer to sequence the 16S rRNA gene V4 variable region in a 30 cycle PCR using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94°C for 3 minutes, followed by 28 cycles of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed. After amplification, PCR products are checked in 2% agarose gel to determine the success of amplification and the relative intensity of bands. Multiple samples are pooled together (e.g., 100 samples) in equal proportions based on their molecular weight and DNA concentrations. Pooled samples are purified using calibrated Ampure XP beads. Then the pooled and purified PCR product is used to prepare DNA library by following Illumina TruSeq DNA library preparation protocol. Sequencing was performed at MR DNA (www.mrdnalab.com, Shallowater, TX, USA) on a MiSeq following the manufacturer's guidelines. Sequence data were processed using MR DNA analysis pipeline (MR DNA, Shallowater, TX, USA). In summary, sequences were joined, depleted of barcodes then sequences <150bp removed, sequences with ambiguous base calls removed. We cleaned the reads using the QIIME software package (34, 35) (with split_libraries.py, deleting chimera with identify_chimeric_seqs.py using usearch61 (36) and filter_fasta.py) before building Operational Taxonomic Units (OTUs). OTUs were retained at >= 97% similarity (pick_open_reference_otus.py using uclust) against the Greengenes database (gg_13_8_99, from August 2013, which contains 202,421 bacterial and archaeal sequences). We annotated these OTUs and built read abundance tables per phylum and genus.

Quantitative PCR

The 16S rRNA gene V4 variable region PCR primers 515/806 with barcode on the forward primer were used in a 30 cycle PCR (5 cycle used on PCR products) using the HotStarTaq Plus Master Mix Kit (Qiagen, USA) under the following conditions: 94°C for 3 minutes, followed by 28 cycles of 94°C for 30 seconds, 53°C for 40 seconds and 72°C for 1 minute, after which a final elongation step at 72°C for 5 minutes was performed.

Taxonomical diversity of the Microbiota

We obtained 5,493,157 reads from the V4 region of the 16SDNA (Table SI-1). These reads were pooled, cleaned to remove chimeras, which retained 87.2 % of the reads (4,789,443 out of 5,493,157 16S rRNA sequences for 62 samples). These clean reads were clustered in OTUs (>= 97%) with QIIME (pick_open_reference.py, keeping only OTU with >= 3 reads).

Once OTUs were clustered and abundances tables built, we calculated alpha- and beta-diversity indices at the phylum and genus level. Alpha-diversity indices (Shannon, Simpson and Chao1) were computed with QIIME (34, 37), with and without multiple rarefaction (using alpha_diversity.py and alpha_rarefaction.py scripts). Alpha-diversity indices of different groups of samples were compared using a Mann-Whitney U-test. Beta-diversity analysis was computed with R (Bray-Curtis measure (38)), and the similarities between samples analyzed by Non-metric MultiDimensional Scaling (NMDS) with R (package vegan). We used an ANOSIM to test the significance of the dissimilarity between groups. The ANOSIM gives two values: the R statistic and the Pvalue. The R statistic informs about the dissimilarity between groups, and the Pvalue about if we can trust the R stastistic. The nearer to 1 R is, the more groups are distinct. If R is near zero no differences between groups exist and if R is near -1 samples within groups are more distinct than samples from different groups.

In order to explore structure in the different *Podarcis sicula* populations based on their gut microbial community we searched for the presence of enterotypes. We followed a published pipeline (http://enterotype.embl.de/enterotypes.html) and visualized the results using BCA (Between Class Analysis) (11).

In order to identify phyla and genera specific to each diet, year and season of sampling, geographical origin (island versus mainland), and the sex of the lizards, we analyzed the phyla and genera abundance matrices by performing Linear Discriminant Analysis with Size effect, using LefSe tool (http://huttenhower.sph.harvard.edu/galaxy) developed by Galaxy (39), default parameters: alpha values of the LefSe set to < 0.05 for the Kruskal-Wallis test among classes and for the Wilcoxon test between subclasses, and threshold on the logarithmic LDA score for discriminative features set to > 2.0).

After finding which phyla and genera from the abundances tables could explain variables (diet, sex, insularity, season, year of sequencing, and geography), we tried to build a model based on these variables to explain the abundances tables. We used RDA (Redundancy Analysis) on distance matrices (40). Inputs are abundances tables (phylum and genus level). We applied the R function decostand to normalize data tables with the “hellinger” option. Then, we applied RDA function from vegan package (https://CRAN.R-project.org/package=vegan).

With the 16S, it is possible to predict the presence of functional content from marker gene with the picrust tool. We used the function 'predict_metagenomes.py' with the default parameters. Next, we used STAMP (Statistical Analysiss of Metagenomic profiles) on the picrust results to analyze the metagenomic profiles and to visualize metabolic pathways

which have significant difference (statistical test: ANOVA) between groups of lizards (e.g. omnivorous and insectivorous lizards).

Results and Discussion

Fifty-one % of the OTUs were annotated at the level of the genus, and 99.1% were annotated at the level of the phylum. The low number of annotated genera can be explained in two ways: by the fact that the reference database (greengene13-8) was lacking representatives of these genera, and by the fact that some of these genera were specific to the lizard gut microbiota and thus novel.

Each lizard hosted on average 3,869 OTUs (Table SI-2), and there were a total of 32,850 OTUs across the 62 individuals. The average archaeal reads/bacterial reads ratio was 1.17%, with significant variations across different groups of lizards (see below). Methanogens feature prominently amongst these archaea (i.e. 52,325 reads out of 54,801 were assigned to *Methanomicrobia*, including 5 OTUs that could not be assigned to known genera).

One hundred and fifty eight OTUs (0.48% of all the OTUs) were shared by all lizards, constituting the core microbiota. This core microbiota belonged to 8 phyla, i.e. by decreasing number of OTUs:*Firmicutes* (110 OTUs), *Bacteroidetes* (24 OTUs), *Proteobacteria* (13 OTUs), *Actinobacteria* (5 OTUs), *Tenericutes* (2 OTUs), *Verrucomicrobia* (1 OTU), *Euryarchaeota* (1 OTU of *Methanobrevibacter*), *Fusobacteria* (1 OTU), and one unassigned OTU. A larger fraction of the microbiota (1414 OTUs) was present in a majority of lizards (Fig. 1).

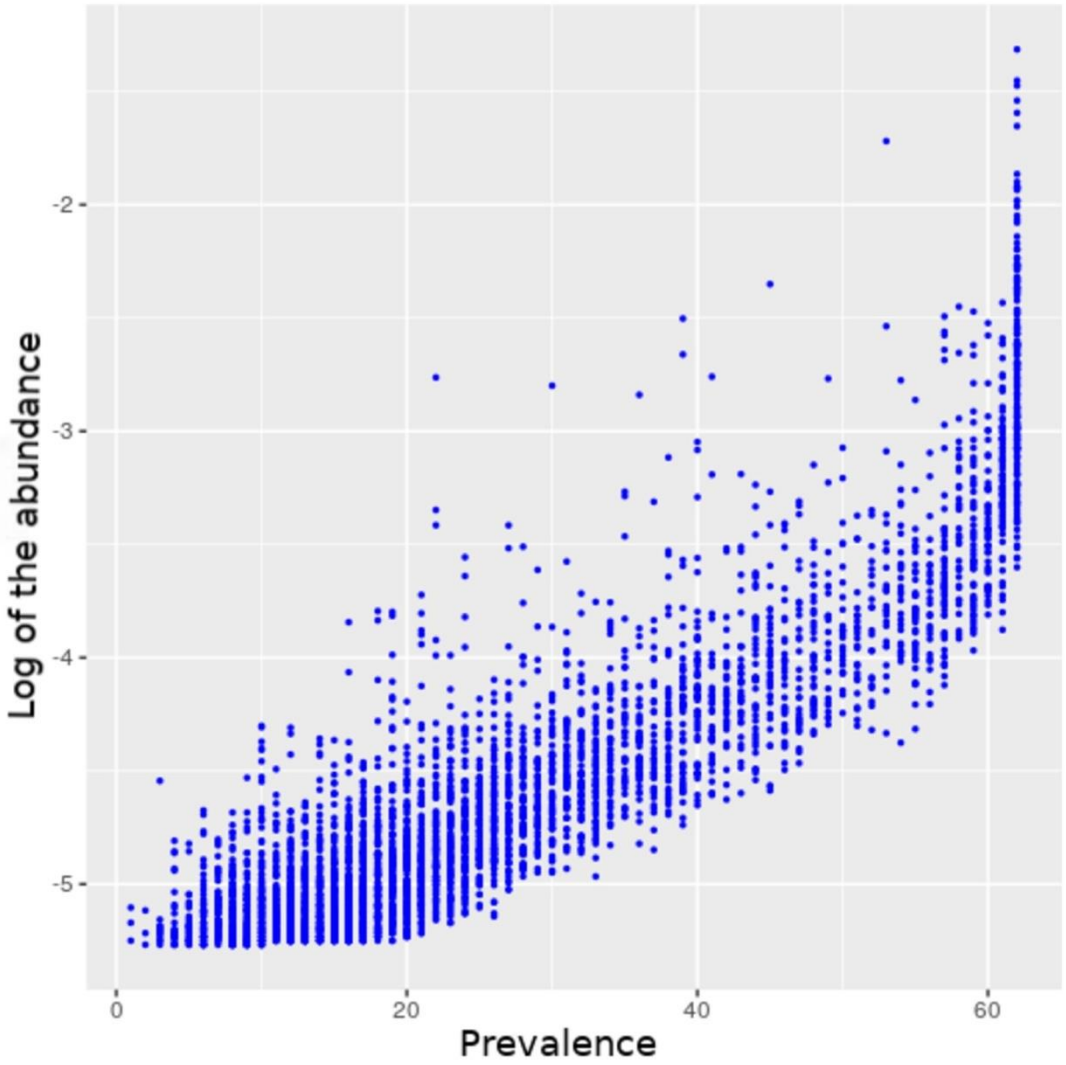


Figure 1: Logarithm of normalized abundance in function of prevalence.

The x-axis is the prevalence of the OTUs, the y-axis is the logarithm of normalized abundances of the OTUs (at 97%). Each blue point is an OTU. The prevalence corresponds to the number of lizards in which the OTU is found. The maximum value of the prevalence is 62 (the number of samples).

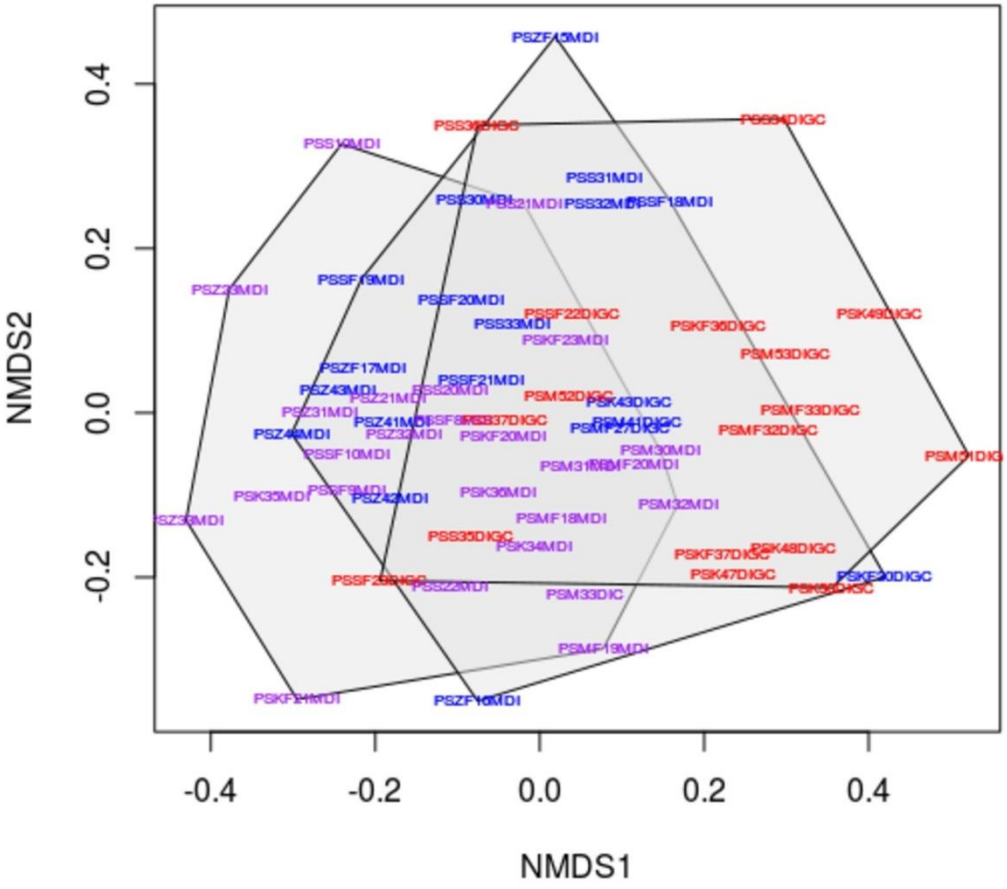
Thus, 4.3% of all the OTUs of the microbiota were prevalent (i.e. present in > 50% individual lizards) in our natural population (Abdul, 2015). Core and prevalent OTUs also contained 22 of the 25 abundant OTUs (i.e. amounting to > 10% of the reads of at least one lizard), i.e. nine *Firmicutes, six Proteobacteria,* two *Bacteroidetes,* two *Verrucomicrobia,* one *Fusobacteria,* one *Euryarchaeota,* and the unassigned OTU. The vast majority of OTUs (>99.9%) were thus rare in our dataset. These reduced numbers may reflect the limited sequencing coverage of the microbiota in each lizard (Figure S1-3), but may also indicate a genuine heterogeneity in the composition of their gut microbiota.

Next, we grouped individual lizards into (eventually overlapping) classes based on their diet (I, insectivorous and O, omnivorous), populations of origin (C, continental, with large populations and Is, insular, with small populations), sex (M, male and F, female), season of sampling (Sp, spring and Su, summer), and year of sampling (2014, 2015, 2016), and computed the Shannon, Evenness and Chao1 indices for each group (Table SI-4). We compared these values to identify significant differences between groups of lizards. There were no significant differences associated with the sex and season/year of sampling. Yet, island lizards had a higher average diversity than mainland lizards according to both the Shannon index (7.14 *vs.* 6.76, respectively, $P = 0.006$, Mann-Whitney U-test) and the Chao1 index (9157 *vs.* 7302, respectively, $P = 3.206e\text{-}05$, Mann-Whitney U-test). This higher diversity may be explained in part by the presence of omnivorous insular lizards, since the diversity of omnivorous lizards was in general higher than the diversity of insectivorous lizards (Shannon index of 7.38 for O *vs* 6.80 for I, $P = 0.001$; Simpson index of 0.94 for O *vs.* 0.94 for I, $P = 0.03$; Chao1 index of 9048.73 for O *vs.* 7889.81 for I, P = 0.01, Mann-Whitney U-test). Moreover, insectivorous insular lizards harbor more diverse communities than continental insectivorous lizards (Chao1 index of 9273.01 for insectivorous lizards on islands *vs.* 7302.99 for insectivorous lizards on the contnent, $P = 0.01$, Mann-Whitney U-test). These results are compatible with the observation that alpha-diversity measures are not significantly different between insular omnivorous and insular insectivorous lizards. Altogether, these measures suggest that diet and populations of origins affect the microbiota. More precisely, the ratio archaea/bacteria was significantly different (Mann-Whitney U-test, $P < 0.001$, higher in omnivorous lizards, in average a ratio of $6x10^{-4}$ for lizards of the insectivorous group *vs.* a ratio $4.6x10^{-2}$ for the lizards of the omnivorous group.) This higher abundance of archaea (also observed in the absolute number of archaeal *vs.* bacterial OTUs, with on average 3.4% of the OTUs being annotated as archaeal in the omnivorous lizards *vs.* 0.05% in insectivorous lizards) was due in part to the presence of methanogens (i.e., within *Methanomicrobia*, members of the *Methanobacteriaceae* family, including two unassigned genera, members of *Methanocorpusculaceae, Methanomassiliicoccaceae*, and the *Methanosarcinaceae* families)

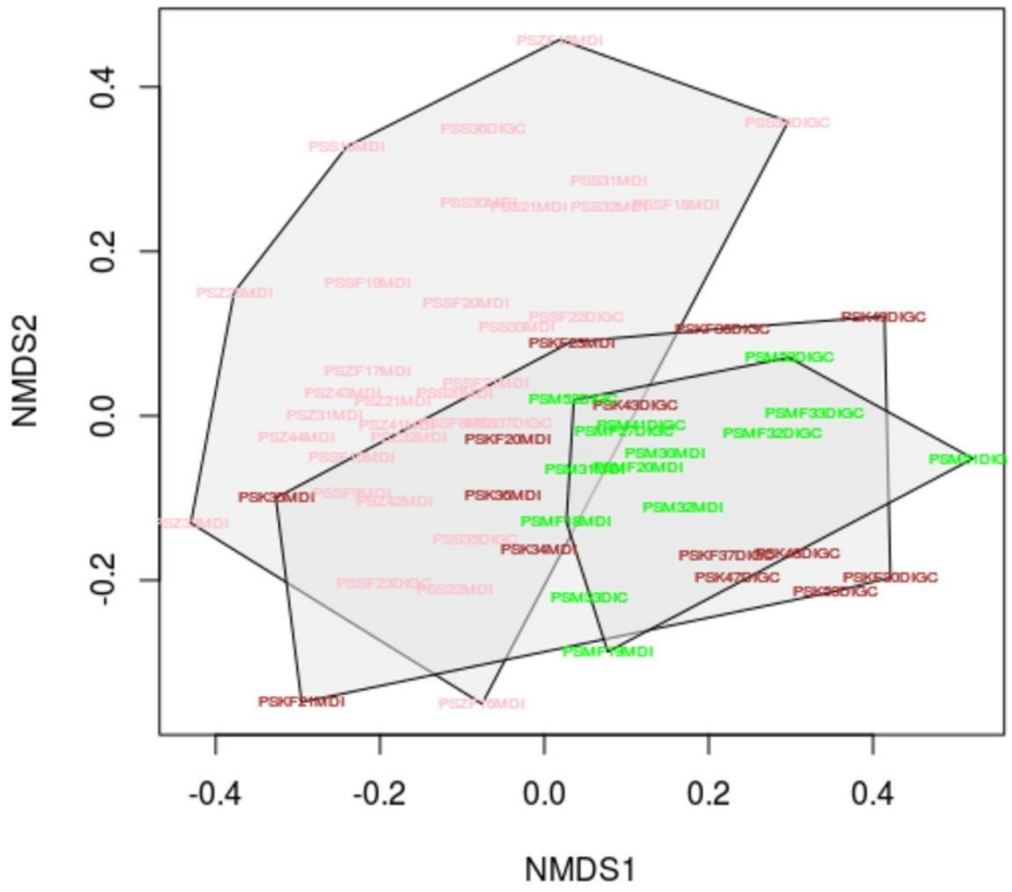
We performed analyses of beta diversity, directly comparing the distributions of OTUs across lizards and groups of lizards. NMDS analyses did not reveal significant differences between the communities of male and female lizards. Likewise, it did not reveal differences between the communities of insectivorous and omnivorous lizards (when an outlier omnivorous lizard, PSM50, which presented an unusual taxonomic distribution, e.g. 84% of its reads were associated with 193 OTUs from the phylum Rickettsiella, was removed from the analysis).

Yet, NMDS analyses showed that continental and island lizards had significantly different communities ($P = 0.001$, albeit with a low stress-value for the NMDS, $R = 0.32$). This difference was not due to the diet of island lizards (e.g. on Pod Kopište, lizards are insectivorous, which may have induced a greater similarity with the insectivorous continental lizards if the diet was impacting NMDS statistics). However, microbial communities of insular omnivorous lizards differed from that of insectivorous continental lizards, suggesting that both

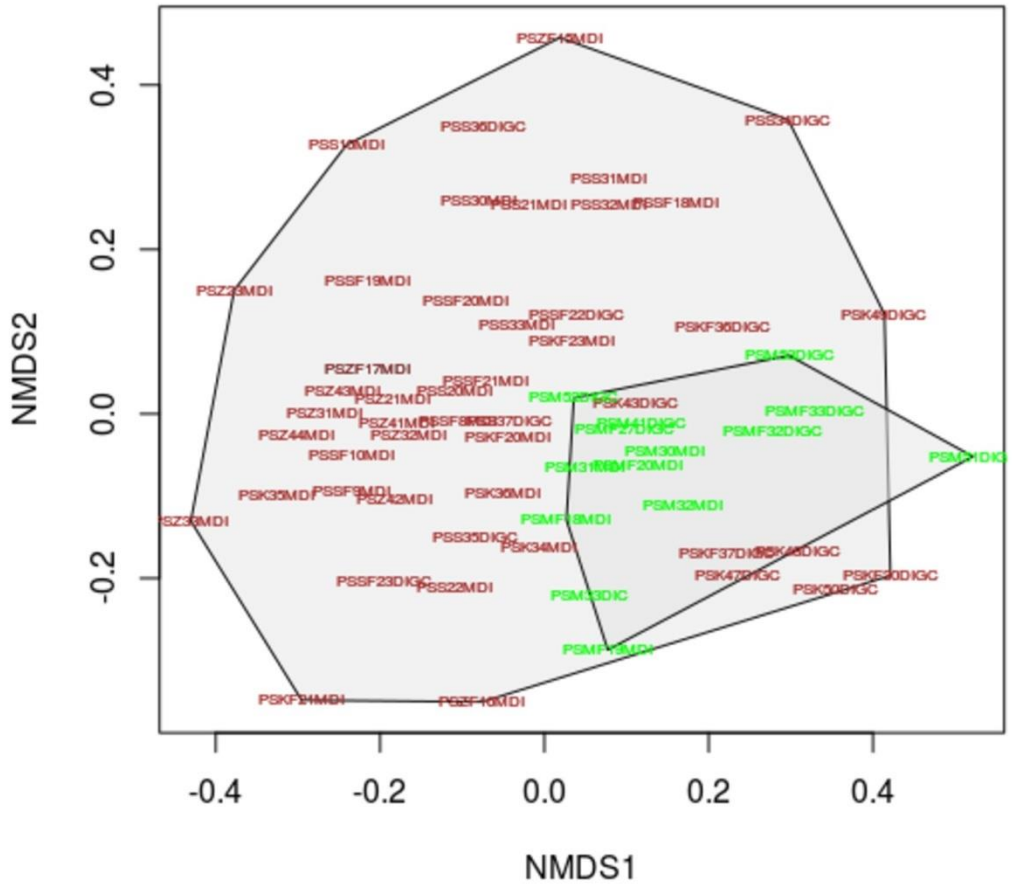
populations of origin and diet affected the microbiota. Gut communities also evolved by year of sampling (but not by seasons), with a significant difference between communities of 2014 and those of 2016 (P = 0.001, albeit with a low stress-value in the NMDS, R = 0.21), suggesting the microbiota composition slightly shifted over time. An alternative interpretation would be that microbial DNA became slightly altered during its conservation, since, although they were sequenced simultaneously, the different samples were inevitably collected at separate times.



a) NMDS on the year of sampling. Samples from 2016 are in purple. Samples from 2015 are in blue. Samples from 2014 are in red.



b) NMDS on the geography. Samples from the continent are in pink. Samples from Pod Kopište are in brown. Samples from Pod Mrčaru are in green.



c) NMDS on diet. In brown are insectivorous lizards and in green are omnivorous lizards.

Characteristics	*R*	*P*
Diet	0.1143	0.053
Year of sampling	0.21	0.001
Insularity	0.36	0.001
sex	-0.032	0.8
Location	0.32	0.001
Season	0.11	0.015

d) Significance analysis of the NMDS (ANOSIM on NMDS results);

Figure 2: Non metric multi-dimensional scaling for beta diversity analysis.

Samples are more similar if their labels are closed. a) NMDS on the year of sampling. Labels are the id of the sample. Samples from 2016 have a purple label, samples from 2015 have a blue label and samples from 2014 have a red label. b) NMDS on the geography. Samples from the continent are in pink, samples from Pod Kopište are in brown, and samples from Pod Mrčaru are in green. c) NMDS on diet. In brown are indicated insectivorous, and in green omnivorous lizards. d) ANOSIM results for each feature. The first column are the features tested. The second column gives the values for the R statistic of the ANOSIM for each feature. This helps to determine if groups are significantly different. The last column is the *P*-value, and gives the significance of the R-statistic value.

Because R are near from zero, there is no difference in beta diversity between groups for these variables.

We further investigated whether the differences between microbial communities were associated with the existence of distinct types of communities or taxa (i.e. enterotype (11–13, 15) or biomarkers (27)), that may themselves correlate with features of the lizards. These notions can be strongly contrasted. On the one hand, each enterotype is expected to correspond to a type of microbial community, i.e. a community with similar proportions of the same taxa. On the other hand, biomarkers correspond to particular species rather than entire community that correlate with a particular phenotypic condition. We performed a between-class analysis (BCA) of the abundance tables of i) genera and ii) phyla, that were assigned by QIIME from the OTUs (at >= 97%). Of note, at the genus level, since 59 % of the OTUs were taxonomically un-annotated, a large proportion of the OTUs were not included in the analysis. At the genus level, the BCA recovered three clusters. At the phylum level, it recovered five clusters. In each case, the two axes of the plots collectively explained only 23.4% of the variance for the BCA at the genus level, and 30.5% of the variance for the BCA at the phylum level. The variance explained by the model is low in both cases. Still, we performed a more detailed analysis of the taxonomic composition of each cluster. It returned a contrasting result with former studies proposing the existence of enterotypes (in other taxa, such as humans, apes, or bees (11–16) and with studies indicating that biomarkers (17, 41) drive the clustering in BCA. Namely, contrasting the taxonomical distribution of the members of the different clusters showed that there were neither dominant taxa (such as *Bacteroides* or *Prevoltella* (6)) nor globally conserved sets of taxa that were specifically associated with each cluster found in the lizard gut microbiota (as also confirmed below by RDA analyses). In that sense, while different groups of microbiota could be statistically distinguished, explaining what taxonomical signal structured these groups was not straightforward. Lizards from all populations, with different diets, sexes, and sampled at different times were typically joined in clusters, so no correlations with hosts features could easily explain the apparent structuring of the microbial communities in our these enterotypes.

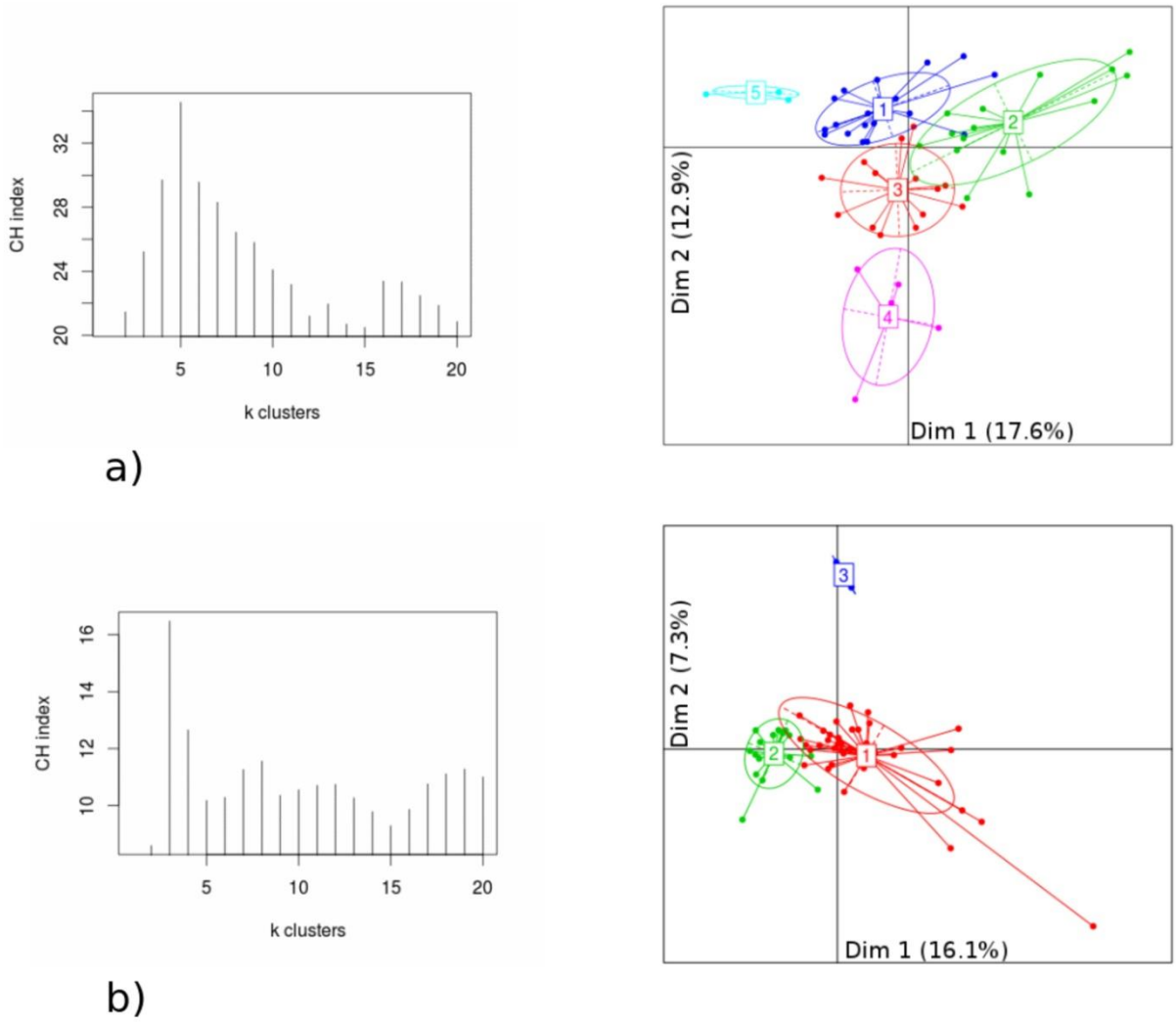


Figure 3: Enterotypes at the phylum and genus level in *Podarcis sicula* gut microbiota

a) Enterotypes at the phylum level. At the left is the CH index graph which allows to choose the optimal number of clusters (for the higher CH index). At the phylum level this is five. At the right is the BCA at the phylum level with five groups. Points are samples. The circles are the clusters and the color of the circles and points are the cluster color. There are no dominant taxa per group. The two axis explain 30.5% of the variance. b) Enterotypes at the genus level. At the left is the CH index graph which allows to choose the optimal number of clusters (for the higher CH index). At the genus level this is three. At the right is the BCA at the phylum level with three groups. Points are samples. The circle are the clusters and the color of the circles and points is the cluster color. There are no dominant taxa per group.

Since no classic enterotypes were detected in our samples, we applied two independent supervised methods (redundancy analyses, RDA and linear discriminant analyses, LDA) to test whether some hosts features correlated with the composition and abundance of microbial taxa. More precisely, RDA tests what host features (or combinations of host features) can explain the abundance of OTUs, whereas the LDA tests whether the abundance of specific microbial phyla allows predicting certain hosts' features.

To assess the influence of explanatory variables (diet, population of origin, sex, season, year) on the microbiota composition, we analyzed our abundance matrices using the RDA function of the R package *vegan,* with RsquaredAdj function. Seven distinct models were constructed to test the relative importance of (1) diet (O *vs.* I), (2) insularity (Is *vs.* C), (3) geography (i.e. the 4 sampling sites), (4) sex (M *vs.* F), (5) season (Sp *vs.* Su), (6) year (the year of sampling), and (7) the complete model (i.e. all explanatory variables were considered together). The sex model explained none of the variation and the season model explained at most 0.98% of the variation in the phyla for these microbiota. The diet model explained at most 4.53% of the variation, the year model 6.74% of the variation, the insularity model 7.86%, and the geography model 9.57% of the variation in the phyla. The complete model explained at most 17.1% of the variation in the phyla. Taken together, these results suggest that the qualitative hosts features studied here only weakly explain the variation in phyla of the lizard gut microbiota. RDA of the abundance tables of the genera and of the OTU abundance table at the genera level supported an identical conclusion: the qualitative hosts features investigated in this study weakly explain variation in genera (18%) and in OTUs (15%) of the lizards gut microbiota. We also tested an eighth 'enterotype' model, which tested whether the 5 clusters of gut microbiota detected at the phylum level by the BCA could explain the differences between our samples. Enterotypes only explained 5,60% of the variation of the phyla in the microbiota, confirming that the members of our BCA clusters cannot be considered as typical communities united by similar and markedly distinct composition of phyla. This result is consistent with the low percentage of variance explained by the axes of the BCA as reported above. Together, these observations indicate that there were no major shifts in the microbial communities associated with the host features investigated here, but most likely changes in targeted specific phyla, likely of limited abundance.

LDA analyses confirmed this conclusion. LDA were separately performed on annotated genera and phyla. They recovered signals indicating that diet, population of origin, and year of sampling, but also season of sampling were associated with distinct variations between microbial communities. By contrast, male and female lizards have comparable compositions of the gut microbiota.

The composition of the gut microbiota slightly changed over time as seven phyla (*Elusimicrobia, Thaumarchaeota, Chloroflexi, Cyanobacteria/Chloroplast, Chlamydiae, Actinobacteria, Fusobacteria*) presented different abundances according to the seasons of sampling. Accordingly, six out of these seven phyla were also recovered as variable phyla amongst the eight phyla (*Elusimicrobia, Spirochaetes, Chloroflexi, Actinobacteria, Cyanobacteria/Chloroplast, Bacteroidetes, Fusobacteria, Chlamydiae*) that changed in abundance during the years of sampling.

Interestingly, four phyla (*Euryarchaeota, Spirochetes, Elusimicrobia* and *Planctomycetes* in order of significance) discriminate between omnivorous and insectivorous lizards. These phyla, which are not amongst the most abundant ones in the microbiota, are more abundant in omnivorous lizards. Three of these phyla have been identified in prior studies of gut microbiomes as being associated with a diet rich in plants (5, 42). *Elusimicrobia*, for example, were reported to be involved in wood digestion in termites (43, 44), *Euryarchaeota* were

reported to be involved in fiber and cellulose digestion in Yaks (45, 46), *Spirochaetes* were reported to be involved in plant digestion in termites (47–50). Finally, *Planctomycetes* are consistently described as low abundance taxa in microbiomes (7, 48). Therefore, differences in the composition of the microbial community associated with the shift in diet concern several, quantitatively minor, phyla.

Eleven phyla (*Euryarchaeota, Spirochetes, Lentisphaerae, Synergistes, Deferribacteres, Planctomycetes* are more abundant in island lizards, and *Cyanobacteria/Chloroplast, Chlamydiae, Actinobacteria, Fusobacteria, Bacteroidetes* were more abundant in continental lizards) discriminate between continental and insular lizards. These differences may be due to the population of origins but also to diet, because island lizards display different diets. We disentangled the effect of insularity from the effect of diet by first comparing lizards with a similar diet, i.e. insectivorous lizards. Eleven phyla (*Woesearchaeota, Spirochetes, Lentisphaerae, Synergistes, Deferribacteres,* more abundant in insectivorous insular lizards; and *Elusimicrobia, Cyanobacteria/Chloroplast, Chlamydiae, Actinobacteria, Fusobacteria, Bacteroidetes* more abundant in continental lizards) discriminate between insectivorous lizards from the continental and those from islands. The discriminating phyla are almost identical to the ones mentionned in the continent versus island comparison, to the exception of the *Woesearchaeota*, *Planctomycetes,* and *Elusimicrobia*, which reinforces the notion that some phyla correlate with the origin of the population rather than diet. We next disentangled the effect of diet from the effect of the population of origins by comparing lizards with a different diet, but similar living areas, i.e. insular lizards. Five phyla (*Euryarchaeta, Spirochaetes, Planctomycetes*, more abundant in omnivorous insular lizards; and *Synergistetes* and *Deferribacteres*, more abundant in insectivorous insular lizards). Altogether, these results suggest that few, specific phyla show significant variation in abundance, and that *Euryarchaeota*, *Spirochaetes* and *Planctomycetes* are possible biomarkers, not of enterotypes identified by BCA, but for an omnivorous diet in lizards.

We also used 16S data to infer what metabolisms may possibly have been associated with these microbiota. Such an inference is potentially problematic because lateral gene transfer can decouple the gene content of genomes, i.e. genomes with similar 16S can have very different metabolic capabilities (51). Therefore, the following results remain speculative and must be interpreted carefully. The picrust procedure identified pathways putatively enriched in lizards with different diets (104 pathways), from different populations of origins (79 pathways), from different seasons (23 pathways). Since picrust infers the presence of metabolic pathways based on 16S information, the numbers of pathways whose abundance may change between groups of lizards reflect the variations detected in the 16S analysis, confirming that diet and population of origins potentially have the highest impact on the gut microbiota, and by extension on the gut microbiome (gene content and metabolism). However, it is difficult to further interpret these lists of pathways. Some of these inferences are clearly artifacts (i.e. Alzeihmer diseases, Huttington's disease, etc.) but others variations are suggestive. For example, the enrichment of peptidoglycan biosynthesis and fatty-acid biosynthesis pathways in omnivorous lizards and an enrichment of other glycan degradation pathways in insectivorous lizards may reflect the properties of the ingested food. These

inferences of enrichment will, however, require further confirmation from shotgun metagenomic studies.

Conclusion

The morphology of *Podarcis sicula* has changed in a few generations in line with a dietary shift (higher bite force, larger body size, evolution of caecal valves; Herrel et al., 2008). By contrast, the dietary shift that occurred in the introduced population of *Podarcis sicula* had a limited impact on the microbiota. Rare rather than abundant microbial taxa changed between the populations of lizards, enhancing the taxonomic diversity in the gut communities of omnivorous lizards with respect to their insectivorous relatives. Interestingly, these rare taxa corresponded to phyla already described for their potential in plant and fiber digestion, and we propose that they may qualify as biomarkers.

Even if studies on humans and mice found that diet is one of the major drivers of the gut microbiota (52–55), our conclusion is consistent with published results of other studies suggesting smaller, targeted changes. Indeed, between wood-fed and grass-fed termites (*Nasutitermes* and *Gnathamitermes*), two phyla vary, but in minor proportions (four to eight %). Few variations in microbiota associated with diet have also been detected in humans (56) and in the greater panda (57–59). This is not surprising, because if the first function of the microbiota is the digestion of food substrates, this function is not the only one realized by the microbiota, since it is also involved in host health (60, 61), immune system development (62–65), and brain and gut development (66). Fulfilling these other functions explain the presence of a diversity of microbial taxa, irrespective of the changes in diet.

Interestingly, former work had proposed that herbivorous hosts harboured a higher taxonomic diversity than omnivorous hosts, which themselves presented richer communities than carnivorous hosts. Our results are compatible with these observations, and further suggest that, even though the impacted taxa are not the most abundant ones, targeted changes in the microbiota can be adaptive when they affect species that encode genes able to help digesting plants and plant fibers. Ultimately our work also shows that the gut microbiota of non-model poikilotherm vertebrates, such as lizards, remains largely underexplored. In particular, the diversity of methanogens deserves greater considerations, and could benefit from further study of such taxa.

Acknowledgments

LABEXBCDIV, European Research Council (FP7/2017-2013 Grant Agreement #615274), funding enviromics, national geographic, Virginie Lemieuxlabonté, Maité Ribère

References

1. Nevo E, Gorman G, Soulé M, Yang SY, Clover R, Jovanović V. 1972. Competitive exclusion between insular Lacerta species (Sauria, Lacertidae). Oecologia 10:183–190.

2. Herrel A, Huyghe K, Vanhooydonck B, Backeljau T, Breugelmans K, Grbac I, Van Damme R, Irschick DJ. 2008. Rapid large-scale evolutionary divergence in morphology and performance associated with exploitation of a different dietary resource. Proc Natl Acad Sci 105:4792–4795.

3. Umu ÖCO, Frank JA, Fangel JU, Oostindjer M, da Silva CS, Bolhuis EJ, Bosch G, Willats WGT, Pope PB, Diep DB. 2015. Resistant starch diet induces change in the swine microbiome and a predominance of beneficial bacterial populations. Microbiome 3:16.

4. Frese SA, Parker K, Calvert CC, Mills DA. 2015. Diet shapes the gut microbiome of pigs during nursing and weaning. Microbiome 3:28.

5. Ravel J, Blaser MJ, Braun J, Brown E, Bushman FD, Chang EB, Davies J, Dewey KG, Dinan T, Dominguez-Bello M, Erdman SE, Finlay BB, Garrett WS, Huffnagle GB, Huttenhower C, Jansson J, Jeffery IB, Jobin C, Khoruts A, Kong HH, Lampe JW, Ley RE, Littman DR, Mazmanian SK, Mills DA, Neish AS, Petrof E, Relman DA, Rhodes R, Turnbaugh PJ, Young VB, Knight R, White O. 2014. Human microbiome science: vision for the future, Bethesda, MD, July 24 to 26, 2013. Microbiome 2:16.

6. Gorvitovskaia A, Holmes SP, Huse SM. 2016. Interpreting Prevotella and Bacteroides as biomarkers of diet and lifestyle. Microbiome 4:15.

7. Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, Gordon JI. 2008. Evolution of Mammals and Their Gut Microbes. Science (80-) 320:1647–1651.

8. Amato KR, Yeoman CJ, Cerda G, A. Schmitt C, Cramer JD, Miller MEB, Gomez A, R. Turner T, Wilson BA, Stumpf RM, Nelson KE, White BA, Knight R, Leigh SR. 2015. Variable responses of human and non-human primate gut microbiomes to a Western diet. Microbiome 3:53.

9. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev A V, Lonsdorf E V, Muller MN, Pusey AE, Peeters M, Hahn BH, Ochman H. 2016. Cospeciation of gut microbiota with hominids. Science 353:380–382.

10. Abdul Rahman N, Parks DH, Willner DL, Engelbrektson AL, Goffredi SK, Warnecke F, Scheffrahn RH, Hugenholtz P. 2015. A molecular survey of Australian and North American termite genera indicates that vertical inheritance is the primary force shaping termite gut microbiomes. Microbiome 3:5.

11. Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, 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 HB, Nielsen T, Pons N, Poulain J, Qin J, Sicheritz-Ponten T, Tims S, Torrents D, Ugarte E, Zoetendal EG, Wang J, Guarner F, Pedersen O, de Vos WM, Brunak S, Dore J, Weissenbach J, Ehrlich SD, Bork P. 2011. Enterotypes of the human gut microbiome. Nature 473:174–180.

12. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen Y-Y, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. Science (80-) 334:105 LP-108.

13. Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. 2012. Chimpanzees and Humans Harbor Compositionally Similar Gut Enterotypes. Nat Commun 3:1179.

14. Moeller AH, Peeters M, Ayouba A, Ngole EM, Esteban A, Hahn BH, Ochman H. 2015. Stability of the gorilla microbiome despite simian immunodeficiency virus infection. Mol Ecol 24:690–697.

15. Li J, Powell JE, Guo J, Evans JD, Wu J, Williams P, Lin Q, Moran NA, Zhang Z. 2015. Two gut community enterotypes recur in diverse bumblebee species. Curr Biol 25:R652–R653.

16. Lim MY, Rho M, Song Y-M, Lee K, Sung J, Ko G. 2014. Stability of gut enterotypes in Korean monozygotic twins and their association with biomarkers and diet. Sci Rep 4:7348.

17. Jeffery IB, Claesson MJ, O'Toole PW, Shanahan F. 2012. Categorization of the gut microbiota: enterotypes or gradients? Nat Rev Microbiol 10:591–592.

18. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, Knight R. 2014. Rethinking "Enterotypes." Cell Host Microbe 16:433–437.

19. Whittaker RH. 1972. Evolution and Measurement of Species Diversity. Taxon 21:213–251.

20. Whittaker RH. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol Monogr 30:279–338.

21. Yang X, Cheng G, Li C, Yang J, Li J, Chen D, Zou W, Jin S, Zhang H, Li D, He Y, Wang C, Wang M, Wang H. 2017. The normal vaginal and uterine bacterial microbiome in giant pandas (Ailuropoda melanoleuca). Microbiol Res 199:1–9.

22. Bennett DC, Tun HM, Kim JE, Leung FC, Cheng KM. 2013. Characterization of cecal microbiota of the emu (Dromaius novaehollandiae). Vet Microbiol 166:304–310.

23. Blasco G, Moreno-Navarrete JM, Rivero M, Pérez-Brocal V, Garre-Olmo J, Puig J, Daunis-i-Estadella P, Biarnés C, Gich J, Fernández-Aranda F, Alberich-Bayarri Á, Moya A, Pedraza S, Ricart W, López M, Portero-Otin M, Fernandez-Real J-M. 2017. The Gut Metagenome Changes in Parallel to Waist Circumference, Brain Iron Deposition, and Cognitive Function. J Clin Endocrinol Metab 102:2962–2973.

24. Lee SC, Tang MS, Lim YAL, Choy SH, Kurtz ZD, Cox LM, Gundra UM, Cho I, Bonneau R, Blaser MJ, Chua KH, Loke P. 2014. Helminth Colonization Is Associated with Increased Diversity of the Gut Microbiota. PLoS Negl Trop Dis 8:e2880.

25. Gomez DE, Arroyo LG, Costa MC, Viel L, Weese JS. 2017. Characterization of the Fecal Bacterial Microbiota of Healthy and Diarrheic Dairy Calves. J Vet Intern Med 31:928–939.

26. Ling Z, Liu X, Cheng Y, Jiang X, Jiang H, Wang Y, Li L. 2015. Decreased Diversity of the Oral Microbiota of Patients with Hepatitis B Virus-Induced Chronic Liver Disease: A Pilot Project. Sci Rep 5:17098.

27. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol 12:R60–R60.

28. Kohl KD, Brun A, Magallanes M, Brinkerhoff J, Laspiur A, Acosta JC, Caviedes-Vidal E, Bordenstein SR. 2017. Gut microbial ecology of lizards: insights into diversity in the wild, effects of captivity, variation across gut regions and transmission. Mol Ecol 26:1175–1189.

29. Kohl KD, Brun A, Magallanes M, Brinkerhoff J, Laspiur A, Acosta JC, Bordenstein SR, Caviedes-Vidal E. 2016. Physiological and microbial adjustments to diet quality permit facultative herbivory in an omnivorous lizard. J Exp Biol 219:1903–1912.

30. Ren T, Kahrl AF, Wu M, Cox RM. 2016. Does adaptive radiation of a host lineage promote ecological diversity of its bacterial communities? A test using gut microbiota of Anolis lizards. Mol Ecol 25:4793–4804.

31. Kohl KD, Amaya J, Passement CA, Dearing MD, McCue MD. 2014. Unique and shared responses of the gut microbiota to prolonged fasting: a comparative study across five classes of vertebrate hosts. FEMS Microbiol Ecol 90:883–894.

32. Hanning I, Diaz-Sanchez S. 2015. The functionality of the gastrointestinal microbiome in non-human animals. Microbiome 3:51.

33. HERREL A, JOACHIM R, VANHOOYDONCK B, IRSCHICK DJ. 2006. Ecological consequences of ontogenetic changes in head shape and bite performance in the Jamaican lizard Anolis lineatopus. Biol J Linn Soc 89:443–454.

34. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–6.

35. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. 2012. Using QIIME to analyze 16s rRNA gene sequences from microbial communities. Curr Protoc Microbiol.

36. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

37. Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. 2012. Using QIIME to analyze 16S rRNA gene sequences from Microbial Communities. Curr Protoc Bioinforma 36:10.7:10.7.1–10.7.20.

38. Borcard D, Gillet F, Legendre P. 2011. Numerical Ecology With RNumerical Ecology with R.

39. Afgan E, Baker D, van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, Grüning B, Guerler A, Hillman-Jackson J, Von

Kuster G, Rasche E, Soranzo N, Turaga N, Taylor J, Nekrutenko A, Goecks J. 2016. The {Galaxy} platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 44:W3–W10.

40. Legendre P, Anderson MJ. 1999. DISTANCE-BASED REDUNDANCY ANALYSIS: TESTING MULTISPECIES RESPONSES IN MULTIFACTORIAL ECOLOGICAL EXPERIMENTS. Ecol Monogr 69:1–24.

41. Knights D, Ward TL, McKinlay CE, Miller H, Gonzalez A, McDonald D, Knight R. 2017. Rethinking “Enterotypes” Cell Host Microbe 16:433–437.

42. Bauer E, Laczny CC, Magnusdottir S, Wilmes P, Thiele I. 2015. Phenotypic differentiation of gastrointestinal microbes is reflected in their encoded metabolic repertoires. Microbiome 3:55.

43. Su L, Yang L, Huang S, Su X, Li Y, Wang F, Wang E, Kang N, Xu J, Song A. 2016. Comparative Gut Microbiomes of Four Species Representing the Higher and the Lower Termites. J Insect Sci 16:97.

44. Herlemann DPR, Geissinger O, Brune A. 2007. The Termite Group I Phylum Is Highly Diverse and Widespread in the Environment . Appl Environ Microbiol 73:6682–6685.

45. Wei YQ, Long RJ, Yang H, Yang HJ, Shen XH, Shi RF, Wang ZY, Du JG, Qi XJ, Ye QH. 2016. Fiber degradation potential of natural co-cultures of Neocallimastix frontalis and Methanobrevibacter ruminantium isolated from yaks (Bos grunniens) grazing on the Qinghai Tibetan Plateau. Anaerobe 39:158–164.

46. Mountfort DO, Asher RA, Bauchop T. 1982. Fermentation of cellulose to methane and carbon dioxide by a rumen anaerobic fungus in a triculture with Methanobrevibacter sp. strain RA1 and Methanosarcina barkeri. Appl Environ Microbiol.

47. Piknová M, Javorský P, Guczynska W, Kasperowicz A, Michalowski T, Pristaš P. 2006. New species of rumen treponemes, p. 303–305. *In* Folia Microbiologica.

48. Santana RH, Catão ECP, Lopes FAC, Constantino R, Barreto CC, Krüger RH. 2015. The Gut Microbiota of Workers of the Litter-Feeding Termite Syntermes wheeleri (Termitidae: Syntermitinae): Archaeal, Bacterial, and Fungal Communities. Microb Ecol 70:545–556.

49. Köhler T, Dietrich C, Scheffrahn RH, Brune A. 2012. High-Resolution Analysis of Gut Environment and Bacterial Microbiota Reveals Functional Compartmentation of the Gut in Wood-Feeding Higher Termites (Nasutitermes spp.). Appl Environ Microbiol 78:4691–4701.

50. Schauer C, Thompson C, Brune A. 2014. Pyrotag Sequencing of the Gut Microbiota of the Cockroach Shelfordella lateralis Reveals a Highly Dynamic Core but Only Limited Effects of Diet on Community Structure. PLoS One 9:e85861.

51. Doolittle WF, Zhaxybayeva O. 2010. Metagenomics and the Units of Biological Organization. Bioscience 60:102–112.

52. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. 2016. Diet-induced extinctions in the gut microbiota compound over generations. Nature 529:212–215.

53. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling A V, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563.

54. Hildebrandt MA, Hoffman C, Sherrill-Mix SA, Keilbaugh SA, Hamady M, Chen Y-Y, Knight R, Ahima RS, Bushman F, Wu GD. 2009. High Fat Diet Determines the Composition of the Murine Gut Microbiome Independently of Obesity. Gastroenterology 137:1712–1716.

55. Zhang C, Zhang M, Pang X, Zhao Y, Wang L, Zhao L. 2012. Structural resilience of the gut microbiota in adult mice under high-fat dietary perturbations. ISME J 6:1848–1857.

56. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489:220–230.

57. Li Y, Guo W, Han S, Kong F, Wang C, Li D, Zhang H, Yang M, Xu H, Zeng B, Zhao J. 2015. The evolution of the gut microbiota in the giant and the red pandas. Sci Rep 5:10185.

58. Wei F, Hu Y, Yan L, Nie Y, Wu Q, Zhang Z. 2015. Giant Pandas Are Not an Evolutionary cul-de-sac: Evidence from Multidisciplinary Research. Mol Biol Evol 32:4–12.

59. Xue Z, Zhang W, Wang L, Hou R, Zhang M, Fei L, Zhang X, Huang H, Bridgewater LC, Jiang Y, Jiang C, Zhao L, Pang X, Zhang Z. 2015. The Bamboo-Eating Giant Panda Harbors a Carnivore-Like Gut Microbiota, with Excessive Seasonal Variations. MBio 6:e00022-15.

60. Huttenhower C, Human Microbiome Project Consortium. 2012. Structure, function and diversity of the healthy human microbiome. Nature 486:207–14.

61. Fujimura KE, Slusher NA, Cabana MD, Lynch S V. 2010. Role of the gut microbiota in defining human health. Expert Rev Anti Infect Ther 8:435–454.

62. Belkaid Y, Hand T. 2014. Role of the Microbiota in Immunity and inflammation. Cell 157:121–141.

63. Selosse MA, Bessis A, Pozo MJ. 2014. Microbial priming of plant and animal immunity: Symbionts as developmental signals. Trends Microbiol.

64. Wu H-J, Wu E. 2012. The role of gut microbiota in immune homeostasis and autoimmunity. Gut Microbes 3:4–14.

65. Eberl G. 2010. A new vision of immunity : homeostasis of the superorganism. Mucosal Immunol 3:450–460.

66. Sampson TR, Mazmanian SK. 2015. Control of Brain Development, Function, and Behavior by the Microbiome. Cell Host Microbe 17:565–576.

Dans ce chapitre, nous avons démontré que le microbiote semblait ne contenir qu'un petit microbiote ubiquitaire (0,48% des OTUs sont dans le microbiote ubiquitaire, soit 158 OTUs) et que peu d'OTUs sont abondantes (seules 25 OTUs sont abondantes). Cela reste néanmoins à nuancer dans la mesure où nos microbiotes sont sous-échantillonnés. Nous avons également montré que les microbiomes de lézards omnivores présentent une plus grande diversité que ceux des lézards insectivores. En considérant que l'insectivorie est une forme spécifique de carnivorie, notre résultat rejoint les conclusions de R.E Ley (Ley et al. 2008) qui proposait que les microbiomes intestinaux de mammifères herbivores présentent une plus grande diversité que les microbiomes intestinaux de mammifères omnivores, qui eux même sont plus diversifiés que les microbiomes intestinaux de mammifères carnivores.

Nous avons également montré l'absence d'entérotypes au niveau du phylum et du genre chez *Podarcis sicula*. En revanche, on a pu trouver grâce à l'outil LefSe que le changement de régime alimentaire chez *Podarcis sicula* est associé à des changements ciblés du microbiote, ne concernant que quelques phyla (Euryarchaeota, Spirochaetes Elusimicrobia, Planctomycetes) étant reconnus comme ayant un rôle dans l'herbivorie.

Enfin, on peut noter qu'une partie non négligeable du microbiote n'a pas pu être annoté au niveau du genre (49% de reads non annotés). Cela rejoint les résultats du Pr Raes (enterotypes of the gut microbiome), dont l'étude des entérotypes présentait 47% de reads non annotés au niveau du genre. Il semblerait qu'il reste de nombreux microorganismes à découvrir au sein des microbiotes intestinaux.

4. Le changement de régime alimentaire chez *Podarcis sicula* est associé à des changements ciblés dans le microbiome

4.1 Présentation de l'ensemble du jeu de données microbiome

Nous disposons actuellement de 50 microbiomes intestinaux de lézards, séquencés en 2014 (16 lézards), en 2015 (16) et en 2016 (18 idem). Le séquençage des microbiomes est non ciblé, et la taille des paires de « reads » obtenus est de 2x300 paires de bases pour les échantillons séquencés en 2014 et 2015. La taille des paires de « reads » obtenus en 2016 est de 2x150 paires de bases.

Cela représente 300 109 100 paires de reads, avec en moyenne environ six millions de paires de reads par échantillon.

Cette thèse proposera des méthodes théoriques afin d'analyser ce grand jeu de données. Les analyses statistiques ont été réalisées sur un sous-ensemble du jeu de données en se restreignant aux échantillons des 12 individus insulaires de l'année 2014. En effet, compte-tenu de la quantité de données (plus de 300 millions de paires de reads pour l'ensemble des 50 microbiomes), nous avons souhaité commencer par développer les analyses sur un jeu de données restreint, puis ensuite étendre les analyses sur l'ensemble du jeu de données, une fois les difficultés identifiées.

4.2 Présentation du jeu de donnée utilisé dans cette étude

Le jeu de données utilisé dans le cadre de ces analyses est l'ensemble des 12 microbiomes insulaires séquencés en 2014 (Figure 13). Nous nous sommes concentrés sur ces 12 échantillons parce que nous avons acquis les autres données de façon échelonnée au cours des trois ans de thèse. Nous avons donc commencé les analyses avec les échantillons dont nous disposions. D'autre part, comme je l'ai exposé en introduction, ces analyses sont limitées en raison du temps de calcul et de l'espace de stockage que nécessiterait l'étude du jeu de données complet (50 microbiomes). Nous avons donc souhaité avoir dans un premier temps une série d'analyses complète qui permette de répondre aux questions que soulève cette thèse sur douze individus, pour ensuite pouvoir l'étendre plus facilement à l'ensemble du jeu de données.

L'échantillonnage s'est déroulé au printemps.

Régime alimentaire	Localisation	Genre	Effectifs	Effectifs
Insectivores	Pod Kopište	Mâle	3	6
		Femelle	3	
Omnivores	Pod Mrčaru	Mâle	3	6
		Femelle	3	

Figure 13 : Répartition des effectifs du jeu de données utilisé dans ce chapitre.

Ce jeu de données représente 62 746 835 paires de « reads » (Figure 14) séquencés en illumina 2*300 paires de base. Tout d'abord, il est important de noter que les microbiomes lézards omnivores contiennent plus de « reads » que les lézards insectivores.

Régime alimentaire	Nombre total de reads	Moyenne du nombre de reads par échantillon	Ecart-type du nombre de reads
Insectivore	29 225 591	4 870 932	552 904
Omnivore	33 521 244	5 586 874	905 574

Figure 14 : Nombre de reads dans les microbiomes de lézards insectivores et omnivores.

La différence de 4 295 653 paires de « reads » entre les omnivores et les insectivores, équivaut à un microbiome supplémentaire dans le pool de microbiomes omnivores (alors qu'en réalité, il y a le même nombre de microbiomes dans chacun des deux groupes). Cette différence est donc non négligeable, elle représente pratiquement 7 % du jeu de données.

La qualité du jeu de données a été évaluée à l'aide de l'outil FastQC (FastQC n.d.), qui est communément utilisé pour analyser la qualité d'un jeu de données (des informations sur les séquences sont fournies, telles que qualité du « read » sur toute sa longueur base azotée par base azotée, le score de qualité phred Q, qui a la propriété d'être reliée de façon logarithmique à la probabilité d'erreur d'identification d'une base P : $Q = -10log_{10}P$, le contenu en bases GC par séquence, le nombre de

bases inconnues, la longueur des séquences, le niveau de duplication et de sur-représentation des séquences) ainsi qu'à l'aide de scripts python que j'ai élaborés. Puis les reads ont été filtrés en fonction de leur qualité à l'aide de l'outil « FASTX-TOOLKIT » (Lab n.d.) (script fastq_quality_filter les reads conservés ont une qualité de score phred d'au moins Q = 20 soit une probabilité d'erreur toutes les 100 paires de base).

4.3 Impact du régime alimentaire du *Podarcis sicula* sur les catégories COGs

Afin d'étudier l'impact du régime alimentaire sur le microbiome intestinal des *Podarcis sicula* au niveau des reads, nous avons choisi d'utiliser le serveur en ligne MGRAST (Wilke et al. 2016). Ce serveur se trouve à l'adresse suivante : http://metagenomics.anl.gov/. Il permet d'analyser taxonomiquement et fonctionnellement les métagénomes. MG-RAST propose les services suivants : contrôle de qualité, annotation taxonomique et fonctionnelle, comparaison des métagénomes entre eux, archivage des métagénomes et des résultats d'analyses réalisées avec MG-RAST.

Dans un premier temps, nous avons regardé l'abondance en reads de chaque grande classe de fonctions (en annotant fonctionnellement - comparaison avec la base de données des COGs (Clusters of Orthologous Groups), avec un pourcentage d'identité ≥ 90% et une couverture d'au moins 100 paires de bases). La base de données des COGs contient à la fois des groupes procaryotes (COGS) et des groupes eucaryotes (KOGs). Elle contient 138 458 protéines provenant de 66 génomes, ce qui représente 4,873 COGs et 4 852 KOGs (Tatusov et al. 2000, 2003; Tatusov, Koonin, and Lipman 1997).

Il existe 25 catégories COGs (Figure 4.3). Ces différentes catégories COGs sont regroupées en 4 grandes classes de fonctions : les processus cellulaires et de signalisation, stockage et processus d'information, métabolisme, fonction faiblement caractérisée.

Identifiant de la catégorie	Catégorie
A	Traitement et modification de l'ARN
B	Structure et dynamique de la chromatine
C	Production et conversion d'énergie
D	Contrôle du cycle cellulaire, division cellulaire, partitionnement chromosomique
E	Métabolisme et transport des acides aminés
F	Transport et métabolisme des nucléotides
G	Transport et métabolisme des carbohydrates
H	Métabolisme des coenzymes
I	Métabolisme des lipides
J	Traduction
K	Transcription
L	Réplication et réparation
M	Paroi cellulaire/Membrane/Biogenèse de l'enveloppe
N	Motilité cellulaire
O	Modification post-traductionnelle, rotation des protéines, fonctions chaperon
P	Transport et métabolisme d'ions inorganiques
Q	Structure secondaire
T	Transduction du signal
U	Trafic intracellulaire, sécrétion et transport vésiculaire

V	Mécanismes de défense
W	Structures extracellulaires
Y	Structure nucléaire
Z	Cytosquelette
R	Prédiction des fonctions générales uniquement
S	Fonctions inconnues

Figure 15 : Table de correspondance des différentes catégories COGs et KOGs de la base de données COGs.

En bleu clair sont renseignées les catégories COGs appartenant à la classe de fonctions des processus cellulaires et de signalisation, en rouge, les catégories COGs appartenant à la classe de fonctions du stockage et processus d'information, en vert, celles appartenant au métabolisme, et enfin en bleu foncé, les catégories COGs dont les fonctions sont faiblement caractérisées.

Les abondances en reads de chaque grande classe de fonction COG dans l'ensemble des microbiomes de lézards insectivores à gauche, de lézards omnivores à droite, sont représentées sur les diagrammes circulaires ci-dessous (Figure 16) :

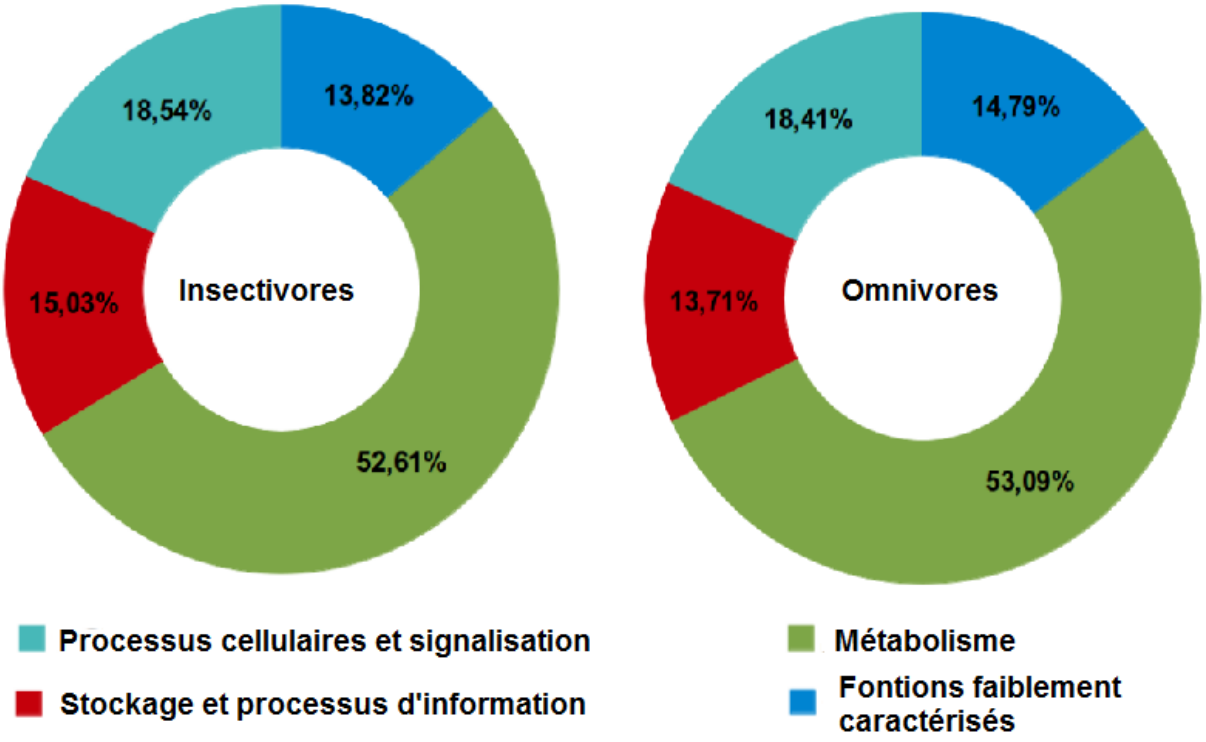


Figure 16 : distribution comparable des reads par classe de fonctions pour les lézards insectivores et omnivores.

On constate que la distribution des reads par classe de fonction est équivalente entre la population de lézards insectivores et la population de lézards omnivores, ce qui était attendu. En effet, on ne s'attend pas à ce qu'un changement de régime alimentaire modifie toutes les fonctions du microbiome intestinal, dans la mesure où le microbiome intestinal est impliqué dans d'autres grands phénomènes que la digestion. Par exemple, ce microbiome est fortement impliqué dans l'immunité chez la souris (Ritchie 2006; Eberl 2010), mais aussi dans la vie des microbes eux-mêmes (Lloyd-Price, Abu-Ali, and Huttenhower 2016).

Suite à ce constat, nous avons choisi d'étudier les microbiomes à un niveau plus détaillé, celui des catégories COGs. Dans un premier temps, nous avons étudié les différences de répartitions des reads au sein des catégories COGs entre les lézards insectivores et les lézards omnivores.

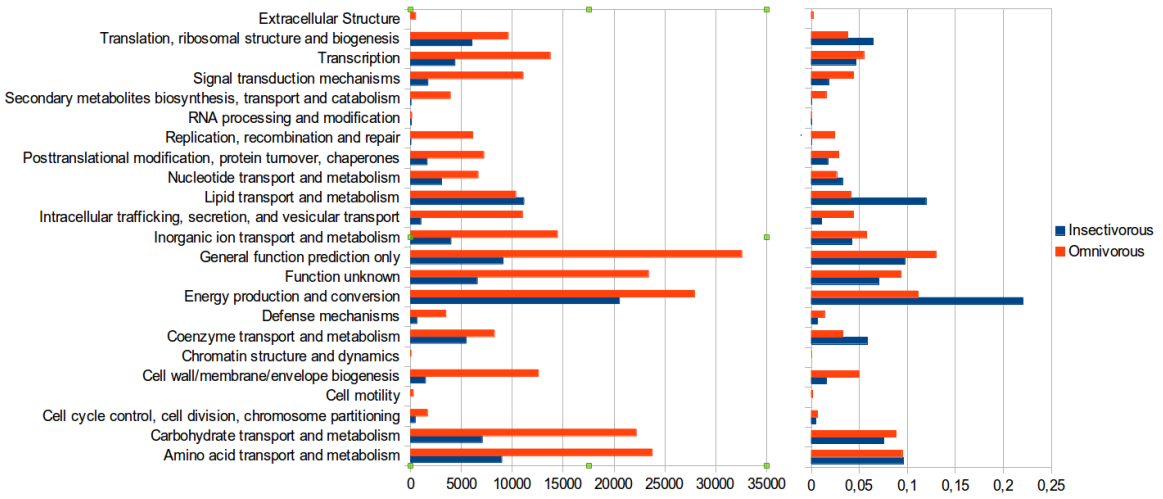


Figure 17 : Nombre (à gauche) et proportion de reads (à droite) par catégorie COG et par population de lézards.

Sur la Figure 17 est représentée à gauche l'abondance absolue de reads par catégorie COG et à droite, l'abondance relative de reads par catégorie COG. On constate des différences notables entre les deux populations. Tout d'abord on peut noter le fait que la population de lézards insectivores possède moins de reads annotés que la population de lézards omnivores. Par ailleurs, en proportion du nombre de reads annotés, les microbiomes de lézards insectivores sont constitués de deux fois plus de reads correspondant à des fonctions liées à la production et à la conversion d'énergie, ainsi qu'au transport et au métabolisme des lipides que les microbiomes de lézards omnivores. En revanche, ces derniers possèdent davantage de reads associés à la membrane et à la paroi cellulaire, à la prédiction de fonctions générales, au transport et au métabolisme des carbohydrates, à la transcription et aux mécanismes de transduction du signal, que les microbiomes de lézards insectivores.

Ces résultats seront à confirmer de deux façons :

- en étendant cette analyse à l'ensemble des 32 échantillons, car l'ajout d'individus permettra d'obtenir des statistiques.

- par l'annotation des ORFs prédites, qui devraient permettre d'annoter davantage de contenu génétique. En effet, il est difficile de prédire des fonctions sur des reads : sur les 79 965 057 reads fournis à MGRAST, seuls 9 274 155 reads ont été annotés, soit 11,6% de l'ensemble.

4.3 Impact du régime alimentaire du lézard sur les voies métaboliques

Nous avons cherché à analyser les fonctions présentant des différences d'abondances observées à l'échelle des catégories COGs entre les lézards insectivores et omnivores, au moyen de cartes de voies métaboliques (Huttenhower and Human Microbiome Project Consortium 2012; Yatsunenko et al. 2012). Nous souhaitons savoir si les microbiomes de lézards insectivores et ceux de lézards omnivores empruntent les mêmes voies métaboliques, ou si certaines voies sont exclusives. Tout d'abord, pour le calcul des abondances par population de lézards a été effectué par MG-RAST. La représentation de ces abondances sur une carte métabolique globale a été aussi réalisée avec ce serveur. Cela est représenté sur la carte Kegg (Kanehisa et al. 2016, 2017; Kanehisa and Goto 2000) suivante, Figure 18 (une carte kegg est une carte représentant des voies métaboliques) :

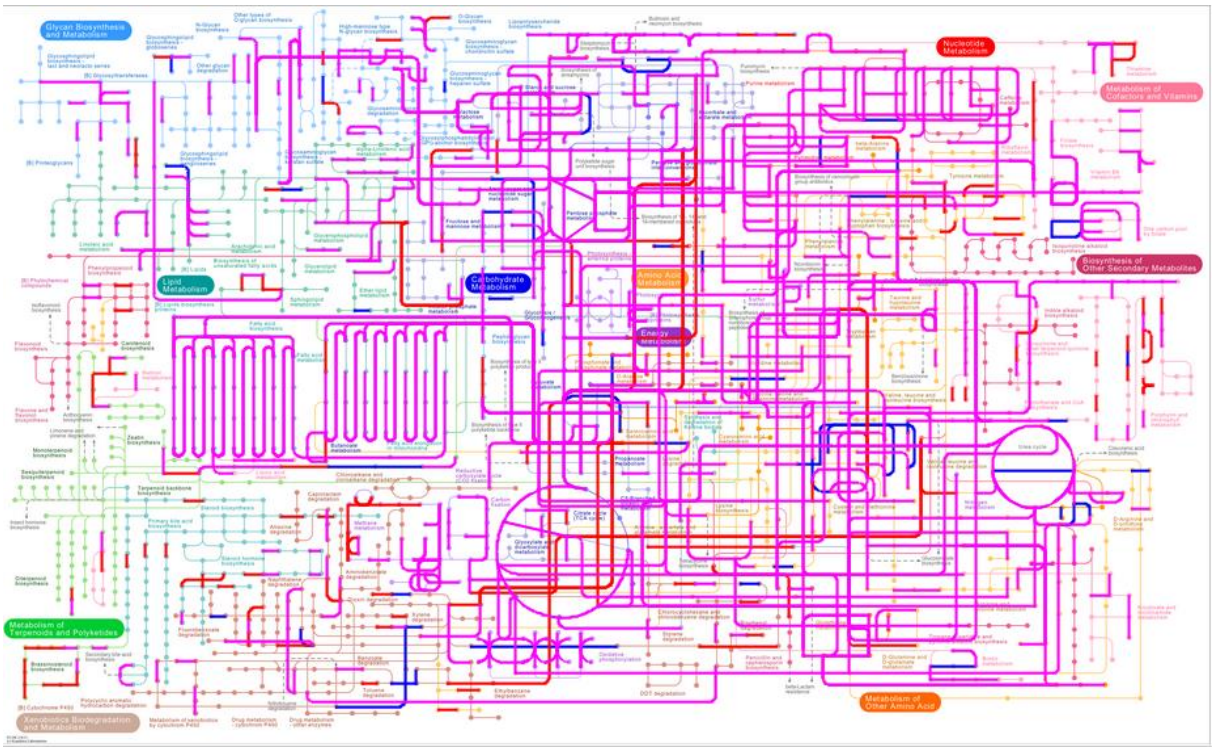


Figure 18 : carte des voies métaboliques pour les 12 individus insulaires.

(Kegg Map MGRAST, basée sur la comparaison avec des génomes de références - pourcentage d'identité ≥ 90%, cover ≥ 100 paires de base, E-value ≤ 10^{-5}). Les voies exclusivement empruntés par les

lézards omnivores sont rouges, celles empruntées exclusivement par les lézards insectivores sont bleues, celles empruntées par les deux populations sont violettes.

On observe sur cette carte qu'une majorité des voies métaboliques sont empruntées par les deux populations de lézards. Cela s'explique certainement par le fait que la majorité de ces voies métaboliques ne sont pas exclusives de la digestion des plantes. Par ailleurs, en regardant les tables d'abondance par enzyme obtenues grâce à MG-RAST, on observe que la majorité des différences entre les microbiomes des lézards insectivores et ceux des lézards omnivores semble être des différences d'abondance. Par exemple, l'exochitinase est plus abondante chez les insectivores de Pod Kopište (24 289 reads) que chez les omnivores de Pod Mcaru (18 617 reads). Cela est assez attendu, dans la mesure où les insectes contiennent beaucoup de chitine.

Ces différences d'abondance peuvent être quantifiées et représentées sur des cartes métaboliques de voies métaboliques individuelles. Pour cela, nous avons fourni à MGRAST les reads non ciblés de nos échantillons. MGRAST a annoté les reads puis compté le nombre de reads de lézards insectivores ainsi que le nombre de reads de lézards omnivores, pour chaque enzyme. A l'aide des tables d'abondance obtenues, nous pouvons ensuite normaliser les abondances en reads par enzyme des lézards insectivores et omnivores.

La normalisation est effectuée de la façon suivante pour chaque enzyme *i* :

$$Ab_{n,Di} = \frac{Ab_{Di}}{Ab_D}$$

où $Ab_{n,Di}$ est l'abondance normalisée en reads de l'enzyme *i* pour le régime alimentaire *D*, où Ab_{Di} est l'abondance en reads de l'enzyme *i* pour le régime alimentaire *D*; et où Ab_D est l'abondance totale en reads des microbiomes du régime alimentaire *D*.

On obtient une table d'abondances normalisées par voie métabolique (Figure 19).

Ecnumber	Abondance normalisée (PSK)	Abondance normalisée (PSM)
5.3.3.8	2,11E-06	1,24E-06
6.2.1.20	2,11E-06	2,49E-06
1.3.8.9	6,87E-06	8,21E-06
1.18.1.1	1,98E-05	1,02E-05
1.14.15.3	4,49E-06	1,12E-05
1.3.8.8	5,55E-06	1,37E-05
2.3.1.21	1,19E-05	1,69E-05
1.3.3.6	2,4844E-05	3,19E-05
1.3.8.6	3,67E-05	5,03E-05
1.1.1.35	2,91E-05	5,28E-05
1.14.14.1	2,51E-05	6,7E-05
1.18.1.3	2,72E-05	8,21E-05
1.3.8.7	0,000165186	0,0001787151
2.3.1.16	0,000249497	0,0002690683
4.2.1.17	0,0003446441	0,0004124387
1.2.1.3	0,0006681444	0,0006583587
1.3.8.1	0,000958079	0,0011409891
2.3.1.9	0,0010947208	0,0014015945
1.1.1.1	0,0026006888	0,0023822876
6.2.1.3	0,003880418	0,0035113291

Figure 19 : Table d'abondances normalisées de la voie métabolique de la dégradation des acides gras.

La première colonne correspond aux identifiants des enzymes. La seconde colonne correspond à l'abondance normalisée des enzymes pour l'ensemble des microbiomes de lézards insectivores. La deuxième colonne correspond à l'abondance normalisée des enzymes pour l'ensemble des microbiomes de lézards omnivores.

Une fois les tables d'abondance normalisées, l'abondance en reads par enzyme est discrétisée, en colorant l'abondance normalisée en fonction de son ordre de grandeur comme sur l'échelle ci-dessous, afin d'obtenir une carte dont les enzymes sont colorées en fonction de l'ordre de grandeur de l'abondance (Figure 20). Nous avons choisi de regrouper les enzymes par ordre de grandeur, parce que la discrétisation permettait de regrouper les enzymes par couleur en fonction d'abondances et donc de pouvoir visualiser les éventuelles différences. Cependant les analyses statistiques qui suivent sont réalisées sur les abondances normalisées et non sur les abondances discrétisées.

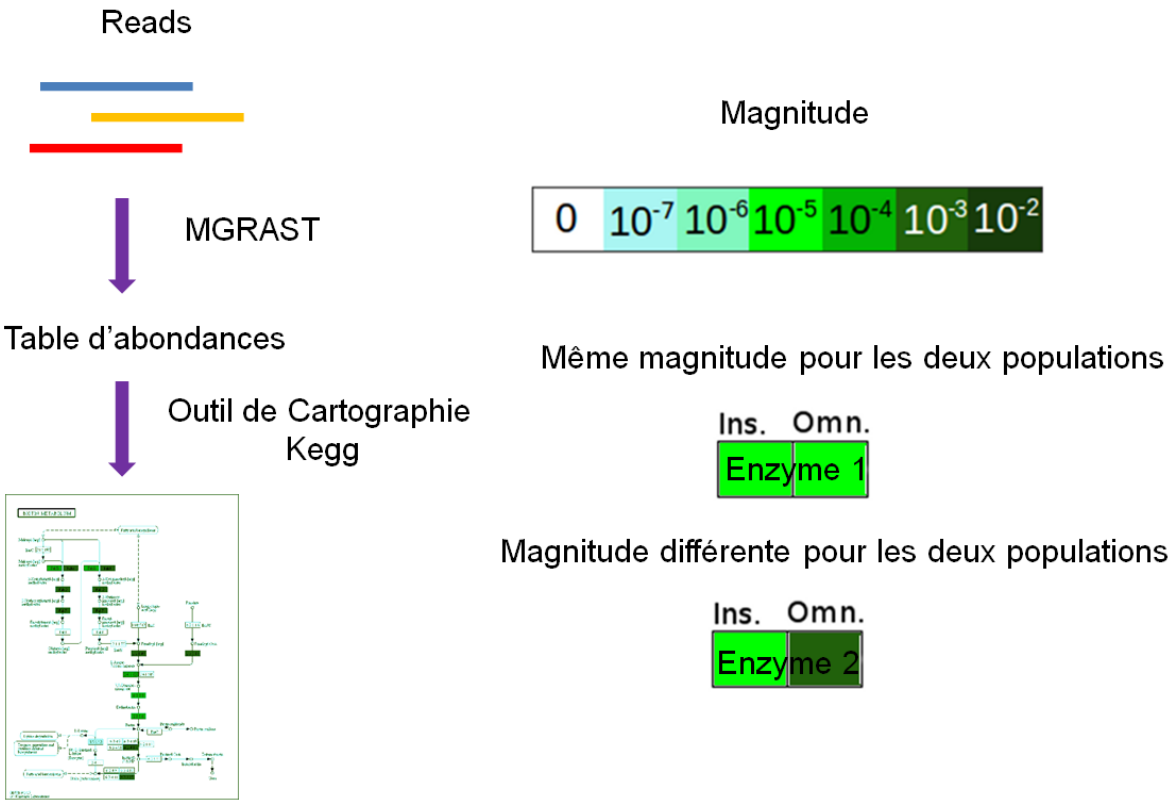


Figure 20 : Construction d'une carte Kegg à partir de reads.

A gauche, sont représentées les étapes permettant d'obtenir une carte métabolique colorée en fonction de l'abondance et du régime alimentaire, à partir des reads des microbiomes intestinaux de lézards. A droite, est donné l'ordre de grandeur des abondances normalisées des enzymes, et est expliqué le code couleur des cartes métaboliques. Pour chaque enzyme, on représente son abondance pour les deux types de régimes alimentaires par un rectangle dont la partie gauche représente l'abondance chez les insectivores, et la droite celle des omnivores.

Cette analyse permet de bien visualiser les différences entre les deux populations (insectivore et omnivore) sur des voies métaboliques données. Par exemple, si on s'intéresse au métabolisme de la pyrimidine (qui est le métabolisme associé à la synthèse et à la dégradation de la base azotée pyrimidine, fonction essentielle des organismes, et indépendante *a priori* du régime alimentaire), on s'attend à peu ou pas de différences entre les deux populations (il n'y a pas de raison que le régime alimentaire ait un impact sur ce métabolisme). Effectivement, on n'observe que deux différences d'ordre de grandeur (Figure 21). L'enzyme 4.2.1.70, qui correspond à la pseudourylate synthase, a une abondance normalisée de l'ordre

de grandeur de 10^{-7} chez les insectivores, et de 10^{-5} chez les omnivores. Elle n'est impliquée que dans cette voie métabolique. L'enzyme 2.7.1.74, qui correspond à la deoxycytidine kinase, a une abondance normalisée de l'ordre de grandeur de 10^{-7} chez les insectivores et de 10^{-6} chez les omnivores. Cette enzyme est aussi impliquée dans le métabolisme de la purine. Ces deux enzymes ne sont impliquées qu'une seule fois dans ce métabolisme.

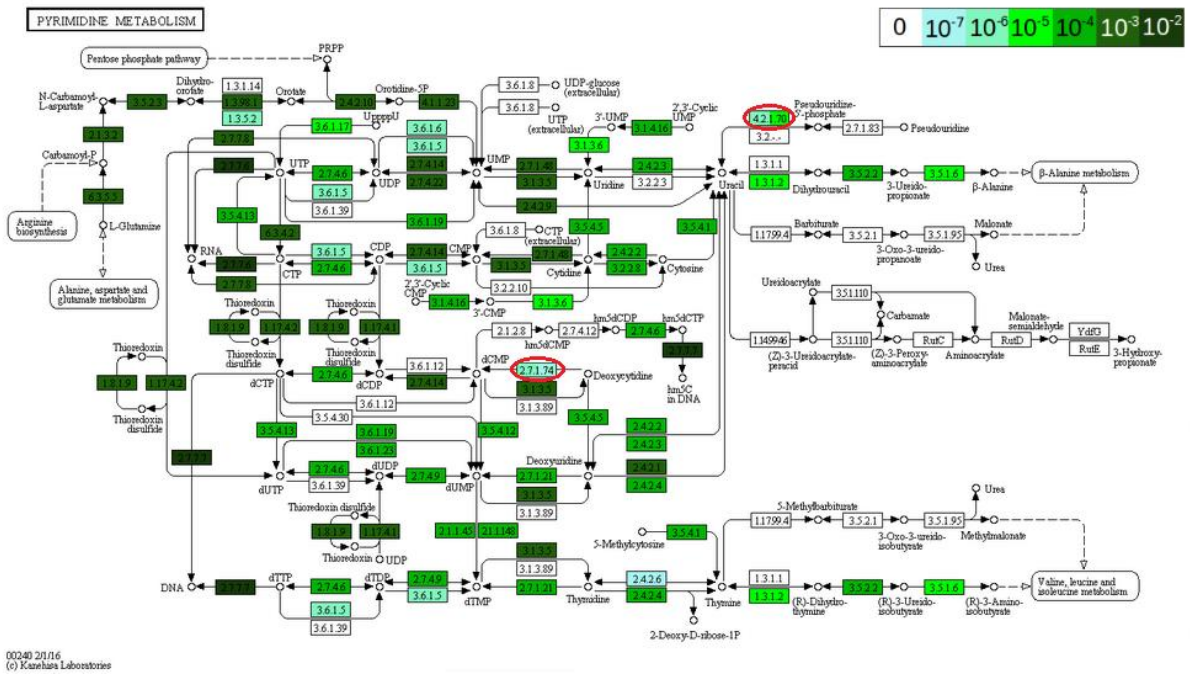


Figure 21 : Carte métabolique Kegg comparant le métabolisme de la pyrimidine des insectivores et des omnivores.

Le code couleur et l'échelle de magnitude de cette carte sont identiques à ceux de la Figure 18. Les cercles rouges signalent la présence d'enzymes dont l'abondance est d'ordre de grandeur différent entre l'ensemble des microbiomes insectivores (indiquée à gauche de chaque rectangle) et l'ensemble des microbiomes omnivores (à droite).

En revanche, on s'attend à ce que la dégradation des acides gras soit affectée par le changement de régime alimentaire. En effet, les polysaccharides provenant des plantes, que l'hôte vertébré ne peut pas dégrader seul, sont pris en charge par les microbes, ce qui génère des chaînes courtes d'acides gras (Holscher 2017; Solden et al. 2017). Ces chaînes courtes d'acide gras présentes dans le milieu peuvent être ensuite dégradées par des microbes. Cette voie métabolique, qui est beaucoup moins

complexe (dans le sens où elle implique moins d'enzymes) que la précédente contient 3 enzymes présentant des différences d'ordre de grandeur d'abondance entre les microbiomes des lézards insectivores et omnivores (Figure 22). Deux de ces enzymes interviennent plusieurs fois dans la voie métabolique (6 fois pour 1.3.8.8, qui est une déhydrogénase n'intervenant que dans cette voie métabolique, et 2 fois pour 1.14.15.3, qui est une alkane 1-monooxygénase présente dans la voie de dégradation des acides gras, mais aussi dans le métabolisme de l'acide arachidonique, dans le métabolisme du rétinol, dans et dans la voie de la dégradation des Caprolactames).

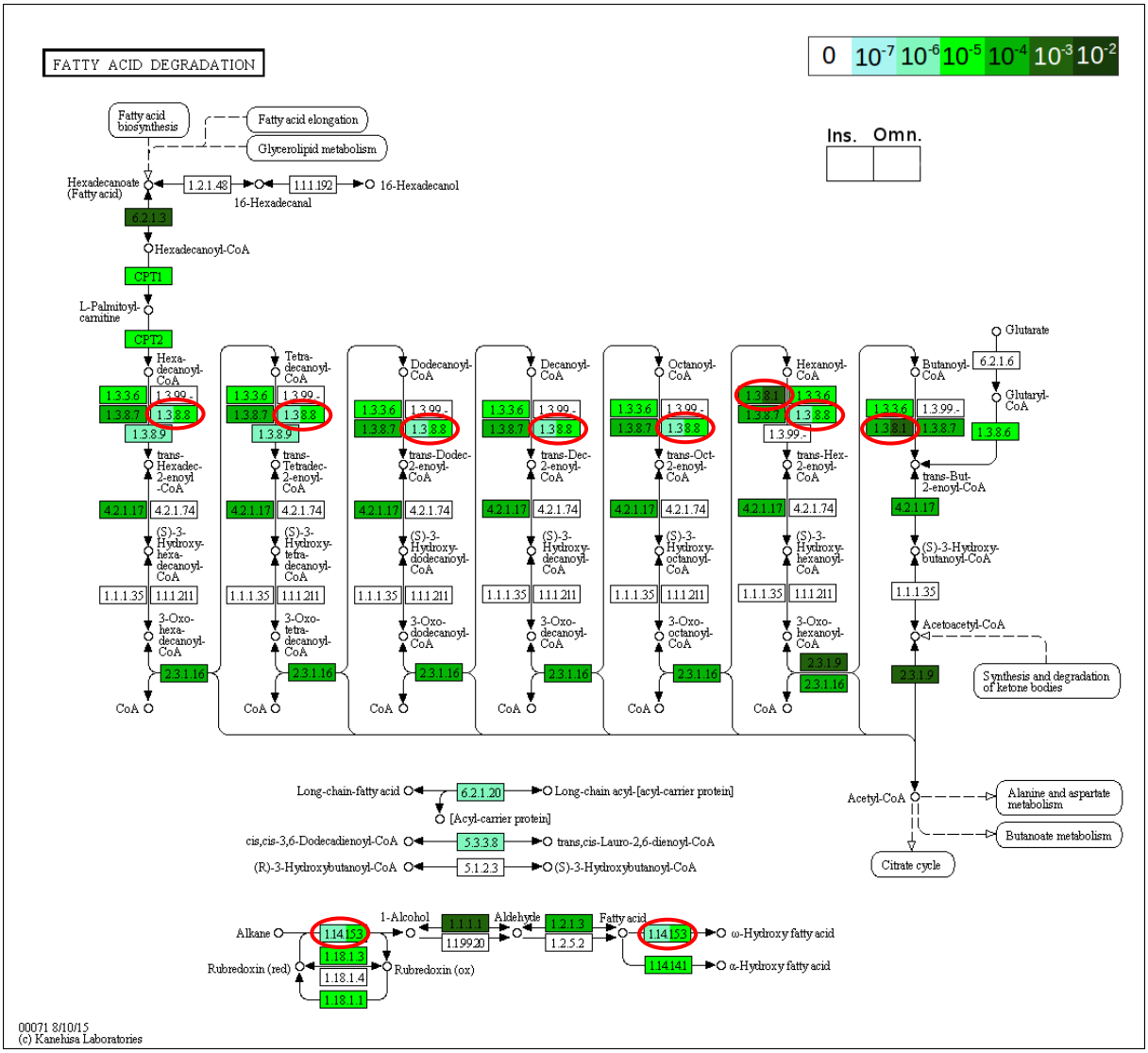


Figure 22 : Carte métabolique Kegg comparant le métabolisme de la dégradation des acides gras des microbiomes de lézards insectivores et des microbiomes de lézards omnivores.

Le code couleur est le même que sur la figure précédente.

Si ces cartes sont adaptées à la visualisation des différences entre insectivores et omnivores, elles ne permettent pas de quantifier cette différence, ni d'estimer sa significativité. Nous avons donc cherché à déterminer quelles sont les enzymes au sein des voies métaboliques qui présentent une différence d'abondance significative entre les deux groupes de microbiomes de lézards.

Pour cela, nous avons choisi d'appliquer des LDA ("Linear Discriminant Analysis", soit analyse linéaire discriminante en français) sur les tables d'abondance des enzymes. Ces abondances ont été normalisées (la somme des abondances pour un régime donné vaut 1) puis transformées par logarithme décimal pour mieux rendre compte des ordres de grandeur (plus exactement, des différences sur des valeurs faibles, ici comprises entre 10^{-2} et 10^{-7}). La matrice obtenue prend pour variables le logarithme décimal des abondances normalisées des enzymes, et pour échantillons, les douze microbiomes des lézards étudiés.

La LDA est une méthode avec laquelle il est possible d'identifier les variables qui discriminent plusieurs groupes les uns des autres. La LDA est une généralisation de l'analyse linéaire de Fisher, et relève de l'apprentissage automatique (« machine learning ») (Tarca et al. 2007). Le résultat de la LDA peut être soit utilisé comme un classificateur linéaire (ce qui est le cas ici), soit comme une étape effectuée en amont d'une classification pour réduire le nombre de dimensions (dimensions déterminées par le nombre de variables, qui est plus important que le nombre d'échantillons, ce qui est un problème pour un certain nombre de classificateurs).

Des LDA pour chaque voie métabolique ont été effectuées. Cependant un certain nombre de ces analyses n'ont pas abouti parce que la variabilité au sein de certaines variables était insuffisante (ce qui est bloquant pour tester les autres variables de la voie métabolique). Sur les 174 voies métaboliques présentes dans nos microbiomes, 92 ont donné des résultats. Les voies métaboliques dans lesquelles il est possible de distinguer les lézards omnivores des lézards insectivores sur la base des abondances des enzymes, sont données dans la table suivante. On considère que la distinction est possible dès lors que la validation croisée donne un score supérieur à 68%, et l'on considère que la contribution de l'enzyme est significative si sa contribution à la LDA est supérieure à 2 en valeur absolue (comme l'outil LefSe). On peut constater que la majorité des enzymes permettant de discriminer les microbiomes

des lézards insectivores et ceux des omnivores sont des enzymes impliquées dans la dégradation de molécules. L'enzyme participant le plus au modèle est l'uréase (3.5.1.5), une enzyme qui catalyse la réaction de transformation de l'urée en dioxyde de carbone et ammoniaque (Figure 23).

Voie métabolique	Enzymes impliquées	Score (validation croisée)	Biologie des molécules
dégradation de l'Atrazine	Uréase (3.5.1.5)	81.3%	Polluant (pesticide)
résistance au Betalactame	beta-lactamase (3.5.2.6)	75%	Antibiotique à large spectre
Biosynthèse des ansamycines	Transkétolase (2.2.1.1)	68.8%	Métabolite secondaire (activité antimicrobienne)
Dégradation du caprolactame	Enzyme de la classe des oxydoréductases (1.1.1.35), Enzyme de la classe des lyases (4.2.1.17)	68.8%	Molécule toxique
Metabolisme du Dglutamine et du Dglutamate	Enzymes de la classe des ligases (6.3.2.8), des hydrolases (3.5.1.2), glutamates et glutamique déhydrogénase (1.4.1.3)	68.8%	Acide aminé (Dglutamine), Neurotransmetteur (Dglutamate)
Biosynthèse des glycosaminoglycanes, héparine, sulfate d'héparane	Enzymes de la classe des Transférases (2.4.1.133, 2.4.1.134, 2.4.1.224),	68.8%	Macromolécules glucidiques (glycosaminoglycanes), polysaccharides

Biosynthèse des alcaloïdes indoles	Enzyme de la classe des hydrolases (3.1.1.78)	68.8%	Hétérocycles souvent d'origine végétale
Biosynthèse des monobactame	Aspartate et aspartique semi-aldehyde déhydrogénase(1.2.1.11), enzymes de la classe des transférases (2.7.2.4 et 2.7.7.4), enzyme de la classe des lyases (4.3.3.7)	75%	antibiotique
Biosynthèse des novobiocines	enzymes de la classe des oxydoréductases (1.3.1.12), de la classe des transférases (2.6.1.1, 2.6.1.5, 2.6.1.57, 2.6.1.9)	68.8%	antibiotiques
Autres dégradations de glycanes	Enzymes de la classe des hydrolases et des glycosylases (3.2.1.18, 3.2.1.23, 3.2.1.24, 3.2.1.25, 3.2.1.45, 3.2.1.51, 3.2.1.52, 3.2.1.96, 3.5.1.26)	68.8%	Polysaccharides essentiels de la membrane cellulaire
Autres types de biosynthèse d'Oglycanes	Enzymes de la classe des hydrolases et des glycosylases (2.4.99.6, 2.4.99.1)	75%	Polysaccharides essentiels de la membrane cellulaire
Phosphorylations oxydatives	Inorganic disphosphatase (3.6.1.1), enzymes de la classe des hydrolases (1.6.5.3,	100%	Processus aérobique

	3.6.3.14, 3.6.3.6), de la classe des oxydoréductases (1.9.3.1, 1.10.2.2, 1.6.99.3), polyphosphate kinase (2.7.4.1), succinate déhydrogénase (1.3.5.1),		
Biosynthèse des phenylpropanoïdes	Enzyme de la classe des hydrolases et des glycosylases (3.2.1.21)	75%	Molécule de défense contre les herbivores et contre des microbes
Métabolisme des phosphonates et phosphinates	Enzyme de la classe des transférases et transaminases (2.6.1.37, 2.7.7.14, 2.7.7.15), de la famille des hydrolases (3.11.1.1), Phosphonopyruvate décarboxylase (4.1.1.82)	68.8%	
Voie de signalisation du récepteur Tcell	Enzymes de la classe des hydrolases (3.1.3.16), de la classe des transférases (2.7.10.2)	75%	
Biosynthèses de Tropane, pipéridine, et alcaloïdes de pyridine	Enzyme de la classe des transférases et transaminases (2.6.1.1, 2.6.1.57, 2.8.3.17),	68.8%	Composé azoté (tropane), alcaloïdes

	lysine décarboxylase (4.1.1.18)		

Figure 23 : Voies métaboliques dont les enzymes permettent de distinguer les microbiomes des lézards omnivores de ceux des lézards insectivores.

Il semblerait que la majorité des enzymes qui présentent des différences notables entre insectivores et omnivores soient des enzymes liées à la défense de l'organisme contre des microbes. Par exemple, la LDA appliquée à la voie métabolique de la résistance au Betalactame (antibiotique à large spectre) discrimine les microbiomes des lézards insectivores et ceux des omnivores en se basant sur l'enzyme bêta-lactamase (3.5.2.6) qui est une enzyme impliquée dans l'hydrolyse des bêta-lactames (cross validation : 75% de bonnes prédictions). En revanche, il ne semble pas y avoir de lien direct entre le régime alimentaire et la biologie de ces molécules.

Afin d'avoir une analyse plus complète, deux solutions sont envisageables : identifier les variables dont la variabilité est insuffisante pour l'analyse, ou alors, effectuer une autre analyse statistique telle que la permanova (Tang, Chen, and Alekseyenko 2016). La permanova est une analyse statistique non paramétrique multivariée, couramment utilisée pour comparer des groupes entre eux, notamment lorsque les conditions nécessaires pour appliquer une anova ne sont pas réunies (normalité des résidus et homogénéité de la variance). Ces deux pistes seront explorées.

4.4 Perspectives

Nous souhaitons par la suite développer une méthode permettant d'obtenir des résultats pour toutes les voies métaboliques, et de disposer d'une plus grande rigueur statistique pour déterminer quelles sont les enzymes permettant de discriminer les microbiomes de lézards insectivores des microbiomes de lézards omnivores.

Dans le cadre de l'annotation taxonomique de l'ARN 16S, nous avions observé qu'environ 50% des OTUs n'étaient pas annotées au niveau du genre (chapitre 3).

Nous nous sommes alors demandé si en assemblant les reads en contigs, en prédisant des ORFs puis en annotant ces ORFs taxonomiquement, nous ne pourrions pas obtenir davantage d'annotations taxonomiques qu'au moyen des reads des régions V4 de l'ARN16S. De la même façon, nous nous demandions s'il était possible d'avoir davantage d'annotations fonctionnelles sur la base des ORFs prédites après assemblage des « reads » en « contigs », qu'en partant des reads comme nous l'avions fait dans les analyses présentées dans ce chapitre. Nous cherchions ainsi à quantifier la matière noire (i.e. matériel métagénomique environnemental de fonctions et de taxonomie inconnues).

Le docteur Guillaume Bernard, post-doctorant au sein de l'équipe AIRE, a alors réalisé les tâches d'assemblage de microbiome. Il existe 3 stratégies différentes pour assembler des reads en contig (Figure 24).

a. b. c.

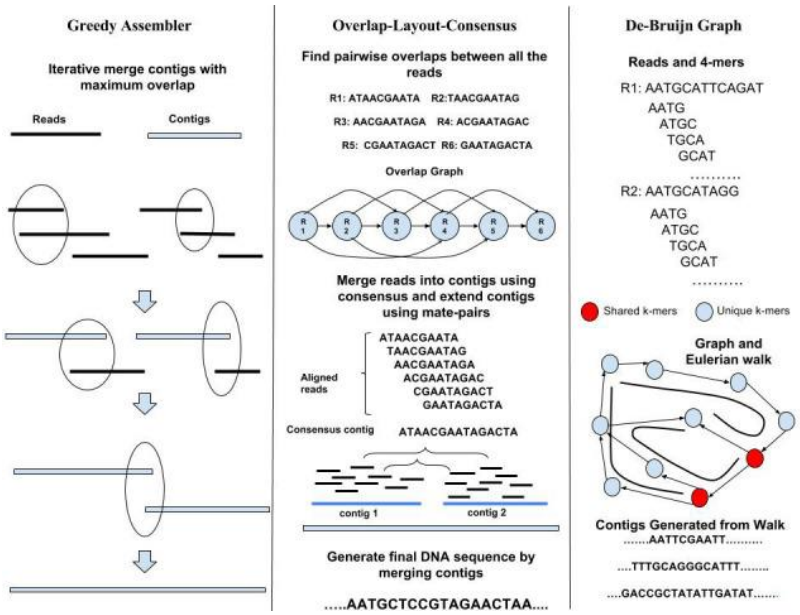


Figure 24 : Aperçu des différents types d'assemblage de novo (Figure tirée de l'article (Ghurye, Cepeda-Espinoza, and Pop 2016)*).*

a. Pour les assembleurs dits « gloutons » (ou « greedy »), les « reads » (en marron) qui se chevauchent le plus sont réunis en contigs (en bleu) de façon itérative. b. Les assembleurs dont l'approche est « Overlap layout-Consensus » qui construisent les contigs en trouvant des chemins sans ramification dans le graphique (b. « Overlap Graph »). L'assembleur prend ensuite la séquence consensus des reads qui se chevauchent, impliqués dans le chemin correspondant. c. D'autres assembleurs se basent sur des graphes de Bruijn: les reads sont coupés en segments courts et chevauchants (les « k-mers »), qui sont

ensuite organisé dans un graphe de de Bruijn en fonction des co-occurrences entre reads. Ce dernier type de graphe est simplifié, pour enlever les artefacts dus à des erreurs de séquençage, et les chemins sans ramification sont considérés comme des contigs.

Le docteur Guillaume Bernard, a ensuite prédit les ORFs (Open Reading Frames, cadres de lecture ouverts) sur l'ensemble des données métagénomiques (sur les 51 échantillons dont on possède le métagénome). Une ORF est un morceau de séquence qui part d'un codon d'initiation et qui se termine au premier codon stop rencontré. Pour cela, il a filtré les reads en fonction de leur qualité avec l'outil Trimmomatic (Bolger, Lohse, and Usadel 2014), puis a assemblé les reads avec Megahit (D. Li et al. 2015) en utilisant les paramètres par défaut. Ensuite, les protéines ont été prédites en utilisant FragGeneScan (Rho, Tang, and Ye 2010). Nous avons ainsi obtenu 461 000 gènes. L'annotation taxonomique de ces ORFs au niveau du phylum (similarité entre les séquences > 85%) a été réalisée avec l'algorithme LCA (« Lowest Common Ancestor ») de MEGAN 6 (Huson et al. 2007). L'annotation fonctionnelle de ces ORFs a été réalisée avec EggNog (Huerta-Cepas et al. 2016) contre l'ensemble de la base de données de cet outil sans seuil limite.

Ces traitements ont permis d'obtenir le résultat suivant pour un microbiome de *Podarcis sicula* (Figure 25). Nous sommes actuellement en train d'étendre cette analyse à tous les microbiomes.

	Porteur connu	Porteur inconnu	Total
Fonction connue	Nos connaissances 23,5%	Nouvelles lignées 50%	73,5%
Fonction inconnue	A quoi ça sert ? 1,4%	Inconnu 25,1%	26,5%
Total	24,9%	75,1%	100%

Figure 25 : Synthèse des gènes annotés.

Le tableau ci-dessus nous montre que 23,5% des gènes sont annotés fonctionnellement et taxonomiquement. 50% des gènes sont annotés

fonctionnellement mais pas taxonomiquement, et une minorité (1,4%) sont connus taxonomiquement mais pas fonctionnellement. En revanche, 1/4 des gènes sont non annotés (fonctionnellement et taxonomiquement). L'information du nombre de gènes non annotés est rarement renseignée. Afin d'avoir un ordre d'idée des pourcentages d'annotation trouvés dans d'autres études, nous avons contacté les responsables du projet TARA. Leurs échantillons provenant de l'océan austral ont permis de découvrir 80 à 90% de nouveaux gènes (donc 80 à 90% des gènes ne sont pas annotés en comparant aux bases de données). Dans notre étude, nous avons moins de fonctions non annotées, puisqu'environ 26% des gènes sont non annotés fonctionnellement. Cela est certainement dû au fait que les microbiomes intestinaux sont des milieux très étudiés, présentant une moins grande diversité de gènes que le projet TARA, mais surtout au fait que l'on n'utilise pas nécessairement les mêmes seuils (ce qui rend les résultats difficilement comparables entre ce projet et notre étude). Dans le projet TARA, 50% des gènes ne sont pas annotés taxonomiquement (au niveau du phylum)(Sunagawa, Coelho, Chaffron, Kultima, Labadie, Salazar, Djahanschiri, Zeller, Mende, Alberti, Cornejo-Castillo, Costea, Cruaud, d\textquoterightOvidio, et al. 2015). Dans le microbiome d'un lézard en revanche, 75,2% des gènes ne sont pas annotés taxonomiquement (au niveau du phylum).

5. Utilisation de réseaux de similarité dans l'étude des microbiomes intestinaux

Dans ce chapitre, nous présentons des méthodes exploratoires permettant d'étudier des processus d'évolution, utilisant des réseaux de similarité dont les développements sont préliminaires.

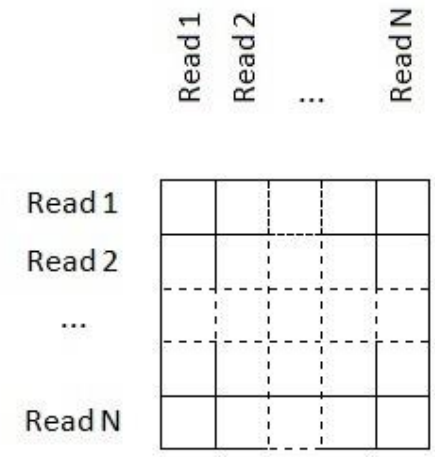
5.1 Les réseaux de similarité de séquences

5.1.1 Présentation des réseaux de similarité de séquences (RSS)

L'utilisation de réseaux de similarité de séquences (RSS) est apparue dans les années 1990. Les RSS avaient été proposés pour analyser les nouvelles données moléculaires obtenues grâce aux avancées en matière de technique de séquençage, ainsi que la réduction du coût de ces techniques. Par exemple, ces méthodes ont été proposées pour prédire les fonctions de gènes et les interactions protéine-protéine (Enright et al. 1999; Tatusov et al. 1997; Watanabe and Otsuka 1995).

Un réseau de similarité est un réseau dans lequel deux nœuds, représentant deux séquences, sont reliés par une arête si les séquences sont considérées comme suffisamment similaires entre elles. Il s'agit donc de comparer deux à deux chaque séquence du jeu de données afin de pouvoir construire le graphe comprenant toutes les séquences du microbiome d'un *P. sicula*. Dans le cadre de notre étude, BLAST (BLAST n.d.; Cock et al. 2015) est utilisé pour effectuer les comparaisons de séquences.

Plus formellement, un RSS est un graphe dans lequel chaque nœud est une séquence et une arête relie deux nœuds qui sont semblables à un seuil donné. Ce seuil concerne la couverture, le pourcentage d'identité et la E-value. La méthode de construction d'un RSS est détaillée dans les figures ci-dessous (Figure 26).



a) Matrice schématisant la comparaison de toutes les séquences d'un jeu de données contre elles-mêmes (blast « all-against-all »).

qseqid	sseqid	evalue	pident	bitscore	qstart	qend	qlen	sstart	send	slen
seq1	seq1	0.0	100.00	952	1	515	515	1	515	515
seq1	seq2	0.0	100.00	837	63	515	515	1	453	530
seq1	seq3	3e-180	98.34	636	1	362	515	362	1	554
seq2	seq2	3e-165	99.69	586	42	361	515	1	320	320
seq2	seq1	7e-157	100.00	558	214	515	515	1	302	379
seq2	seq3	2e-156	100.00	556	191	491	515	1	301	301
seq3	seq3	5e-153	99.34	545	58	358	515	1	301	301
seq3	seq2	7e-137	100.00	492	250	515	515	1	266	578
seq3	seq1	6e-118	99.57	429	281	515	515	1	235	300

b) Exemple de fichier de sortie BLAST tabulaire. Les comparaisons sont effectuées dans les deux sens : les résultats de BLAST ne sont pas symétriques. Sont encadrées en rouge les comparaisons d'une séquence contre elle-même.

Descripteurs d'alignements locaux fournis par l'outil BLAST	**abbréviations**
Identifiant de la séquence requête (query sequence identification)	qseqid
Identifiant de la séquence cible (subject sequence identification)	sseqid
E-value	evalue
Pourcentage d'appariements identiques	pident
Début de l'alignement sur la séquence requête (query start)	qstart
Fin de l'alignement sur la séquence requête (query end)	qend
Longueur de l'alignement sur la séquence requête (query length)	qlen
Début de l'alignement sur la séquence cible (subject start)	sstrart
Fin de l'alignement sur la séquence cible (subject end)	send
Longueur de l'alignement sur la séquence cible (subject length)	slen

Nombre d’appariements identiques	nident
Bit score	bitscore
Longueur	length
Nombre d’appariements de score positif	positive
Pourcentage d’appariements de score positif	ppos
Nombre de mésappariements	mismatch
Nombre de trous ouverts	gapopen
Nombre de trous	gaps

c) Descripteurs d’alignements locaux fournis par BLAST

Figure 26 : Méthode de construction d'un réseau de similarités de séquences à l'aide de l'outil BLAST.

La sortie de BLAST peut donc être vue comme un graphe, dans laquelle chaque ligne est une arête potentielle. Par exemple, dans le cas de l’étude de la diversité génétique à l’aide de réseaux de reads, on considère qu'une arête est tracée entre deux nœuds (représentant deux reads) si ces reads sont similaires à 90% ou plus (pourcentage d’identité), sur au moins 80% de la longueur de l’un des reads (couverture = longueur de l’alignement/longueur totale de la séquence) avec une E-value inférieure ou égale à 10^{-5}. Ou encore, dans le cas où les nœuds représentent des ORFs, une arête est tracée entre deux ORFs, si ces ORFs présentent au moins 95% de similarité sur au moins 80% de la longueur de chacune des ORFs avec une E-value inférieure ou égale à 10^{-5}. Cela implique donc de filtrer les sorties BLAST obtenues pour ne conserver que les arêtes satisfaisant ces trois exigences.

Une visualisation graphique de l’exemple de sortie BLAST proposé dans la Figure 26 b) est présentée ci-dessous (Figure 27) :

a) b)

Figure 27 : Représentations graphiques de la sortie BLAST présentée dans la Figure 26.

a) Représentation sous forme de graphe orienté de la sortie BLAST (Figure 26)

b) Représentation de la sortie après symétrisation (élimination des boucles et des arêtes multiples)

Les boucles représentées sur la Figure 27 a) correspondent aux similarités encadrées en 26 b). Il s'agit des similarités d'une séquence à elle-même, ce qui n'est pas informatif, et justifie la suppression de ces boucles.

Par ailleurs, dans la Figure 27. a), les arêtes du graphe sont orientées, car la comparaison de BLAST d'une séquence requête à une séquence cible n'est pas symétrique. Cela signifie que les comparaisons effectuées dans les deux sens ne donnent pas les mêmes résultats (zones de similarités, scores tels que la E-value, bitscore,... légèrement différents d'une comparaison à l'autre). Cette asymétrie n'ayant aucun sens biologique, les réseaux sont donc symétrisés : sur les deux arêtes reliant deux nœuds, celle possédant les meilleurs scores est conservée, et la notion d'orientation des arêtes est supprimée (Figure 27 b).

Dans le cadre de cette thèse, nous avons exploité deux types de réseaux de similarités de séquences : les réseaux de similarités d'Open Reading Frames (ORFs), et les réseaux de similarités de reads (Boon et al. 2015) que nous avons développés au sein de l'équipe AIRE en 2015 (Völkel et al. 2016) [ref volkel]. Les réseaux de similarité d'Open Reading Frames (ORFs) sont une façon de décrire le microbiome et de détecter les transferts latéraux de gènes (Bapteste, Bicep, and Lopez 2012). Ils permettent aussi de mesurer la diversité microbienne et les relations au sein de communautés microbiennes (Bapteste, Bicep, and Lopez 2012; Forster et

communautés microbiennes (Bapteste, Bicep, and Lopez 2012; Forster et al. 2015). Ces réseaux se concentrent sur la diversité génétique présente dans les microbiomes. Par exemple, ils permettent de trouver des variants d'un même gène dans un microbiome (séquences suffisamment proches pour être reliées entre elles dans le réseau, mais qui ne sont pas identiques). Ils permettent en outre d'analyser la diversité des contextes génomiques. Par exemple, on peut se demander si l'on retrouve une même séquence ou un même gène dans différents contextes génomiques (cf. la transposase de la Figure 35 se trouve dans les différents contextes génomiques représentés par les longs filaments partant de la boucle centrale)(Völkel et al. 2016).

Dans ce chapitre, nous souhaitons donc répondre aux questions suivantes : quelle est la diversité génétique et génomique présente dans chacun des microbiomes intestinaux de *Podarcis sicula* ? Y-a-t-il une différence en terme de diversité génétique entre les microbiomes intestinaux de *Podarcis sicula* insectivores et omnivores ? Y-a-t-il une différence en terme de diversité génomique entre les microbiomes intestinaux de *Podarcis sicula* insectivores et omnivores ?

5.1.2 Les réseaux de similarités d'ORFs

Une des premières utilisations des réseaux de similarités d'ORFs, était la définition de groupes COGs de familles d'homologues et faciliter la prédiction de fonctions sur un grand nombre de gènes, en se basant sur l'homologie (Tatusov et al. 2000, 1997). Nous nous intéressons ici à la définition de familles de gènes à l'aide de réseaux de similarités d'ORFs. Une famille de gènes est un ensemble de plusieurs gènes similaires, formés par réplication d'un seul et même gène d'origine. Dans la mesure où les gènes d'une famille de gènes sont similaires, ils sont donc tous connectés les uns aux autres dans le réseau d'ORFs, soit de façon directe (i.e. les deux séquences sont reliées par une arête) soit de façon indirecte (i.e. les deux séquences ne sont pas reliées par une arête, mais il existe un chemin passant par d'autres séquences permettant de parcourir le réseau pour aller d'une séquence à l'autre). Une famille de gènes, étant donc un ensemble de nœuds connectés entre eux, forme une composante connexe.

Une utilisation des RSS présentée dans cette thèse correspond à la quantification de la mobilité des familles de gènes d'un microbiome par des EGM (Éléments Génétiques Mobiles). Les éléments génétiques mobiles sont des séquences d'ADN capables de se déplacer de façon autonome aussi bien au sein d'un génome que d'un génome à un autre. La taille des EGMs varie de quelques centaines de paires de bases à plus de 100 000 paires de bases (Binnewies et al. 2006). Il existe différents types d'EGM dont les virus, les plasmides et les intégrons pour n'en citer que trois. Cette étude est présentée plus en détail dans la partie 5.1.4.

5.1.3 Les graphes bipartis

Un graphe biparti comprend deux types prédéfinis de nœuds : les nœuds de type I et les nœuds de type II. Seuls des nœuds de type différent peuvent être connectés par une arête. Les graphes bipartis permettent de résumer quels gènes sont partagés par quels génomes (Watson et al. 2017) si on représente les génomes par des nœuds de type I et les familles de gènes (définies dans le réseau d'ORFs) par des noeuds de type II.

On peut également choisir de représenter à l'aide d'un graphe biparti des hôtes (nœuds de type I) et les classes de microbes (nœuds de type II) présentes dans les microbiomes de ces hôtes. Cela permet de déterminer quel est le microbiote ubiquitaire (les microbes partagés par tous les hôtes étudiés), quels sont les microbes partagés uniquement par des lézards insectivores, ceux partagés uniquement par des lézards omnivores, et donc d'étudier la transmission de gènes et de microbes au sein des holobiontes (Corel et al. 2016).

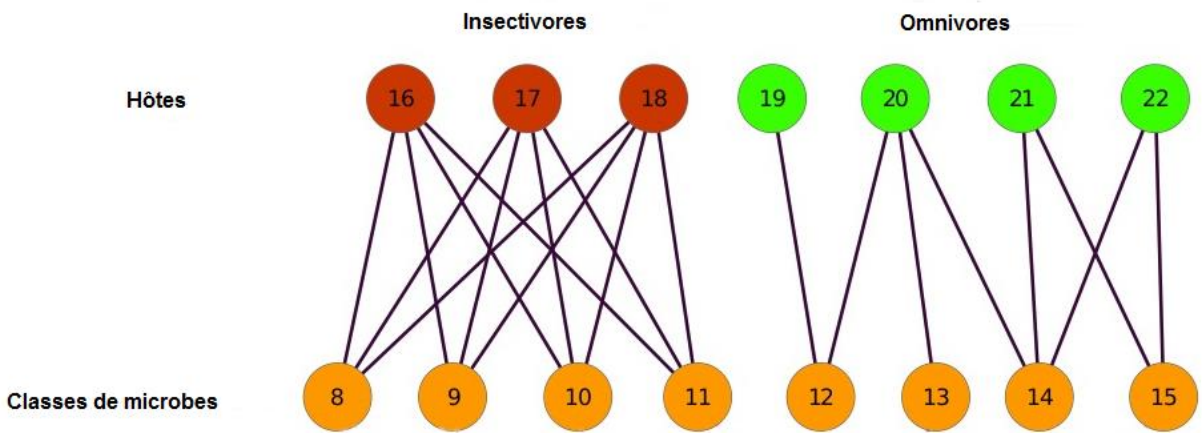


Figure 28 : Exemple de graphe biparti Hôtes-microbes.

Les nœuds rouges correspondent aux lézards insectivores, les verts aux omnivores, les nœuds jaunes correspondant à des OTUs.

Dans l'exemple ci-dessus (Figure 28), les classes microbiennes 8, 9, 10 et 11 sont spécifiques d'un régime insectivore et représentent le microbiote ubiquitaire de la population de lézards insectivores, ce qui signifie que tous les microbiomes de lézards insectivores possèdent ces classes microbiennes, et que ces dernières sont exclusives des microbiomes de lézards insectivores. Les classes 12, 13, 14, et 15, bien que spécifiques d'un régime omnivore, ne sont pas ubiquitaires.

Nous proposons l'emploi de tels graphes pour étudier les microbiomes de lézards sous différents aspects, dans la partie 5.1.4.

5.1.4 Etude des règles d'introgression et de transmission avec des réseaux (chapitre de livre n°2)

Nous présentons ici le chapitre que nous avons rédigé pour l'ouvrage 'Experimental and theoretical modes of transmission' édité par Theresa Coque et Fernando Baquero aux éditions ASM Press. L'ouvrage étant sous presse à l'heure actuelle, le formatage n'est pas définitif.

5. Experimental and Theoretical Modes of Transmission

TRACKING THE RULES OF TRANSMISSION AND INTROGRESSION WITH NETWORKS

Chloé Vigliotti*, Cédric Bicep*, Eric Bapteste, Philippe Lopez and Eduardo Corel

Institut de Biologie Paris-Seine. UMR 7138 CNRS-UPMC Evolution. Paris, France

Abstract

Understanding how an animal organism and its gut microbes form an integrated biological organisation, known as holobiont, is becoming a central issue in biological studies. Such an organisation inevitably involves a complex web of transmission processes that occur on different scales in time and space, across microbes and hosts. Network-based models are introduced in this chapter to tackle aspects of this complexity, and to better take into account vertical and horizontal dimensions of transmission. Two types of network-based models are presented, sequence-similarity networks and bipartite graphs. One interest of these networks is that they can consider a rich diversity of important players of microbial evolution that are usually excluded from evolutionary studies, like plasmids and viruses. These methods bring forward the notion of 'gene externalization', which is defined as presence of redundant copies of prokaryotic genes on Mobile Genetic Elements (MGE), and therefore emphasizes a related although distinct process from lateral gene transfer between microbial cells. This chapter introduces guidelines to the construction of these networks, their analysis, and illustrates their possible biological

interpretations and uses. The application to human gut microbiomes shows that sequences present in a higher diversity of MGE have both biased functions and a broader microbial and human host range. These results suggest that an 'externalized gut metagenome' is partly common to humans and benefits the gut microbial community. We conclude that testing relationships between microbial genes, microbes and their animal hosts, using network-based methods, could help to unravel additional mechanisms of transmission in holobionts.

Introduction

It has been proposed that an organism and its microbes form an assemblage called a holobiont (1, 2). The human body and human genome along with gut microbes and their genomes can be seen as a dynamic holobiont system (3–5), e.g. a superorganism amalgamating microbial and human attributes (6). In this multipartite holobiont, the host genome provides the primary genome, while microbial genes constitute the ‘second human genome’, which is in fact a prokaryotic pangenome (3, 7, 8). Whether the holobionts are units of selection is actively debated (9–11), yet, other aspects of their biology are less controversial, and holobionts are becoming a major object of study in biology. Among these uncontroversial features lies the observation that, by definition, a holobiont is home to several different modes of genetic transmission. In nuclear transmission, the genetic material is inherited from one individual (in parthenogenesis for example) or, most of the time, from two individuals whereas in organelle transmission (of mitochondria, for example), the material is mostly inherited from the mother, in animals (12). Both

types of transmissions result directly from the reproduction of the host. This stands in contrast with the transmission of the microbiota, that is, the acquisition (or loss) of microbes between host generations. In mammals, at birth, the microbiota is inherited from the mother, but this is not always the case for other animal groups, where it could also be inherited from the environment (12, 13). During the life of the individual, the microbiota may even evolve, depending on different factors, which are currently not well characterized (host constraints, diet, environment, transmission between different hosts…) (14). The transmission of microbiomes differs in turn from the transmission of microbiota, since it is no longer (or at least not only) microbes that are exchanged, acquired or lost, but genes themselves. These genes may be carried by microbes, but also by viruses, plasmids, or other classes of mobile genetic elements. For example, transmissions in the gut microbiome are in part due to horizontal gene transfer (HGT) (4, 15, 16) because of the high cell density in microorganisms, and mediated by viruses – especially temperate prophages – (17, 18), integrases, recombinases (19), and conjugative transposons (20). Finally, the transmission of microbes from the environment to the host has not been systematically taken into account (11). As with any transmission, microbial transmission can be transient or permanent (Figure 1).

These complex transmissions endow the holobiont with characteristics inherited either from a macro-organism, as in nuclear transmission, or from its microbes, as in microbiome transmission, and even from the coexistence of the host and its microbes (21, 22). The host can exert some control on the microbial species in its microbiota (e.g. by genetic regulation)(23), and it can indirectly influence the genetic content of these microbes, its distribution and transmission. For example, a

mutation in a gene MEFV (which encodes a protein involved in regulation of innate immunity) affects the gut microbiota composition at the taxa level (e.g. proportions of *Enterobacteriaceae*, *Acidaminococcaceae*, *Ruminococcus* and *Megasphaera* are affected by the human gene mutation) and the gut microbiota diversity (23). Reciprocally, microbes may play a role in the host ability to digest food (e.g. cellulose in aquisition of a plant-based diet), in host protection, and in host development (24). The gut microbiome encodes indispensable metabolisms for human life, contributes to human nutrition and affects the development of our immune system and protection against pathogens (20, 25)**.** Co-evolution however can be difficult to distinguish from mere co-existence of a host and its microbes (11), (26), (12), (3).

Studying already established transmissions may allow us to find associations between a host (maybe with specific characteristics) and its microbiota and its microbiome. This chapter first introduces sequence similarity networks, and illustrates, with an application on human gut microbiomes, how these networks can be integrated into the study of genetic transmission. Next, it introduces bipartite networks, and illustrates how they may be theoretically used to enhance analyses of microbes and of gene transmissions in holobionts, with a particular focus on the microbiomes of populations of lizards with different diets. Overall, network analyses enhance the focus on an important process of genetic transmission, i.e. gene externalisation between mobile elements and cellular chromosomes.

Introducing sequence-similarity networks

Network-based methods are useful to study the evolution of gene family (27), gene transfer, composite genes and genomes, evolutionary transitions and holobionts (26, 28–31). More specifically, networks can be used to search for rules of association between genes and their microbial, viral or animal hosts, and to identify putative cases of gene introgression. Gene introgression occurs when the genetic material of a particular evolutionary unit propagates into different host structures and is replicated within these host structures(26, 32). By contrast, tree-based methods are more routinely used to describe divergences from a last common ancestors. Stated simply, the tree representation of evolutionary processes does not show the same thing as a network representation. Indeed, the tree representation allows observation of the evolution from one individual to many individuals, while the network representation includes the evolution from several individuals to one individual. These methods do not show the same paths of transmission, therefore it can be argued that network-based methods are more adapted than trees to study the processes where organisms form a collective system of complex genetic transmission, including patterns of vertical inheritance when they exist, since the tree is included in the network (33).

A sequence-similarity network is a network representation of sequence similarities, where each node of the network represents a sequence, and two nodes are related by an edge if the sequences have a higher similarity than a predefined threshold of identity and cover (34). Such network is built using the BLASTP algorithm (in a BLASTp all-against-all run on the set of sequences). In this process, an amino acid sequence is considered similar to another (i.e. the two corresponding nodes are linked by an edge) if at least 80% of each sequence is included in a match

(mutual cover criterion) and if the sequence similarity over the covering region (percentage of identity criterion) is higher than a given percentage, e.g. 95% (34, 35). Because of this thresholding scheme, in a sequence similarity network some groups of nodes have no connection, even an indirect one. Such groups are called connected components and can be used to approximate gene families (figure 2 (i)) (27, 36).

Building networks at different thresholds allows one to see differences in transmissions (1, 2). Although, the network is filtered for 80% mutual cover, several networks can be produced at different thresholds of percentage of identity: 70%, 75%, 80%, 85%, 90% and 95%. If one considers, in a first approximation, that sequences evolve by point mutation (rather than by recombination), and that these mutations accumulate linearly over time, then sequences that have been diverging for a longer time period, should contain more mutations than sequences derived from a recent last common ancestor. While such a schematic molecular clock cannot be realistically assumed, since the mutation rates of different gene families can be very different, and so similar percentages of identity can be found for families with very different ages, the use of % identity thresholds can nonetheless serve as a proxy for a relative dating of transmissions (37). Then, it provides a lower bound, since this procedure will admittedly underestimate the age of sequences that were affected by recent recombination events leading to gene conversion (i.e. since these events changed different, diverged sequences into identical ones).

APPLICATION OF SEQUENCE SIMILARITY NETWORKS ON HUMAN MICROBIOME DATA. Even if there is a growing interest in the mobilome of the

mammalian and human gastrointestinal tract (17, 38, 39), little is known about the population of Mobile Genetic Elements (MGE) residing in the human gut. In other words, investigations of the 'human + gut microbiota' holobiont (3, 15, 40) could turn into studies of a tripartite 'mobile elements + gut microbes + human individual' system, taking into account 3 types of potential partners: gut mobile genetic elements, their direct cellular hosts (i.e. the microbial community within which these MGE circulate), and the human individuals hosting these microbes.

To study this tripartite system, a simple bioinformatics protocol can be used to organize the predicted ORFs from 31 human gut microbiomes (18 from American, 13 from Japanese individuals) into 21,525 clusters encompassing at least 4 sequences with significant similarity (see Figure 3). Then, genes from the gut microbiota of these humans can be classified into resident families (defining a 'resident metagenome'), and into externalized and highly externalized families (both defining the gut externalized metagenome). This latter terminology indicates that a gene family has undergone horizontal transfer at some point between different types of genomes, i.e. between cells and viruses, between cells and plasmids, or between cells, viruses and plasmids. This partition allows us to propose a testable hypothesis about genetic transmission: "genes externalized on more types of MGE are shared more broadly both in gut microbial communities and across their human hosts". The methodology underlying the construction and analyses of these networks is detailed below, in order to allow the reader to perform similar studies.

CLUSTERING MICROBIOME SEQUENCES INTO GROUPS WITH DIFFERENT DEGREES OF EXTERNALIZATION. Using nucleotide sequences from

168 (20) and (41) as inputs for MetageneAnnotator (42) (default parameters), 311,265
169 ORFs were predicted from the gut microbiomes of 13 Japanese individuals and
170 195,521 ORFs were predicted from the gut microbiome of 18 US individuals, thus
171 yielding 506,786 similarly predicted ORFs. Unquestionably, assembly problems could
172 create chimeric sequences, which could potentially be misclassified. However, the
173 study below is based on assembled and unassembled reads from two independent
174 datasets, which corroborate one another. These ORFs were compared with 748,688
175 sequences, corresponding to all the proteins from phages, plasmids and integrons
176 publically available at the time of the analysis from the NCBI and ACID (43), and with
177 all spacers from 52,267 sequences from human gut prokaryotes' CRISPRs obtained
178 from (44), because spacers in these CRISPRs are expected to correspond to
179 fragments of viruses infecting gut microbes.
180 All sequences described above were compared by an all-against-all BLASTP
181 with –e set to 1e-20 and other parameters set to default for the non-unique 506,786
182 predicted ORFs and 748,688 mobile sequences, and the 506,786 predicted ORFs
183 were compared to the 52,267 sequences from human gut prokaryotes' CRISPRs by
184 a BLASTn (parameter –e set to 1e5, and other parameters set to default) to assign a
185 label to the gut predicted ORFs identical to a spacer. ORFs present in multiple copies
186 were not removed from the dataset before performing the analysis. Indeed, some
187 genes might be abundant in the human gut microbiome, such as genes involved in
188 metabolism of carbohydrates, which might be frequently transferred and/or duplicated
189 (45), and this important information was therefore not lost. BLAST outputs were
190 converted to a sequence similarity network using EGN (35). In this network, each
191 protein sequence corresponded to one node, and two nodes were directly connected

when their best hit displayed >= 20% identity for a BLAST E-value =< 1e-20. This threshold is sufficiently low to identify divergent homologs, even though it is true that sequences with greater divergence could still be holomogs. However, below 20%, BLAST detection is closer to a grey zone and false positives cannot be excluded. These analyses, which took 4 days to be completed, were performed on a computer with 2 quadcore Intel Xeon E5430 CPUs running at 2.66 GHz.

This inclusive protocol produced 74,615 connected components, defining a first set of clusters of sequences. Each human gut predicted ORF was characterized by a label reflecting its origin: ‘virus’ (for phages or sequences with 100% identity to a spacer), ‘plasmid’, ‘integron’ (altogether corresponding to ‘MGE’) and ‘gut predicted ORF’. These labels were then used to classify sequences clusters based on their content. Clusters of sequences exclusively encompassing sequences from gut predicted ORFs were considered as a proxy of the ‘resident gut metagenome’, and their gut predicted ORFs were referred to as ‘resident’. We found 13,259 such ‘resident’ clusters. Clusters of sequences that included both sequences from predicted ORFs and from MGE were considered as a proxy of the ‘gut externalized metagenome’. These latter clusters were further distinguished into two groups: those with only one type of MGE (only virus, or only plasmid, or only integron), for which gut predicted ORFs were referred to as ‘potentially externalized’, and those with > 1 type of MGE, for which gut predicted ORFs were referred to as ‘potentially highly externalized’, because more than one type of vector could intervene in the transfer of their genes. We found 7,468 ‘potentially externalized’ clusters and 798 ‘potentially highly externalized’ clusters.

SEQUENCES WITH A HIGHER DEGREE OF EXTERNALIZATION DEGREE ENCODE BIASED FUNCTIONS. The sequence similarity network was then simplified, by removing nodes corresponding to MGE, further splitting 'resident', 'potentially externalized' and 'highly externalized' clusters into clusters exclusively comprised of gut predicted ORFs. We then focused on the 21,525 clusters of 'resident' ORFs and of ORFs from gut externalized metagenome that comprised >= 4 ORFs (i.e. the minimal number of sequences for significant dN/dS analyses). Each ORF in these 21,525 clusters was individually, taxonomically and functionally annotated, using RPS-BLAST against the COG database (27, 46) and MGRAST (47). A majority rule was used to assign a COG category to each cluster (i.e. the most frequent COG category associated with ORFs from the cluster determined a general functional assignation for that cluster). This categorization allows one to test whether clusters with 'resident', 'potentially externalized', and 'potentially highly externalized' genes were enriched in different functional COG categories (hypergeometric test, p-values threshold of 0.01, adjusted for multiple testing using a Bonferroni correction). The taxonomic diversity was computed based on MGRAST annotations at different ranks of the taxonomy, from phyla to genera, as the number of different taxa represented in the gut microbiota, using the Vegan package (48). The number of human hosts with sequences in each cluster was quantified for all clusters of sequences.

As a result, we found that resident, potentially externalized and highly externalized clusters encompass genes encoding significantly distinct functions (Hypergeometric test, $p < 0.01$, Figure 4). The functional profiling shows that a homology-based separation of genes into resident, potentially externalized and

239 potentially highly externalized clusters is compatible with former knowledge of the
240 functions of genes from the mobilome. Genes involved in translation, ribosomal
241 structure and biogenesis, and transcription, as well as genes with poorly predicted
242 functions were significantly over-represented in the resident clusters. Accordingly,
243 such informational genes are generally considered less transferred (49–51), while the
244 limited distribution of genes with poorly predicted functions fits well with their lack of
245 functional annotation. Most externalized genes, especially when adaptive, are
246 present in more than one genome, enhancing chances for externalized genes to
247 have been annotated. Consistently, the externalized clusters of the gut microbiome
248 were enriched in genes involved in various metabolic pathways and microbial
249 interactions. More precisely, potentially externalized clusters were significantly
250 enriched in defense mechanisms, energy production and conversion, metabolism
251 and transport of amino acids, carbohydrates, coenzymes and inorganic ions.
252 Potentially highly externalized clusters were significantly enriched in genes encoding
253 replication and repair, as well as nucleotide metabolism and transport. These
254 differences are in line with previous knowledge on the gut mobilome. For example,
255 HGT is thought to largely explain CAZyme convergence across gut bacterial
256 genomes (52) and in our dataset, genes diagnosed as potentially externalized that
257 encode enzymes involved in carbohydrate metabolism and transport, are indeed
258 overrepresented. Likewise, there is a documented selective pressure in the gut to
259 enrich the microbial community in genes involved in DNA repair since ingested food,
260 secondary bile acids and nitroso compounds synthetized by gut microbes increase
261 the amount of genotoxic substances in the intestine (20). Such genes are
262 overrepresented in the most externalized clusters. For example, well known "genetic

goods" (53) includes ABC-type multidrug transport systems found on contigs associated with TN1549-like conjugative transposons (20), and plasmid genes coding community functions, mitigating the toxic effects of bile acids, or promoting adherence to host epithelial cells (15).

MGE can also be seen as providing a second type of community service, for the potentially externalized and highly externalized clusters detected in our analyses. These clusters could contribute to the functional stability of the gut microbiota, because the presence of their sequences on MGE genomes generates functional redundancy (3). The risk of losing these functions when a particular bacterial lineage gets eliminated from the competitive gut environment or fails to survive phage attacks (16) is reduced both through the externalization of gene copies. This dynamic MGE-mediated genetic redundancy is a key feature of the ‘mobile elements + gut microbes + human individual’ system, as will be shown below.

THE GUT EXTERNALIZED METAGENOME ENCOMPASSES GENES BENEFITING THE GUT MICROBIAL COMMUNITY. Amongst the externalized clusters, genes that can be considered public genetic goods are recovered (53). Since they are externalized, they are likely to benefit a broader range of microbial hosts than strictly vertically inherited genes, and are also likely to favor survival of the gut community. Consistently, genes already reported as beneficial for gut microbes and over-represented in microbiomes featured within the most abundant of our potentially externalized and highly externalized clusters. For example, the most abundant clusters within the defense mechanism category encoded an ABC-type multidrug transport system, ATPase and permease components (COG1132) and an

ABC-type antimicrobial peptide transport system, permease component (COG0577), commonly enriched both in Japanese adult microbiomes (20) and in obese monozygotic twins (41), as well as a cation/multidrug efflux pump (COG0841), which is also enriched in these obese twins. Since host intestinal cells produce various cationic antimicrobial peptides, and many microorganisms also do so to compete with other microbes, the enrichment of antimicrobial peptide transporters and the multidrug efflux pump possibly plays a primary role in stable colonization of gut microbes in the adult intestine by conferring resistance to cationic antimicrobial peptides (20).

SEQUENCES WITH DIFFERENT EXTERNALIZATION DEGREES ARE UNDER SIMILAR SELECTIVE PRESSURE IN THE GUT. The ratio of non-synonymous over synonymous mutations (i.e. substitution in DNA leading to a change in amino acids) in clusters of gut microbial sequences can be computed. Since non-synonymous mutations can be deleterious, and tend to be eliminated by purifying selection, a dN/dS ratio <1 indicates that sequences are under this kind of selective pressure. However, dN/dS analyses are ideally performed at an even finer level of granularity than the clusters defined above, i.e. from clusters of sequences presenting even stronger similarities, so that all sequences can be aligned together over most of their length. Thus, we used the BLASTClust program (ftp://ftp.ncbi.nlm.nih.gov; minimum sequence similarity threshold>70%; bidirectional coverage L >90%, other parameters by default) to construct stringent clusters of very similar predicted ORFs, which were aligned with MUSCLE (54) (default parameters) when they comprised >= 4 ORFs. Phylogenetic trees were reconstructed from these

311 protein alignments with PhyML (55) (-d aa -m WAG -f e v e). Corresponding
312 nucleotides were aligned, based on these templates, with transAlign (56). Aligned
313 nucleotides and phylogenetic trees were used as input in PAML (57) for estimating
314 one dN/dS ratio for each stringent cluster by maximum likelihood. This basic model
315 was fitted by specifying model = 0, NSsites = 0, in the codeml control file.
316 This protocol allowed us to compare the dN/dS distributions for the 3 classes
317 of clusters. While resident, potentially externalized, and potentially highly externalized
318 clusters showed distinct functional profiles, their molecular sequences were in stark
319 contrast, under comparable selective pressures (Mann-Whitney Wilcoxon test, $p > 0.01$, Figure 5). Most clusters in the gut microbiome were under purifying selection.
321 Thus, just like resident sequences, potentially externalized and highly externalized
322 sequences did not appear to be pseudogenes, nor to represent inactivated
323 prophages which no longer contribute to the active gut externalized metagenome.
324 Instead, sequences from potentially externalized and highly externalized clusters
325 were likely exploited by their microbial hosts. This observation does not mean that in
326 general, externalized sequences cannot undergo pseudogenization; but that, in our
327 study the bioinformatic pipeline effectively discarded pseudogenes. That most such
328 clusters were under purifying selection does not mean that sequences within a
329 cluster would not show genetic divergence. When we estimated genetic variation for
330 each cluster, by computing the mean %ID between all its pairs of connected
331 sequences in the similarity network, we observed that resident clusters were
332 comprised of significantly more similar sequences than potentially externalized and
333 highly externalized clusters (Figure 3, Mann-Whitney Wilcoxon test, $p < 0.01$). Thus,
334 sequences from the gut-externalized metagenome were in general more divergent

than sequences from the resident microbiome but not less affected by purifying selection. This latter observation is consistent with the description of fast evolving phages in gut microbial communities (58, 59): these phages and their genes are under selection and diverge faster in this environment.

SEQUENCES WITH HIGHER EXTERNALIZATION DEGREE HAVE BROADER MICROBIAL AND HUMAN HOST RANGES. The distribution of ORFs across inferred microbial host phyla and microbial host genera strongly suggests that partitioning sequences from human gut microbiomes into resident, potentially externalized and highly externalized clusters effectively captured aspects of the differential mobility of these sequences. Each predicted ORF in a cluster was assigned to its best matching microbial phylum and genus, using RAST annotation server (http://rast.nmpdr.org/, default parameters). Thus, the taxonomical diversity within each cluster could be estimated as the number of distinct microbial host phyla or genera associated with sequences for this cluster. Remarkably, clusters with the highest externalization degree were also the ones with the significantly broadest taxonomical diversity. Clusters of sequences with similarity to more than one type of MGE were more ubiquitous in the gut community (Figure 6). Some potentially highly externalized clusters were even found in >= 4 phyla, a host range compatible with findings in (60), where, using the most significant BLAST alignment to determine the origin of phage-encoded bacterial genes, found that 97 % of these mobile genes were attributed to the 4 dominant phyla known in the gut (60). These results (Figure 7) are compatible with the literature**.** Recent works representing gene-sharing networks of the mobilome (61–63), e.g. what genomes of MGE share what genes or

gene fragments with what other MGE genomes, and even broader gene-sharing networks of both the mobilome and the cellular genomes (64), reported that some genetic material can eventually be shared between different types of MGE (i.e. between viruses and plasmids, between viruses and virophages, etc.). Thus, some sequences from the microbiome are possibly carried around by a greater number of types of MGE (see for example, (15)), and more successful in a broader diversity of genomic contexts than others.

Remarkably, just like for microbial hosts, sequences inferred to be more externalized in the gut community were also the ones with the broader human host range clusters (Mann-Whitney Wilcoxon test, $p < 0.01$, Figure 8). Resident clusters were typically found in a smaller number of individuals than potentially externalized clusters, while potentially highly externalized clusters were found in the largest amount of humans. 0.8 % of these potentially highly externalized clusters were even present in >= 28 out 31 individuals. 75 % of these clusters from the gut externalized metagenome was shared between >= 4 individuals. Thus, the shared microbiome encoded on the gut-externalized metagenome was larger than the shared resident microbiome at the level of human hosts, for this dataset.

THE GUT EXTERNALIZED METAGENOME IS PARTLY COMMON TO HUMANS. The more sparse distribution of resident genes in comparison to externalized genes across humans may seem counterintuitive. However, it is likely explained by two possibilities. First, the gut microbiome is obviously under sampled, which means that *bona fide* core genes, present in all prokaryotic lineages and in all human individuals, were missed since they were not sequenced. However,

importantly, this undersampling could have equally affected resident and externalized genes. The two broader hosts distributions for externalized genes than for resident genes is therefore consistent with the notion that gene externalization, producing numerous copies of the same genes in the gut microbial community, introduces genetic redundancy. If so, externalized genes are likely to be more prone to being sequenced than resident genes, and this may explain why they appear to be more broadly distributed. Importantly, all sorts of genes can be externalized: for example, while housekeeping genes are likely to be part of a core microbiome because prokaryotes host these genes in their genomes, it does not mean that housekeeping genes are not being externalized, and are exclusively resident. Typically, some recA or gyrB sequences were classified as highly externalized, and present in up to 19 and 22 human hosts respectively. The second possibility is that externalized genes could be expected to be prominent members of the core microbiome, since not only prokaryotes, but also mobile genetic elements, i.e. multiple types of genomes, host them and multiply these genes.

This second result is better interpreted in the light of studies focusing on mobile genetic elements, which reported contrasting observations about the sharing of the gut mobilome between holobionts. On the one hand, MGE compositions were shown to display high inter-individual variation (3). Remarkable interpersonal variations in gut viromes and their encoded functions were described (58), demonstrating that gut viromes were unique to individuals (65), even when individuals have similar bacterial community structures (17, 41). But, on the other hand, prophages were also reported to be universal in the human gut, their genes amounting to 5% of the minimal gut metagenome (66), and some convergence of

viral populations was described for individuals following the same diet (59). Moreover, the persistence over more than 2.5 years of a small portion of the global virome within individual guts was also established (58), suggesting the possible existence of at least a very partial shared gut mobilome, should viruses with similar gene content be retained across multiple holobionts. In agreement with this conclusion, some gut plasmids and their ORFs were found in human gut microbiomes over large geographical distances (40). Likewise, Kurokawa *et al.* indicated that Tn1549-like conjugative transposons were enriched in most of the gut microbiomes present in their study, as well as in two fecal samples from American individuals (20), establishing a connection for this gene family between genetic mobility and its distribution in humans.

Network analyses support and amplify the notion that gut mobilome can be largely spread across humans. Could such externalized clusters further benefit, beyond the gut microbial community, to its human host (4, 15, 33)? Horizontal gene transfer from marine bacteria colonizing dietary seaweeds, into the genomes of gut bacteria were proposed to have introduced genes required for the use of seaweeds glycans, benefiting humans with a typical Japanese diet (45). Likewise, for antibiotic treated mice, direct evidence that phages contribute functional advantages to their gut microbial hosts (e.g. detected by an enrichment in genes related to the synthesis of cell wall constituents, or replication and repairrelated pathways) and possibly to their animal host (e.g. diagnosed by an enrichment in genes involved in cofactor and vitamins synthesis, starch, cellulose, lactose and fructans metabolism, and carbohydrate active enzymes) was also reported (60). Such studies must be considered with caution because samples can be biased, and one cannot exclude

the possibility of reagent contamination for the library prep and sequencing kits. While the indications of network analyses do not allow us to make such strong assumptions about the benefits of some externalized genes for the human host, they clearly suggest that the gut externalized metagenome has a broad host distribution across humans. The spread of gut externalized metagenome genes, within and across ‘mobile elements + gut microbes + human individual’ systems could contribute to explain why humans host a functional core microbiome (25), (41), (66). It stresses on the importance of considering multiple transmission channels to explain animal phenotypes.

Introducing bipartite graphs

Beyond the study of shared gene family with sequence-similarity networks (by considering in which genome a given sequence is present), another graph structure, namely bipartite graphs, can also be used to study genes and microbes transmissions in holobionts. A bipartite graph is a graph with two types of nodes (type I and II nodes) such that an edge only connects nodes of one type with nodes of the other type. For example, in host-gene family bipartite graphs, type I nodes are host organisms, and type II, gene families. These graphs can be used for instance to determine gene families that are shared by individuals (26), and possibly related to characteristics from the host (e.g. male/female or dietary groups) and to find gene families exclusive to each group. One advantage of this type of graph, is that, while they are equivalent to a presence-absence heatmap on the same data, it is straightforward to apply some concepts and algorithms of graph theory. It is easy to

rapidly identify which groups of genes are shared by which groups of genomes (67) (Figure 2 panel (ii)).

In theory, metagenomic data, such as the microbiome and microbiota of lizards with different diets, could be investigated using bipartite graphs and we will briefly indicate how. This example was chosen because such a dataset exist, and will soon be made available. In 1969, ecologists introduced 10 insectivorous lizards from the Adriatic island of Pod Kopiste (Croatia) to that of the neighbouring island of Pod Mrcaru. It was later realized that the lizards from the species *Podarcis sicula* on Pod Mrcaru had become omnivorous (80 % herbivorous) and even changed in morphology. However, changes of their gut microbiome and microbiota, and the transmissions of microbial genes and of microbes between lizards, were not investigated. Bipartite graph analyses would allow to consider the distributions of pairs of entities as diverse as genes, gene families, microbial taxa, and individual lizards, simply by introducing these objects either as type I or type II nodes, and thus to gain knowledge about the effect of transmission in such holobionts.

DETECTING THE TRANSMISSION OF GENES BETWEEN HOSTS. One of the striking features of microbiomes is that their gene content is functionally biased (20). The rules of transmission between hosts' microbiomes could be deciphered by considering “host-gene family” graphs, where type I nodes are hosts (i.e. lizards), and type II nodes are gene families. As explained above, a gene family can be defined as a connected component in a Sequence-Similarity Network (i.e. at >= 80% mutual cover, >= 30% identity, E-value <= 1e-5). These thresholds were empirically tested in multiple publications (31, 35, 68, 69). They recovered clusters of homologous

sequences, that are consistent in terms of COG annotation (27) and Pfam annotations, and therefore their use (in particular that of the >= 80% cover) providing a good proxy to define gene families, and to assign a family to each gene. Therefore, a “host-gene family” bipartite graph can be constructed as follows. An edge is drawn between a type I and a type II node if a member of the gene family represented by the type II node is present in the microbiome of the lizard represented by the type I node. Type I nodes may be further coloured by characteristics of the hosts (i.e. diet, gender,...). Gene families that are exclusively present in lizards displaying one characteristic (insectivorous or omnivorous), or shared by the two groups of lizards, or even the core genome, can then be detected (i.e. all the gene families shared by all lizards in the network (figure 2 (ii)). Importantly, because such an analysis would exploit metagenomic data, the quality of network analysis will depend on the quality of the predicted ORFs. Datasets with higher depth of coverage will in particular produce less false negatives in the predicted ORFs, and therefore will introduce less artificial nodes and relationships in the networks.

TRACKING THE TRANSMISSION OF MICROBES BETWEEN HOSTS.

Transmission can also take place at the level of the microbes themselves: in this case, the rules of transmission can be investigated by constructing a bipartite graph describing what microbes (type II nodes) are present in what microbiomes (type I nodes).

In this case, type I nodes are still hosts (i.e. lizards), but type II nodes are now the microbial assignation of the genes (either to a given OTU (70), or to a given prokaryotic taxa, i.e. a species, a genus or a phylum, or to a mobile genetic element

503 such as a virus or a plasmid – figure 2 (iii)). An edge links a type II node to a type I
504 node if the microbial assignation has been found in the microbiota of the lizard
505 corresponding to the type I node. These assignations can be achieved by various
506 strategies, such as by BLASTing reads (or predicted ORFs) against a reference
507 database containing both prokaryotes, viruses and plasmids (64), or by delineating
508 OTUs using QIIME (70).
509 The structure of the bipartite graph can then be exploited by decomposing the
510 network into connected components (or CC), *i.e.* all sets of nodes for which there is
511 always an interconnecting path. CC in bipartite graphs are informative partitions of
512 the data. They can be studied at different stringency levels, using different thresholds
513 of percentage of identity (with the same molecular clock interpretation than for
514 sequence-similarity networks). Bipartite graphs (and CC themselves) can be further
515 decomposed into twins, *i.e.* groups of nodes having exactly the same set of
516 neighbours in the graph. Applying these two notions to “genome-gene family”
517 bipartite graphs simultaneously define groups of gene families and groups of
518 genomes (referred to as the *support* of the CC or the twin). Computing connected
519 components splits the set of genomes into groups that have no genomic content in
520 common, and defines the corresponding disjointed pools of gene families. By
521 definition, gene families associated with disjoint groups of genomes have not been
522 transmitted between these genomes; or, if these genes have moved across the
523 distinct sets of genomes, they have been lost since. In other words, the definition of
524 CC in the “genome-gene family” network hints at the existence of barriers of genetic
525 transmission between genomes. Identifying twins splits the set of gene families into

groups that are present in exactly the same genomes, and thus, on the contrary, characterize groups of genomes that have an exclusive genomic content in common.

Both operations have been implemented specifically for genomic data (13, 26). A particularly interesting distinction when diverse annotations are additionally available (*e.g.* taxonomical, ecological, dietary...) can be done between CCs and twins having a *homogeneous* or *heterogeneous* support. In the theoretical application to the lizard dataset, homogeneous twins might be further subdivided between those whose support contains all lizards of one type, and those that contain only a subset of them. Accordingly, the partition in twins could show what microbial phyla are shared by all lizards or on the contrary, which microbial phyla are specific to lizards with a given diet. This may hint at preferential routes of microbes and genes transmissions between hosts with similar ecologies.

DETECTING GENE TRANSMISSIONS IN THE MICROBIAL COMMUNITIES. One step further in unravelling the processes governing the previous transmission mechanisms would be the study of the transmissions within microbial communities themselves, that is, of genes between microbes, and how these transmissions are made. In this setting, we are interested by the patterns of inheritance, either vertical or horizontal. The latter corresponds not only to the well-characterized notion of Lateral Gene Transfer (LGT) (30), but also to the process of 'gene externalization' described above.

To this aim, a sound approach would be to construct a “microbiota-gene family” bipartite graph, which describes the microbe (in the same broad sense of « microbial assignation at a given level» as before) that is associated to a particular

550 sequence (Bipartite graphs (ii)). As before, the definition of gene families requires us
551 to have built a sequence similarity network (Bipartite graphs (i)). Indeed, type I nodes
552 are here defined as microbial classes and type II nodes, as gene families, while an
553 edge connects a gene family to a microbial class whenever at least one gene in this
554 family is present in at least one microbe belonging to this microbial class.
555 Patterns of inheritance may be subsumed under broad classes that
556 characterize the availability of a given gene to be transmitted. Consistently with the
557 study of the human microbiome, three kinds of mobility class can be identified: a
558 gene may be externalized, if the gene family is shared by at least one bacterium and
559 one kind of mobile genetic elements. A gene may be highly externalized, when it is
560 shared by at least one bacterium and more than one kind of mobile genetic elements.
561 A gene may also be resident, if the gene family it belongs to is not shared at all by
562 mobile genetic elements. With this representation, it is easy to compute how many
563 phyla, genera or species of bacteria/archea share a given gene family (Figure 2 (iv)).
564 Moreover, the previous mobility information can be related with functional
565 annotation, since each gene family can be associated with a COG category, derived
566 from ORFs predicted on contigs, using the RPSblast tool (71–73). Thus, the bipartite
567 graph allows also to find what functions and COG categories are represented in
568 which microbial taxonomic levels (in the specific form under which they are present in
569 the lizards microbiome), and in particular their mobility pattern. COG categories (or
570 functions) which are highly externalized, externalized or resident can therefore be
571 detected, and correlations between gene externalization and gene functions can be
572 calculated. This is particularly important to get a more accurate understanding of the

possible biases in the transmissions of genes with different functions in the microbiome.

EXTENSION TO MULTIPARTITE GRAPHS. After having constructed and analyzed bipartite graphs, it may be worthwhile to extend even further the methodology to the consideration of multipartite graphs that connect more than two levels of information (Figure 9). Such a structure may help to unravel more intricate mechanisms of transmission. In the same illustrating example, a tripartite graph can be built, involving the three levels of hosts (lizards), microbiota, and gene families. In principle, this structure allows to detect patterns that are not accessible to the previously presented bipartite analysis: for instance some gene families are found to be exclusively shared by insectivorous lizards (gene families 1, 2, 3 and 4 in figure 9), but some of them are also shared by the same microbes (for example microbial classes 8 and 9 share gene families 1 and 2), while others are present in different microbes (for example gene families 3 and 4). In the implied “lizard-gene family” bipartite graph in figure 9, the group of gene families: 1, 2, 3, 4 will be considered as a whole (because these families are equally present in hosts 16, 17 and 18), while in the “microbiota-gene family”, gene families will be split into 3 groups: the first group contains gene families 1 and 2 (shared by microbial classes 8 and 9), the second group contains gene family 3 and the third, gene family 4. Only by considering the three levels (figure 9) is it possible to find differentially shared gene families. A possible conclusion might be that the first group of genes is shared only because the microbiota have been transmitted, while the second group of genes (i.e. genes present in different microbes yet exclusive to lizards with a specific diet) may indicate

that genes coding these functions are present in similarly constrained environments (the insectivorous lizards' gut), irrespective of what microbes carry these genes. Then, the transmission of these genes is decoupled from that of the microbes.

Conclusion

Holobionts are home to numerous biological transmissions. In this chapter, we highlighted the ways of transmission in microbial communities, in particular, the process by which mobile genetic elements contribute to the dissemination of genes in an environment, that we have called *gene externalization*. We propose that gene families present in a higher diversity of MGE have a broader bacterial and animal host range, and that gene externalization plays a key role in propagating similar material across hosts at two distinct levels of biological organization within this multilevel dynamic system. At the animal level, this gene externalization may therefore contribute to explain the sharing of a core functional microbiome. Under this hypothesis, this core is largely made of externalized genes necessary for microbes to survive in the gut. This claim encourages searching for some structure in the 'microbiome + animal' holobiont from a processual perspective (e.g. by characterizing the externalization degree for each shared or not shared gene family), and to reason within the conceptual framework of a 'mobile elements + gut microbes + human individual' system.

Accordingly, focusing on the externalized gut metagenome that is shared across humans could encourage future studies taking into account the implication of gut MGE for human health (4, 18). If gene externalization proves to be a mechanism involved in the constitution of a core microbiome across microbes and humans for

example, variations in introgressive processes and gene mobility are likely to play a role in deviations from the functional core microbiome. Such deviations have already been associated with different physiological states such as leanness or obesity (41), and differences in the abundance of bacteriophages and in pTRACA22, a plasmid specific of the human gut microbiome, have also already been reported between healthy individuals and those suffering from Crohn disease and inflammatory bowel disease (3, 4, 74). Highly externalized genes, considering their broad host range, both within microbes and humans, especially when involved in antibiotic drug resistance, might prove deleterious to humans (3, 18). These elements of the gut externalized metagenome deserve greater attention in models evaluating antibiotics uses (20, 40), because the number of treatment options for many clinical infections makes it critical to understand how the gut externalized metagenome spreads (5).

Moreover, future studies could evaluate whether the number of MGE carrying externalized genes, and not only the number of types of MGE on which genes are externalized, contributes to the constitution of the core microbiome. To this end, we have outlined the novel use of bipartite graph structures for the study of the genetic transmission in microbial communities. This approach gives a global and synthetic view of the transmissions: what are the externalized gene families and functions, what are the gene families exclusively associated with a particular lifestyle, and what are the microbes exclusively present in a set of microbiomes. Future studies will test the relevance of the network framework in transmissions studies.

Acknowledments:

644 E.C. and E.B. were funded by the European Research Council (FP7/2007-2013
645 Grant Agreement 615274) and CV by the LabexBCDIV**.** We thank Dr Paul Dean for
646 kindly proofreading this chapter.
647

648 **References**

649 1. Bosch TCG, McFall-Ngai MJ. 2011. Metaorganisms as the new frontier.
650 Zoology114(2011):185-190.
651 2. Bosch TCG. 2012. Understanding complex host-microbe interactions in Hydra.
652 Gut Microbes,3,345-51.
653 3. Jones B V. 2010. The human gut mobile metagenome, a metazoan
654 perspective. Gut Microbes 1,415-31.
655 4. Lepage P, Leclerc MC, Joossens M, Mondot S, Blottière HM, Raes J, Ehrlich
656 D, Doré J. 2013. A metagenomic insight into our gut's microbiome. Gut
657 62:146–58.
658 5. Broaders E, Gahan CGM, Marchesi JR. 2013. Mobile genetic elements of the
659 human gastrointestinal tract: Potential for spread of antibiotic resistance genes.
660 Gut Microbes,4:271-81.
661 6. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, Samuel BS, Gordon JI,
662 Relman D a, Fraser-Liggett CM, Nelson KE. 2006. Metagenomic analysis of
663 the human distal gut microbiome. Science (80-) 312:1355–1359.
664 7. Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F, Falony G, Almeida M,
665 Arumugam M, Batto JM, Kennedy S, Leonard P, Li J, Burgdorf K, Grarup N,
666 Jorgensen T, Brandslund I, Nielsen HB, Juncker AS, Bertalan M, Levenez F,
667 Pons N, Rasmussen S, Sunagawa S, Tap J, Tims S, Zoetendal EG, Brunak S,

Clement K, Dore J, Kleerebezem M, Kristiansen K, Renault P, Sicheritz-Ponten T, de Vos WM, Zucker JD, Raes J, Hansen T, Bork P, Wang J, Ehrlich SD, Pedersen O. 2013. Richness of human gut microbiome correlates with metabolic markers. Nature 500:541–546.

8. Relman DA. 2012. 3. Microbiology: Learning about who we are. Nature 486:194–195.

9. Guerrero R, Margulis L, Berlanga M. 2013. Symbiogenesis: The holobiont as a unit of evolution. Int Microbiol 16:133–143.

10. Lloyd EA. 2016. Holobionts as Units of Selection: Holobionts as Interactors, Reproducers, and Manifestors of Adaptation Elisabeth 1:1–38.

11. Moran NA, Sloan DB. 2015. The Hologenome Concept: Helpful or Hollow? PLoS Biol 13:e1002311.

12. Bordenstein SR, Theis KR. 2015. Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. PLoS Biol 13:e1002226.

13. Phillips ML. 2009. Gut Reaction: Environmental Effects on the Human Microbiota. Environ Health Perspect 117:A198–A205.

14. Spor A, Koren O, Ley R. 2011. Unravelling the effects of the environment and host genotype on the gut microbiome. Nat Rev Micro 9:279–290.

15. Ogilvie LA, Firouzmand S, Jones B V. 2012. Evolutionary, ecological and biotechnological perspectives on plasmids resident in the human gut mobile metagenome. Bioeng Bugs 3:13-31.

16. Xu J, Mahowald MA, Ley RE, Lozupone CA, Hamady M, Martens EC, Henrissat B, Coutinho PM, Minx P, Latreille P, Cordum H, Van Brunt A, Kim K, Fulton RS, Fulton LA, Clifton SW, Wilson RK, Knight RD, Gordon JI. 2007.

692 Evolution of Symbiotic Bacteria in the Distal Human Intestine. PLoS Biol
693 5:1574–1586.
694 17. Duerkop B a, Hooper L V. 2013. Resident viruses and their interactions with
695 the immune system. Nat Immunol 14:654–659.
696 18. Ogilvie LA, Bowler LD, Caplin J, Dedi C, Diston D, Cheek E, Taylor H, Ebdon
697 JE, Jones B V. 2013. Genome signature-based dissection of human gut
698 metagenomes to extract subliminal viral sequences. Nat Commun 4:2420.
699 19. Tasse L, Bercovici J, Pizzut-Serin S, Robe P, Tap J, Klopp C, Cantarel BL,
700 Coutinho PM, Henrissat B, Leclerc M, Doré J, Monsan P, Remaud-Simeon M,
701 Potocki-Veronese G. 2010. Functional metagenomics to mine the human gut
702 microbiome for dietary fiber catabolic enzymes. Genome Res 20:1605–1612.
703 20. Kurokawa K, Itoh T, Kuwahara T, Oshima K, Toh H, Toyoda A, Takami H,
704 Morita H, Sharma VK, Srivastava TP, Taylor TD, Noguchi H, Mori H, Ogura Y,
705 Ehrlich DS, Itoh K, Takagi T, Sakaki Y, Hayashi T, Hattori M. 2007.
706 Comparative metagenomics revealed commonly enriched gene sets in human
707 gut microbiomes. DNA Res 14:169–181.
708 21. Bordenstein SR, O'Hara FP, Werren JH. 2001. Wolbachia-induced
709 incompatibility precedes other hybrid incompatibilities in Nasonia. Nature
710 409:707–710.
711 22. Theis KR, Venkataraman A, Dycus J a, Koonter KD, Schmitt-Matzen EN,
712 Wagner AP, Holekamp KE, Schmidt TM. 2013. Symbiotic bacteria appear to
713 mediate hyena social odors. Proc Natl Acad Sci U S A 110:19832–7.
714 23. Khachatryan ZA, Ktsoyan ZA, Manukyan GP, Kelly D, Ghazaryan KA, Aminov
715 RI. 2008. Predominant role of host genetics in controlling the composition of

30

716 gut microbiota. PLoS One 3:e3064.
717 24. Selosse MA, Bessis A, Pozo MJ. 2014. Microbial priming of plant and animal
718 immunity: Symbionts as developmental signals. Trends Microbiol. 22:607-613.
719 25. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012.
720 Diversity, stability and resilience of the human gut microbiota. Nature 489:220–
721 230.
722 26. Corel E, Lopez P, Méheust R, Bapteste E. 2016. Network-Thinking: Graphs to
723 Analyze Microbial Complexity and Evolution. Trends Microbiol. 24(3):224-237.
724 27. Tatusov RL, Koonin E V, Lipman DJ. 1997. A genomic perspective on protein
725 families. Science (80-) 278:631–637.
726 28. Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative
727 impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U
728 S A 105:10039–44.
729 29. Kloesges T, Popa O, Martin W, Dagan T. 2011. Networks of gene sharing
730 among 329 proteobacterial genomes reveal differences in lateral gene transfer
731 frequency at different phylogenetic depths. Mol Biol Evol 28:1057–1074.
732 30. Skippington E, Ragan MA. 2011. Lateral genetic transfer and the construction
733 of genetic exchange communities. FEMS Microbiol Rev. 35:707-735.
734 31. Cheng S, Karkar S, Bapteste E, Yee N, Falkowski P, Bhattacharya D. 2014.
735 Sequence similarity network reveals the imprints of major diversification events
736 in the evolution of microbial life. Front Ecol Evol 2:72.
737 32. Bapteste E, Lopez P, Bouchard F, Baquero F, McInerney JO, Burian RM.
738 2012. Evolutionary analyses of non-genealogical bonds produced by
739 introgressive descent. Pnas 109:18266–18272.

31

740 33. Zhang S-B, Zhou S-Y, He J-G, Lai J-H. 2011. Phylogeny inference based on
741 spectral graph clustering. J Comput Biol 18:627–637.
742 34. Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, Lopez P,
743 Stoeck T, Bapteste E. 2015. Testing ecological theories with sequence
744 similarity networks: marine ciliates exhibit similar geographic dispersal patterns
745 as multicellular organisms. BMC Biol 13:1–16.
746 35. Halary S, McInerney JO, Lopez P, Bapteste E. 2013. EGN: a wizard for
747 construction of gene and genome similarity networks. BMC Evol Biol 13:146.
748 36. Bittner L, Halary S, Payri C, Cruaud C, de Reviers B, Lopez P, Bapteste E.
749 2010. Some considerations for analyzing biodiversity using integrative
750 metagenomics and gene networks. Biol Direct 5:47.
751 37. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, Fourment
752 M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria.
753 Microb Genom, 2(11):e000094.
754 38. Breitbart M, Haynes M, Kelley S, Angly F, Edwards RA, Felts B, Mahaffy JM,
755 Mueller J, Nulton J, Rayhawk S, Rodriguez-Brito B, Salamon P, Rohwer F.
756 2008. Viral diversity and dynamics in an infant gut. Res Microbiol 159:367–373.
757 39. Cadwell K. 2015. Expanding the role of the virome: commensalism in the gut. J
758 Virol 89:1951–3.
759 40. Jones B V, Sun F, Marchesi JR. 2010. Comparative metagenomic analysis of
760 plasmid encoded functions in the human gut microbiome. BMC Genomics
761 11:46.
762 41. Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE,
763 Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC,

764 Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins.
765 Nature 457:480–484.
766 42. Noguchi H, Taniguchi T, Itoh T. 2008. Meta gene annotator: Detecting species-
767 specific patterns of ribosomal binding site for precise gene prediction in
768 anonymous prokaryotic and phage genomes. DNA Res 15:387–396.
769 43. Joss MJ, Koenig JE, Labbate M, Polz MF, Gillings MR, Stokes HW, Doolittle
770 WF, Boucher Y. 2009. ACID: annotation of cassette and integron data. BMC
771 Bioinformatics 10:118.
772 44. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. 2012. CRISPR targeting reveals a
773 reservoir of common phages associated with the human gut microbiome.
774 Genome Res 22:1985–1994.
775 45. Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. 2010.
776 Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut
777 microbiota. Nature 464:908–912.
778 46. Galperin MY, Makarova KS, Wolf YI, Koonin E V. 2015. Expanded Microbial
779 genome coverage and improved protein family annotation in the COG
780 database. Nucleic Acids Res 43:D261–D269.
781 47. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, Paczian T,
782 Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards R, Venter J,
783 Remington K, Heidelberg J, Halpern A, Rusch D, Eisen J, Wu D, Paulsen I,
784 Nelson K, Nelson W, Tyson G, Chapman J, Hugenholtz P, Allen E, Ram R,
785 Richardson P, Solovyev V, Rubin E, Rokhsar D, Banfield J, Huse S, Huber J,
786 Morrison H, Sogin M, Welch D, Overbeek R, Begley T, Butler R, Choudhuri J,
787 Diaz N, Chuang H-Y, Cohoon M, Crécy-Lagard V de, Disz T, Edwards R,

McNeil L, Reich C, Aziz R, Bartels D, Cohoon M, Disz T, Edwards R, Gerdes S, Hwang K, Kubal M, Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M, Field D, Morrison N, Selengut J, Sterk P, Altschul Sf, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D, Jarvie T, DeSantis T, Hugenholtz P, Larsen N, Rojas M, Brodie E, Keller K, Huber T, Dalevil D, Hu P, Andersen G, Cole J, Chai B, Farris R, Wang Q, Wuyts J, Peer Yv de, Winkelmans T, Wachter R De, Leplae R, Hebrant A, Wodak S, Toussaint A, Meyer F, Overbeek R, Rodriquez A, Tringe S, Mering C von, Kobayashi A, Salamov A, Chen K, Chang H, Podar M, Short J, Mathur E, Detter J, Rodriguez-Brito B, Rohwer F, Edwards R, McHardy A, Martin H, Tsirigos A, Hugenholtz P, Rigoutsos I, Edwards R, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson D, Saar M, Alexander S, Alexander E, Rohwer F, Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards R, Felts B, Rayhawk S, Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F, Mou X, Edwards R, Hodson R, Moran M, Dinsdale E, Edwards R, Hall D, Angly F, Breitbart M, Brulc J, Furlan M, Desnues C, Haynes M, Li L, Krause L, Diaz N, Bartels D, Edwards R, Puhler A, Rohwer F, Meyer F, Stoye J, Liang F, Holt I, Pertea G, Karamycheva S, Salzberg S, Quackenbush J, Rohwer F. 2008. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9:386.

48. Dixon P. 2003. VEGAN, a package of R functions for community ecology. J

812 Veg Sci 14:927–930.
813 49. Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited:
814 Connectivity Rather Than function constitutes a barrier to horizontal gene
815 transfer. Mol Biol Evol 28:1481–1489.
816 50. Leigh JW, Schliep K, Lopez P, Bapteste E. 2011. Let them fall where they may:
817 Congruence analysis in massive phylogenetically messy data sets. Mol Biol
818 Evol 28:2773–2785.
819 51. Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes:
820 the complexity hypothesis. Proc Natl Acad Sci U S A 96:3801–3806.
821 52. Lozupone C. A., Hamady M, Cantarel BL, Coutinho PM, Henrissat B, Gordon
822 JI, Knight R. 2008. The convergence of carbohydrate active gene repertoires in
823 human gut microbes. Proc Natl Acad Sci U S A 105:15076–15081.
824 53. McInerney JO, Pisani D, Bapteste E, O'Connell MJ. 2011. The public goods
825 hypothesis for the evolution of life on Earth. Biol Direct 6:41.
826 54. Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy
827 and high throughput. Nucleic Acids Res 32:1792–1797.
828 55. Guindon S, Delsuc F, Dufayard JF, Gascuel O. 2009. Estimating maximum
829 likelihood phylogenies with PhyML. Methods Mol Biol 537:113–137.
830 56. Bininda-Emonds ORP. 2005. transAlign: using amino acids to facilitate the
831 multiple alignment of protein-coding DNA sequences. BMC Bioinformatics
832 6:156.
833 57. Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol
834 Evol 24:1586–1591.
835 58. Minot S, Bryson A. 2013. Rapid evolution of the human gut virome. Proc. Natl.

35

836 Acac. Sci. U.S.A. 110:12450–12455.
837 59. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman
838 FD. 2011. The human gut virome: Inter-individual variation and dynamic
839 response to diet. Genome Res 21:1616–1625.
840 60. Modi SR, Lee HH, Spina CS, Collins JJ. 2013. Antibiotic treatment expands the
841 resistance reservoir and ecological network of the phage metagenome. Nature
842 499:219–22.
843 61. Desnues C, La Scola B, Yutin N, Fournous G, Robert C, Azza S, Jardot P,
844 Monteil S, Campocasso A, Koonin E V, Raoult D. 2012. Provirophages and
845 transpovirons as the diverse mobilome of giant viruses. Proc Natl Acad Sci U S
846 A 109:18078–18083.
847 62. Yutin N, Raoult D, Koonin E V. 2013. Virophages, polintons, and transpovirons:
848 a complex evolutionary network of diverse selfish genetic elements with
849 different reproduction strategies. Virol J 10:158.
850 63. Iranzo J, Krupovic M, Koonin E V. 2016. The double-stranded DNA virosphere
851 as a modular hierarchical network of gene sharing. MBio 7(4):e00978-16.
852 64. Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E. 2010. Network analyses
853 structure genetic diversity in independent genetic worlds. Proc Natl Acad Sci
854 107:127–132.
855 65. Reyes A, Haynes M, Hanson N, Angly FE, Andrew C, Rohwer F, Gordon JI.
856 2010. Viruses in the fecal microbiota of monozygotic twins and their mothers.
857 Nature 466:334–338.
858 66. Qin J, Li R, Raes J, Arumugam M, Burgdorf S, Manichanh C, Nielsen T, Pons
859 N, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H,

860 Zheng H, Xie Y, Tap J, Lepage P, Bertalan M, Batto J, Hansen T, Paslier D Le,
861 Linneberg A, Nielsen HB, Pelletier E, Renault P, Zhou Y, Li Y, Zhang X, Li S,
862 Qin N, Yang H. 2010. A human gut microbial gene catalog established by
863 metagenomic sequencing. Nature 464:59–65.
864 67. Rivera CG, Vakil R, Bader JS. 2010. NeMo: Network Module identification in
865 Cytoscape. BMC Bioinformatics 11 Suppl 1:S61.
866 68. Haggerty LS, Jachiet PA, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ,
867 Pisani D, Wilkinson M, Bapteste E, McInerney JO. 2014. A pluralistic account
868 of homology: Adapting the models to the data. Mol Biol Evol. 31:501-16.
869 69. Méheust R, Zelzion E, Bhattacharya D, Lopez P, Bapteste E. 2016. Protein
870 networks identify novel symbiogenetic genes resulting from plastid
871 endosymbiosis. Proc Natl Acad Sci U S A 113:3579–84.
872 70. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello
873 EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST,
874 Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD,
875 Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J,
876 Yatsunenko T, Zaneveld J, Knight R. 2010. QIIME allows analysis of high-
877 throughput community sequencing data. Nat Methods 7:335–6.
878 71. Cock PJA, Chilton JM, Grüning B, Johnson JE, Soranzo N. 2015. NCBI
879 BLAST+ integrated into Galaxy. Gigascience 4:1.
880 72. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K,
881 Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics
882 10:421.
883 73. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-

Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Lu F, Marchler GH, Mullokandov M, Omelchenko M V, Robertson CL, Song JS, Thanki N, Yamashita RA, Zhang D, Zhang N, Zheng C, Bryant SH. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 39:D225.

74. Riley PA. 2004. Bacteriophages in autoimmune disease and other inflammatory conditions. Med Hypotheses 62:493-498.

Figure Legends

Figure 1. Different types of transmissions in holobionts. This figure presents different transmissions between holobionts, and between an holobiont and its environment. The holobiont is composed of two parts : the host, and its microbial communities (microbiome from gut, skin, oral …). When a host gives birth to another, parent bring to it offspring mitochondria, cytoplasm and genes (pink arrows). The transmission is from the parent to the progeny. In some cases, like in Mammals, the mother gives microbiota and microbiome to her children (pink dashed arrow). An host from an holobiont may also bring to the microbial community of another holobiont microbes, mobile genetic elements (viruses, plasmids), or genes, or metabolites (black arrow between holobionts).

907 Otherwise holobiont may exchange with the environment mobile genetic elements
908 (plasmids, viruses), microbes (blacks arrows between holobiont and environment)...
909 Then, in a microbial community there are transmissions between the different
910 elements : transmissions between microbes (for example by reproduction – purple
911 arrows- or Lateral Gene transfer – green arrow), transmission between microbes and
912 mobile genetic elements (red arrows). In this case we talk about externalization.
913
914 **Figure 2. Networks as tools for describing relationship between holobionts and**
915 **transmissions in lizards' gut microbiome.** (i) Sequence-Similarity Network (SSN).
916 The SSN is built by comparing ORFs from all lizards'gut microbiome, using an all-
917 against-all BLASTP. The SSN contains 5 connected components, also called Gene
918 families. Nodes are ORFs and they are colored depending on their taxonomic
919 annotation (see legend). If two nodes are similar at a determined percentage of
920 identity (e.g. 95%) then they are linked by an edges.
921 (ii) Bipartite graph Lizards-Gene families. Type I nodes are lizards, colored
922 depending on their diet and bottomtype II nodes are the 5 gene families described in
923 (i). There is an edge between a type I node and a type II node if in the gene family
924 you can find a sequence which is contained in the lizards'gut microbiome of the type I
925 node.
926 (iii) Bipartite graph Lizards-microbial classes. Type I nodes are lizards, and type II
927 nodes are microbial classes. If in one lizard's gut microbiome a microbial class is find,
928 then, there is a link between the type I node associated and the type II node
929 associated. Type I nodes are colored depending on the diet of the lizard, and type II
930 nodes are colored depending on the microbial class of ORFs.

(iv) Bipartite graph Microbial classes – Gene families. Type I nodes are microbial classes and type II nodes are the 5 gene families described in (i). There is an edge between a type I node and a type II node if at least one ORF of the gene family of the type II node is from the microbial class of the type I node.

Figure 3. Distribution of resident, potentially externalized and highly externalized clusters according to their average pairwise identity percentage. Functional distributions were plotted for the 3 classes of clusters; resident clusters are in black, potentially externalized clusters are in grey, and potentially highly externalized clusters in white.

Figure 4. Functional distributions of resident, potentially externalized and highly externalized clusters. Each cluster was assigned a COG annotation. (A) RNA processing and modification; (B) Chromatin structure and dynamics; (C) Energy production and conversion; (D) Cell cycle control and mitosis; (E) Amino acid metabolism and transport;(F) Nucleotide metabolism and transport; (G) Carbohydrate metabolism and transport; (H) Coenzyme metabolism; (I) Lipid metabolism; (J) Translation; (K) Transcription; (L) Replication and repair; (M) Cell wall/membrane/envelop biogenesis; (N) Cell motility; (O) Post-translational modification, protein turnover, chaperone functions; (P) Inorganic ion transport and metabolism; (Q) Secondary structure; (T) Signal transduction; (U) Intracellular trafficking and secretion; (Y) Nuclear structure; (Z) Cytoskeleton; (R) General Functional Prediction only; (S) Function Unknown. Functional distributions were plotted for the 3 classes of clusters; resident clusters are in black, potentially

externalized clusters are in grey, and potentially highly externalized clusters in white. For each class of clusters, significantly enriched functional categories ($p < 0.01$; Hypergeometric test, after adjusting for multiple testing) are identified by # (resident), + (potentially externalized) and * (potentially highly externalized).

Figure 5. Distributions of dN/dS ratio for resident, potentially externalized and highly externalized clusters. Resident, potentially externalized and highly externalized clusters are colored in black, grey and white, respectively. The 3 distributions are not significantly different (Mann-Whitney Wilcoxson test, $p < 0.01$). dN/dS ratios > 2 were pooled to simplify the display.

Figure 6. Distributions of the taxonomic diversity for resident, potentially externalized and highly externalized clusters. Left panel: distribution of clusters across microbial host genera, right: distribution of clusters across microbial host phyla. Clusters are color-coded as above. Potentially externalized clusters have a significantly broader host range than resident clusters, and potentially highly externalized have a significantly broader host range than resident clusters and potentially externalized clusters (Mann-Whitney Wilcoxson test, $p < 0.01$). Taxonomic diversity > 4 were pooled to simplify the display.

Figure 7. A schematic representation of gene sharings within and between three 'mobile elements + gut microbes + human individual' holobionts. Combinations of mobile genetic elements, and of their direct hosts, e.g. the gut microbial cells, as well as of the human body, itself hosting gut microbes form an

integrated, multilevel, multipartite, dynamic biological system, also referred to as a holobiont. For each individual human, the black square represents a close-up of its gut microbiota (e.g. with gut microbial cells in brown and mobile genetic elements in purple) and of its gut microbiome (e.g. the genes contained within these microbial cells and mobile genetic elements). Genes are represented by colored rectangles; genes with the same color belong to the same gene family. The process of horizontal gene transfer (HGT), mediated by mobile genetic elements, is responsible for mobilizing genes between microbial cells. We demonstrated that gene families carried by a larger diversity of mobile genetic elements are not only more widely shared between gut microbes, but also between individual human hosts, as shown for the red gene family. Thus HGT is a key process for introducing genetic similarity at multiple consecutive host levels within and between holobionts.

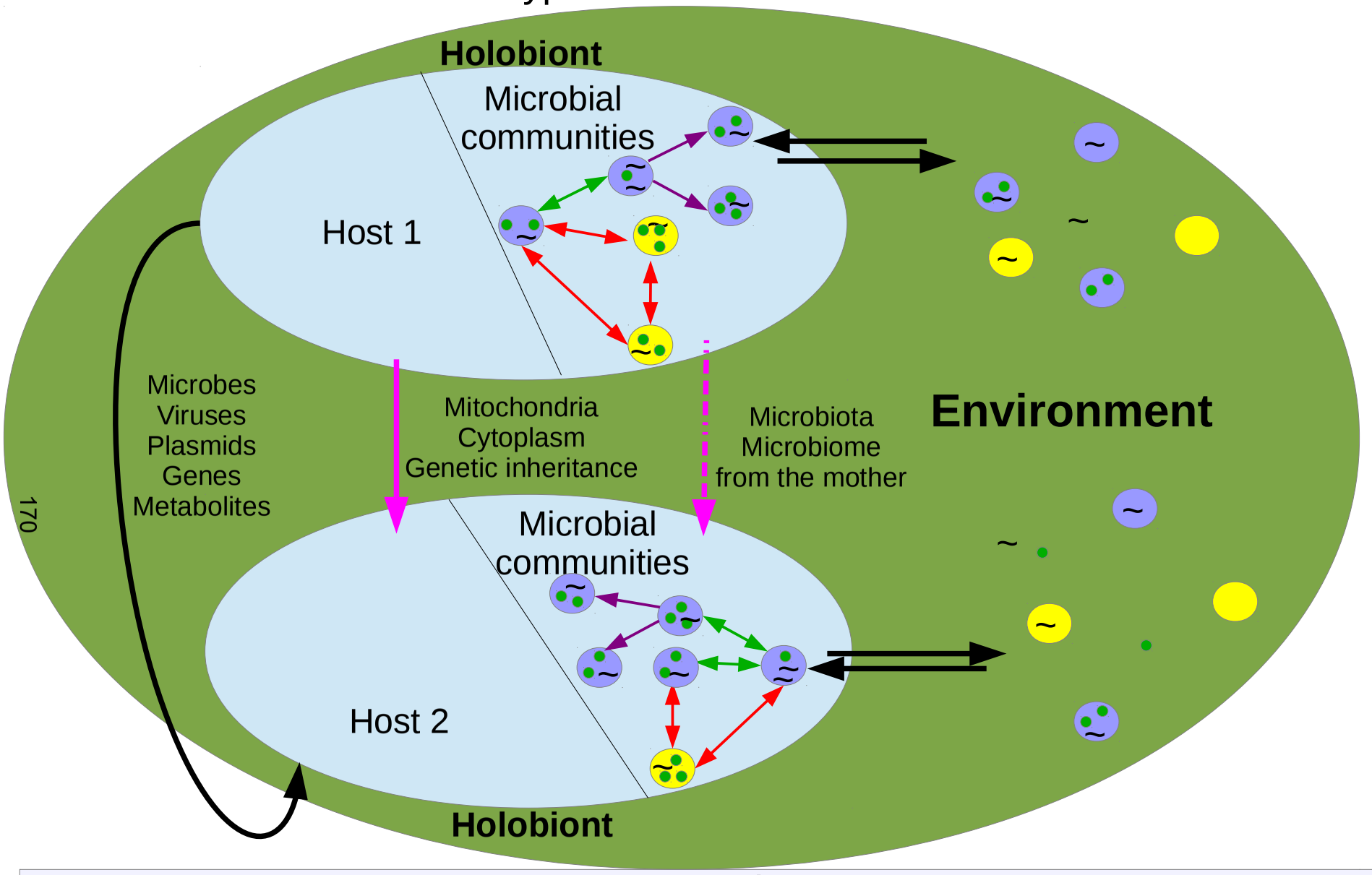
Figure 8. Distribution of resident, potentially externalized and highly externalized clusters across human hosts. Clusters are color-coded as above. Potentially externalized clusters have a significantly broader host range than resident clusters, and potentially highly externalized have a significantly broader host range than resident clusters and potentially externalized clusters (Mann-Whitney Wilcoxson test, $p < 0.01$).

Figure 9. Tripartite graphs allow to distinguish gene and microbial transmissions. In this tripartite graph type I nodes are the hosts (lizards) and are colored depending on their diet. Middle nodes are microbial classes, and type II nodes are gene families. Type II nodes colored in red are gene families shared by

insectivorous lizards only. Type II nodes colored in blue are gene families shared by omnivorous lizards only, and then, those colored in purple are shared by insectivorous and omnivorous lizards. The tripartitre graph allows to divide gene families specific from a diet in two categories : gene families which are shared by microbial classes specific from a diet (group 1 : gene families 1 and 2, encircled in red), and gene families which are shared by microbial classes non specific from a diet (group 2 : gene families 3 and 4, encircled in red). In the first group, gene families are not necessarily involved in the diet of the lizards, they are in insectivorous lizards because all the microbial classes which contain them are specific from insectivorous lizards. In the second group, gene families are more likely involved in the diet, because they are present in microbial classes which are not exclusive to insectivory. These two groups corresponds to two different ways of transmission, the first group is a microbial transmission whereas the second group is a gene transmission.

Figure 1

Different types of transmissions in holobionts

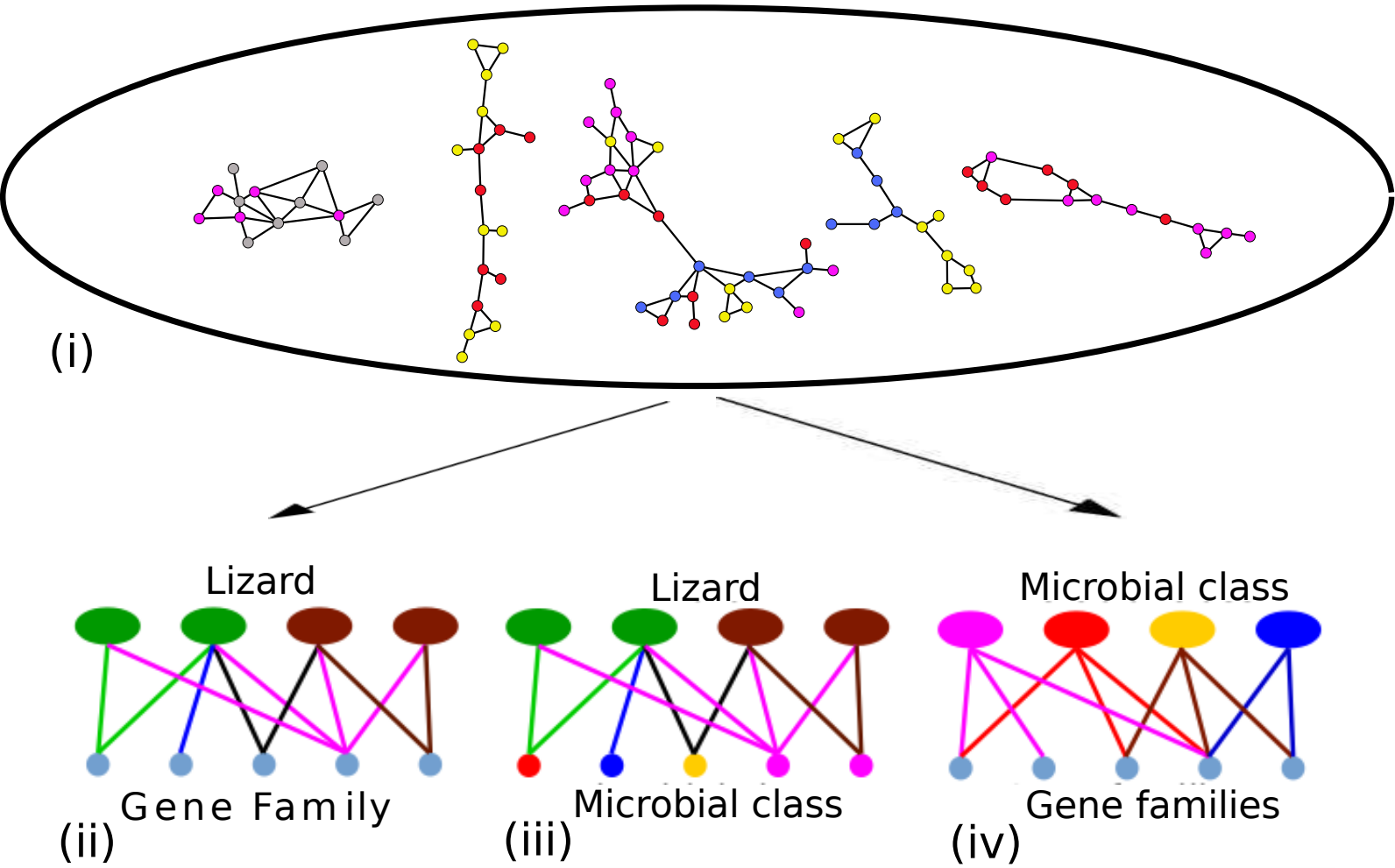


Legend :

- Plasmid
- Virus
- Prokaryote
- ~ Genes
- Transmission by parental inheritance
- Transmission independent of parental inheritance between hosts (from the same species or not)
- Optional transmission by parental inheritance
- Transmission by Lateral Gene Transfer
- Transmission by Lateral Gene Transfer by externalization
- Transmission by prokaryote division

Networks as tools for describing relationship between holobionts and transmissions in lizards' gut microbiome

Figure 2



- Color code for (i), (ii), (iii), (iv):
  - omnivorous lizards
  - insectivorous lizards
  - firmicutes
  - virus
  - plasmid
  - bacteroidetes

– (ii) gene families/(iii) microbes shared by:
  - some insectivorous and omnivorous lizards
  - omnivorous lizards
  - all lizards
  - insectivorous lizards
  - only 1 lizard

-(iv) gene families shared by :
  - firmicutes
  - plasmid
  - virus
  - bacteroidetes

Figure 3

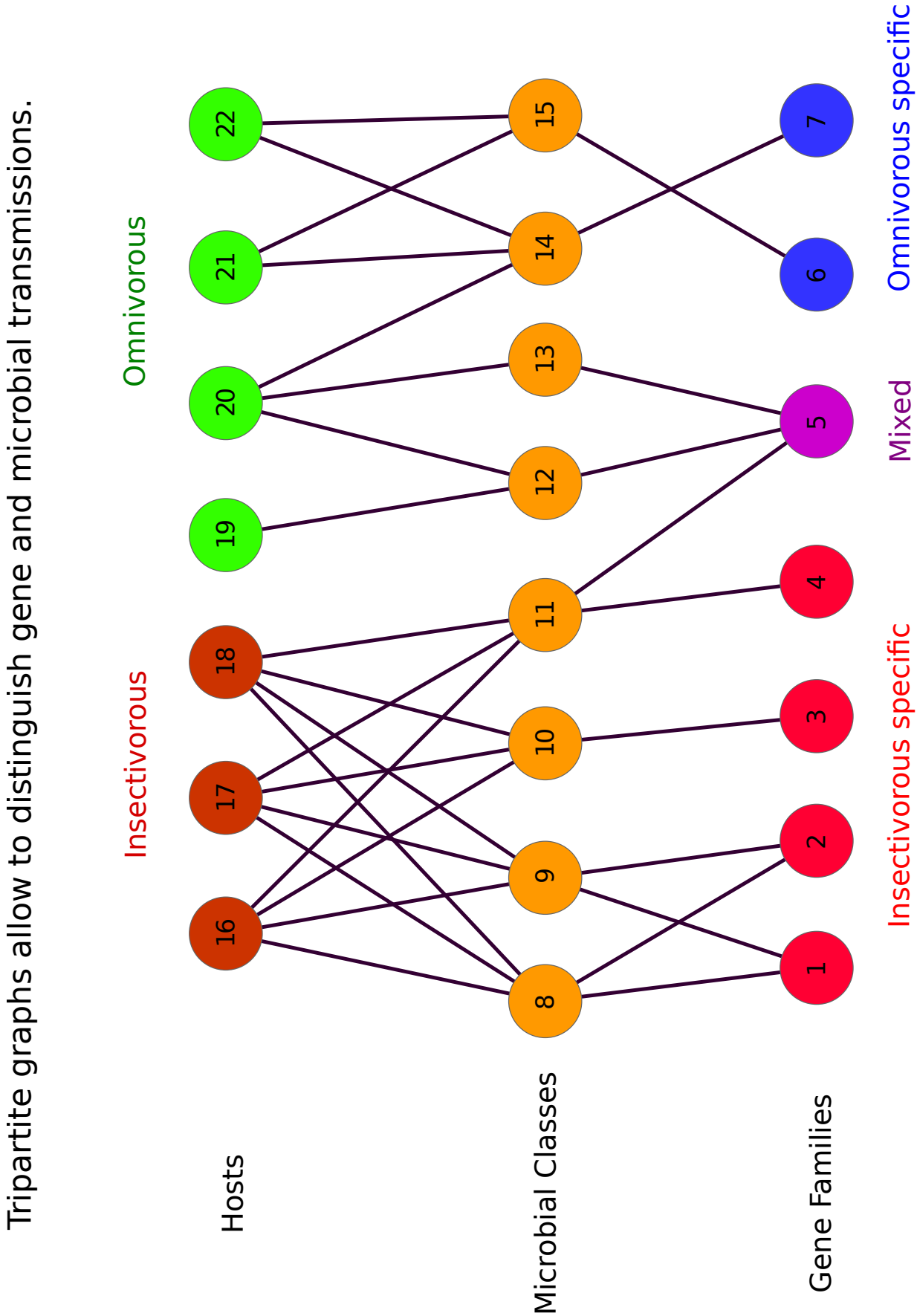

Tripartite graphs allow to distinguish gene and microbial transmissions.
Omnivorous
Insectivorous
Hosts
Microbial Classes
Gene Families
Omnivorous specific
Mixed
Insectivorous specific
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1

Figure 4

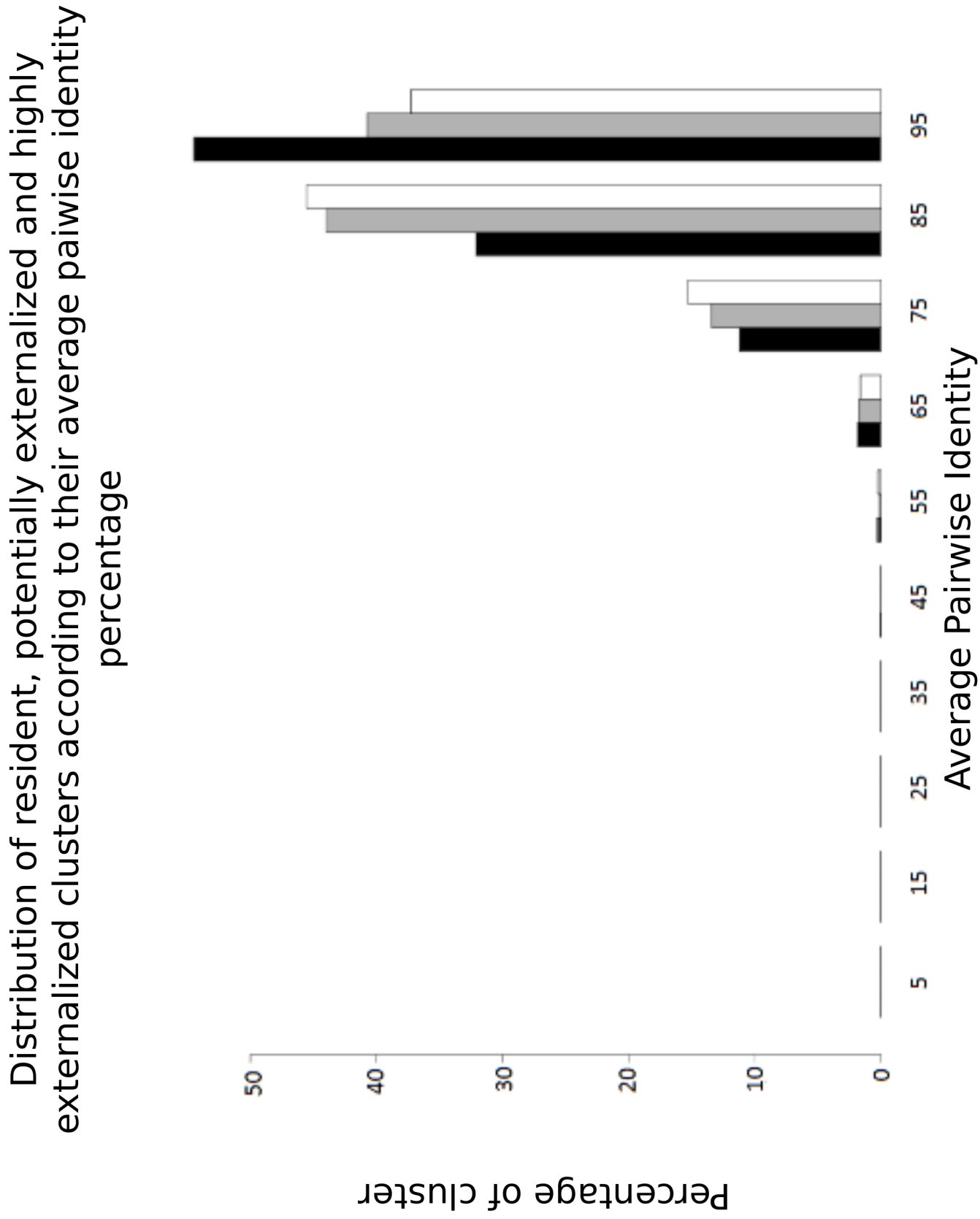

Distribution of resident, potentially externalized and highly externalized clusters according to their average paiwise identity percentage
Percentage of cluster
0
10
20
30
40
50
5
15
25
35
45
55
65
75
85
95
Average Pairwise Identity

Figure 5

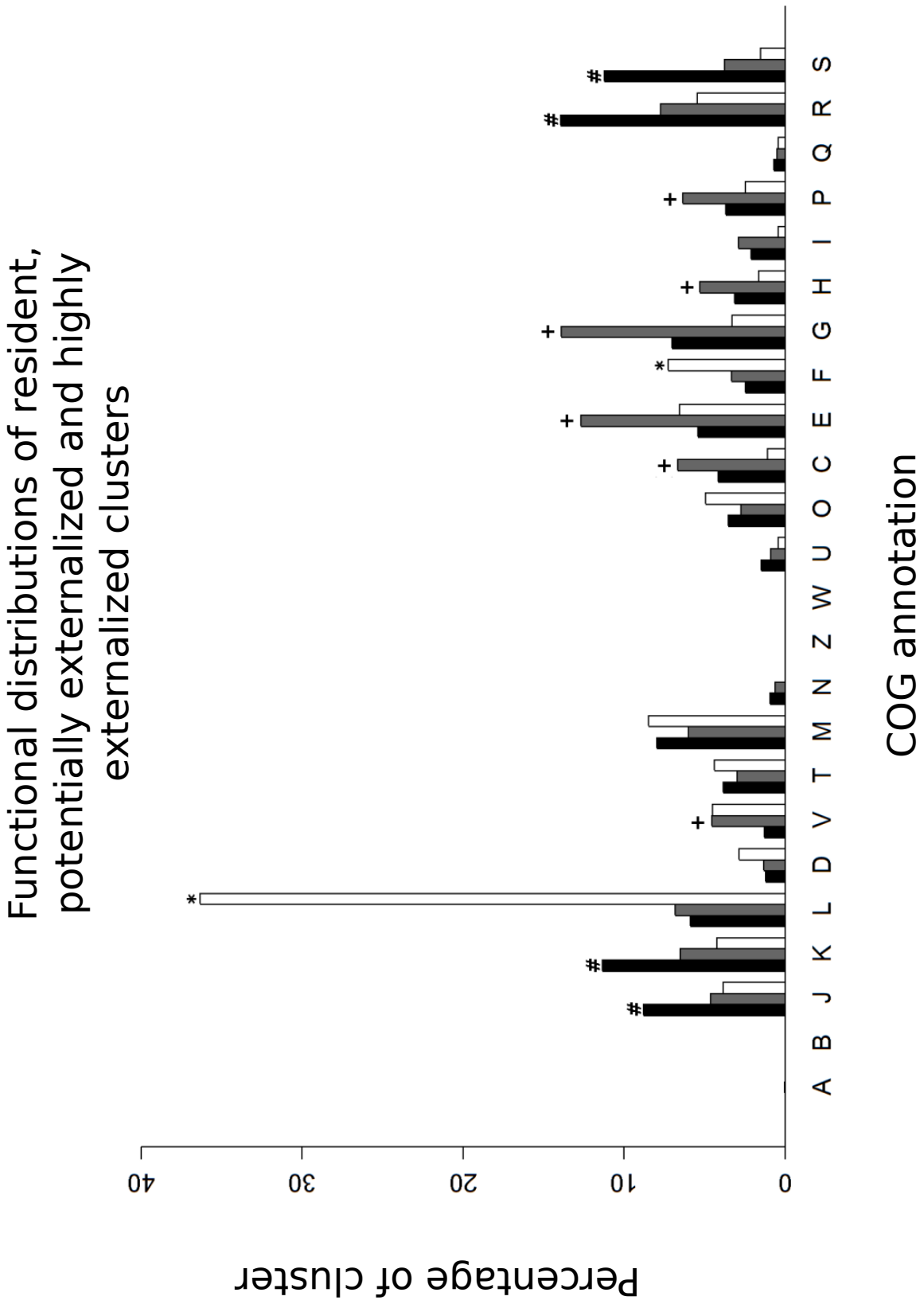

Functional distributions of resident, potentially externalized and highly externalized clusters
Percentage of cluster
0
10
20
30
40
A B J K L D V T M N Z W U O C E F G H I P Q R S
COG annotation

Figure 6

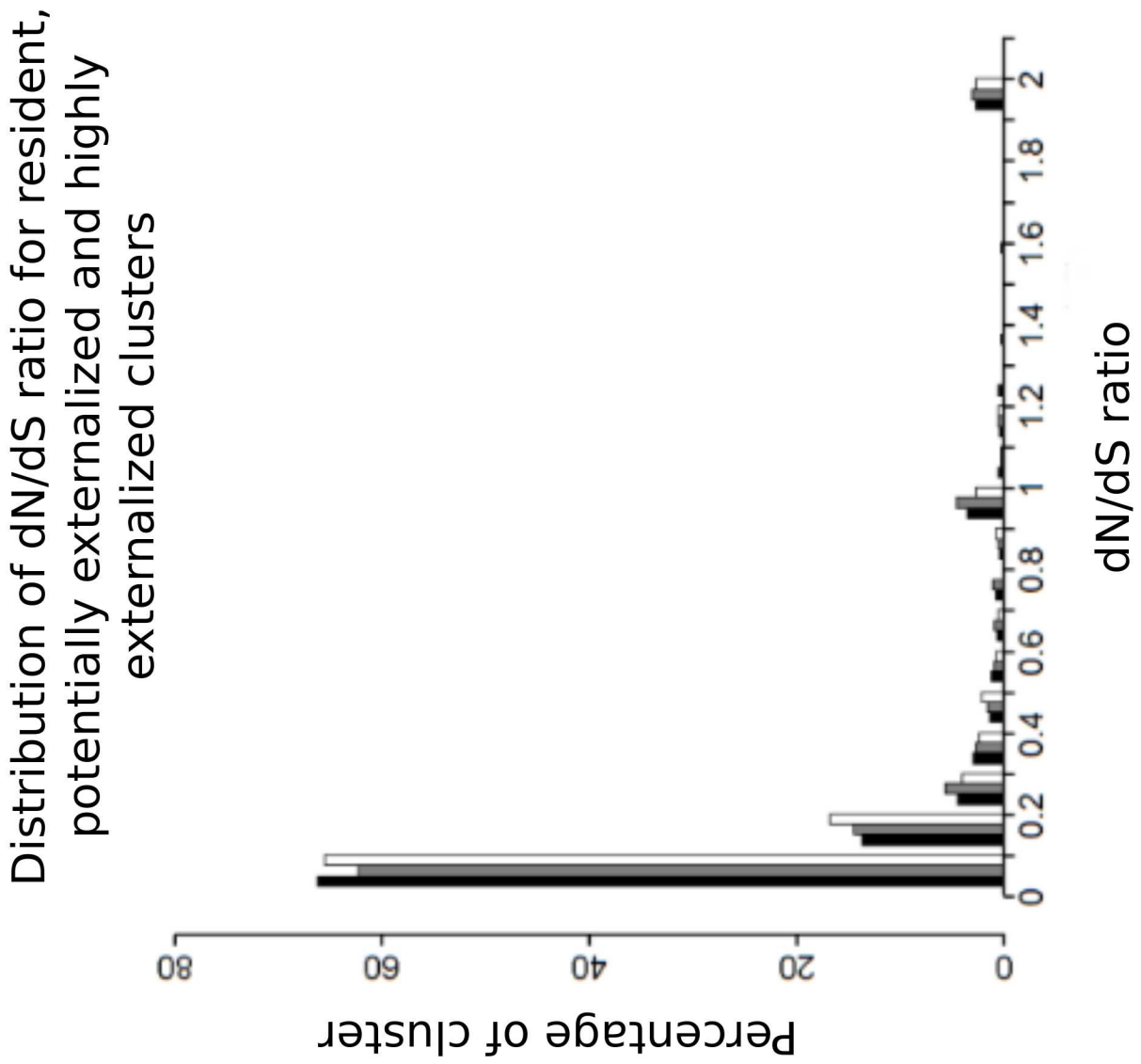

Distribution of dN/dS ratio for resident, potentially externalized and highly externalized clusters
Percentage of cluster
0
20
40
60
80
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
dN/dS ratio

Figure 7

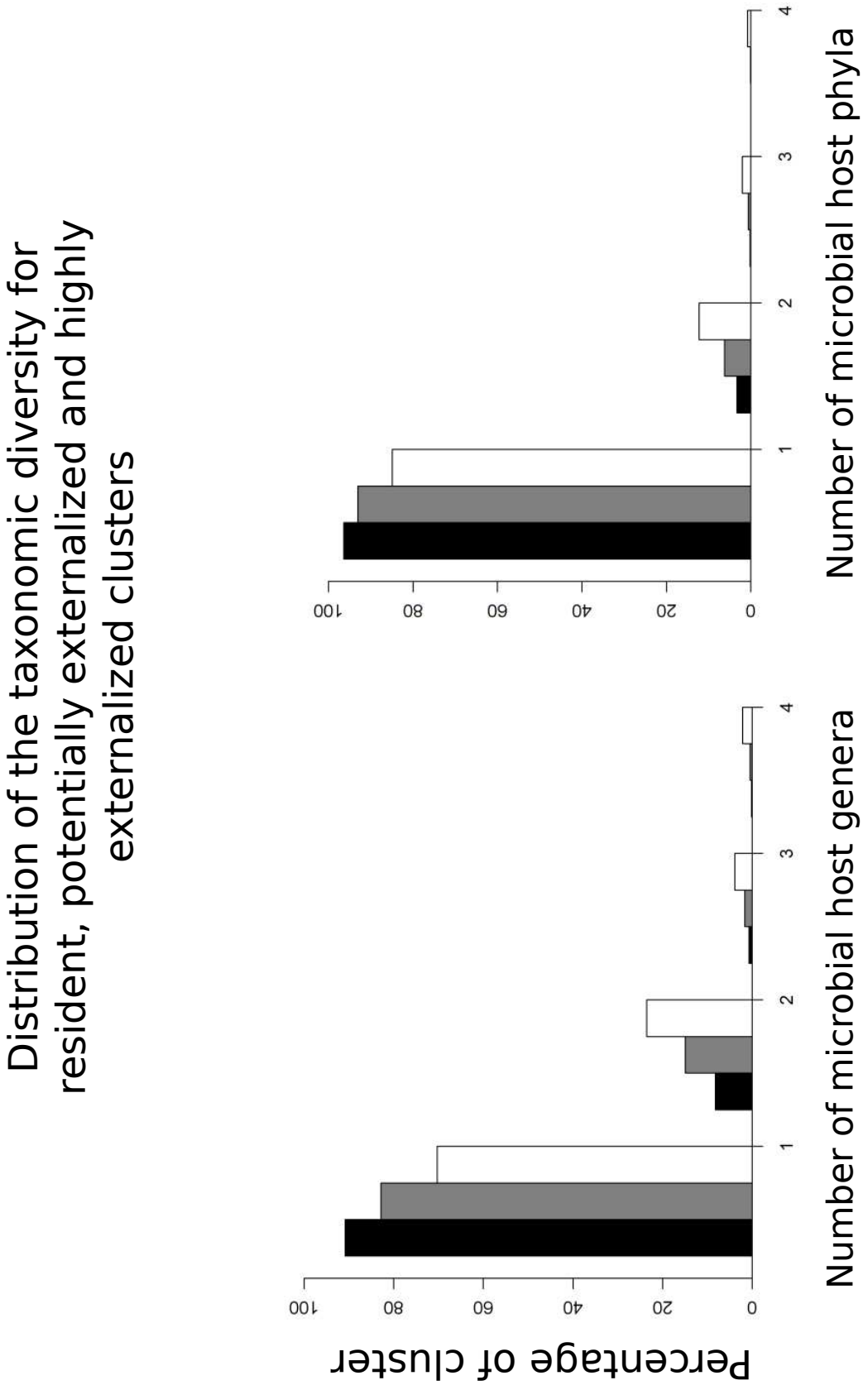

Distribution of the taxonomic diversity for resident, potentially externalized and highly externalized clusters
Percentage of cluster
0
20
40
60
80
100
1
2
3
4
Number of microbial host genera
Number of microbial host phyla

Figure 8

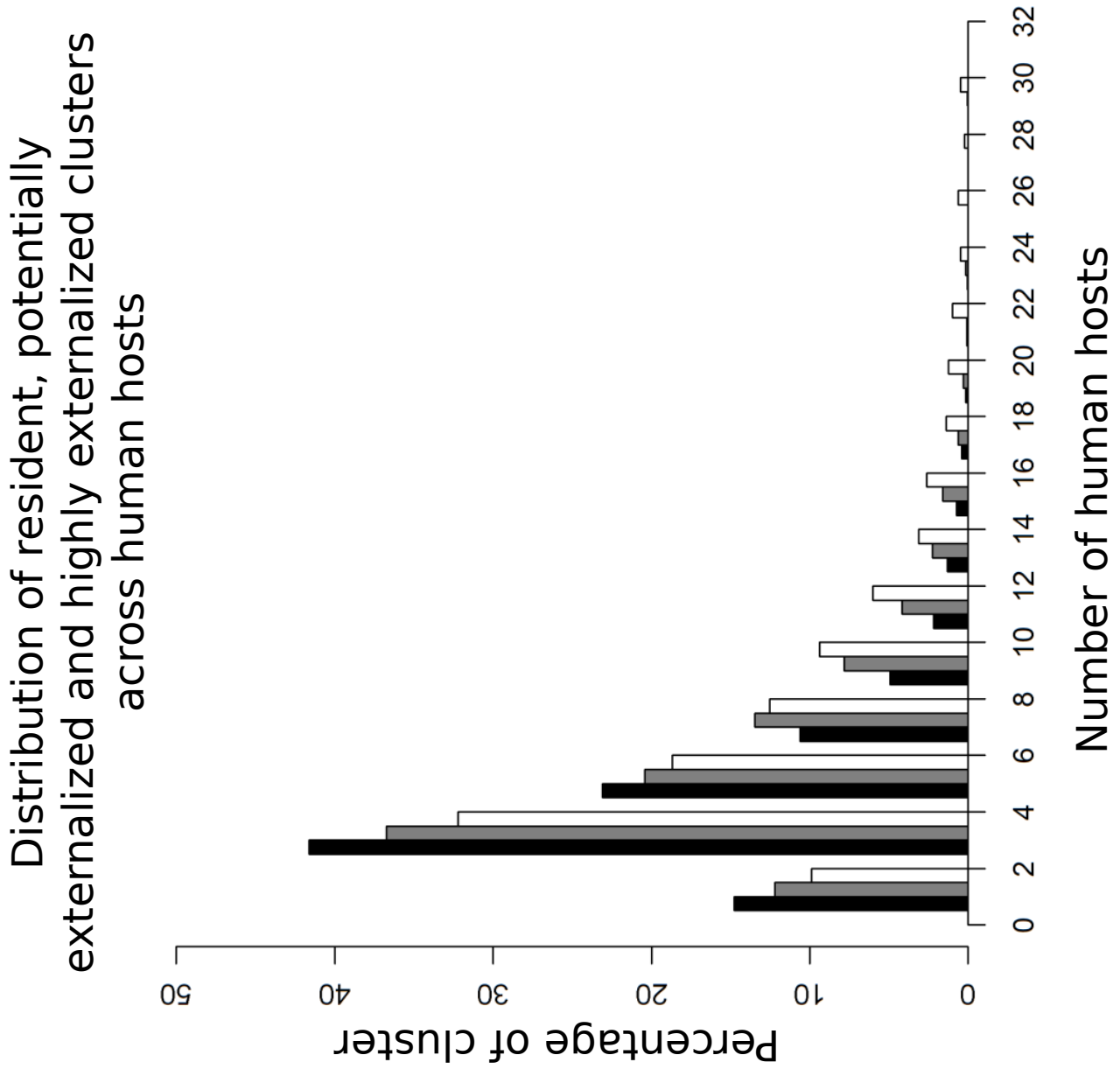
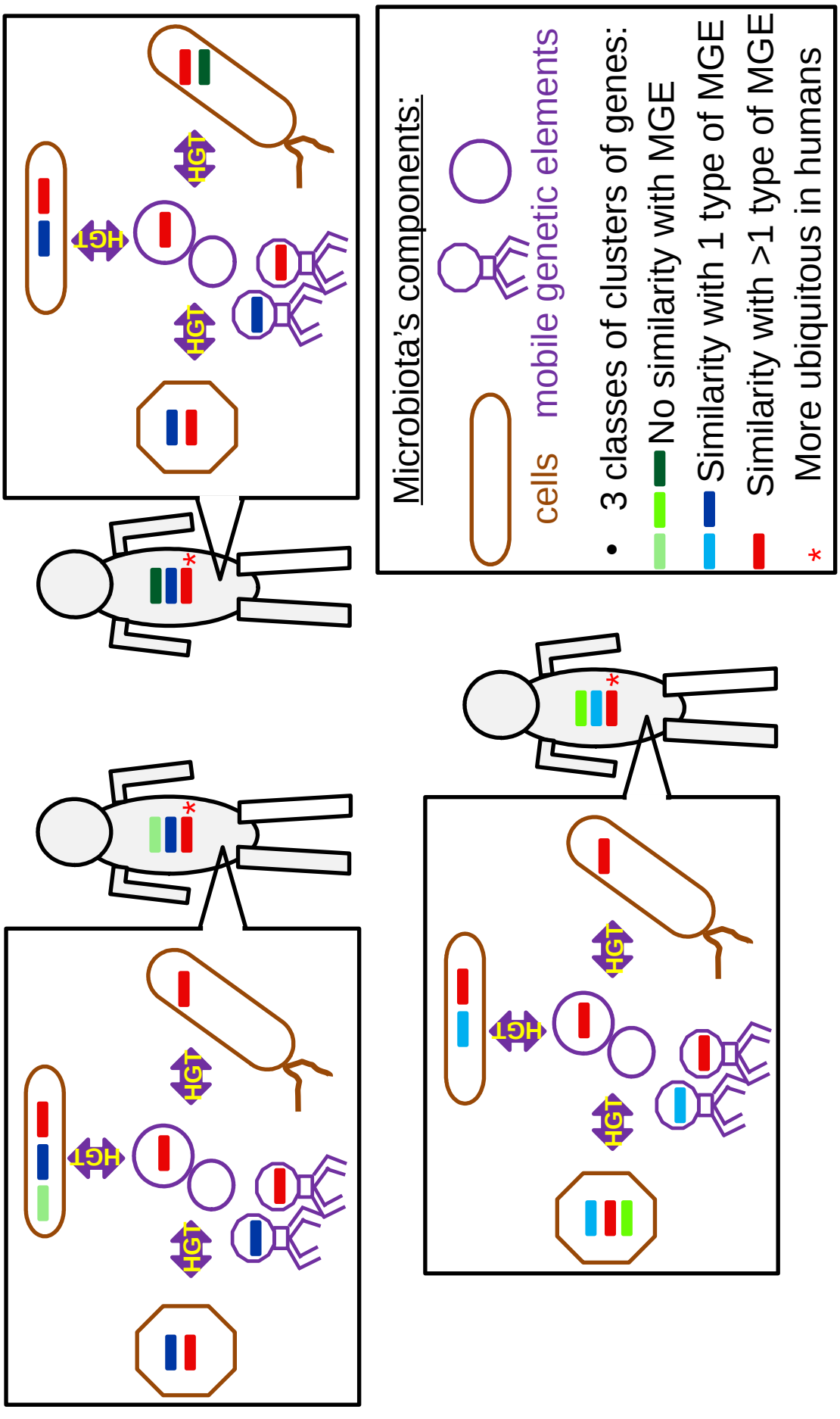

Distribution of resident, potentially externalized and highly externalized clusters across human hosts
Percentage of cluster
0
10
20
30
40
50
Number of human hosts
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32

Figure 9


Microbiota's components:
cells
mobile genetic elements
3 classes of clusters of genes:
No similarity with MGE
Similarity with 1 type of MGE
Similarity with >1 type of MGE
More ubiquitous in humans
*
HGT
HGT
HGT
HGT
HGT
HGT
*
*

On peut donc retenir de ce chapitre que les graphes bipartis sont un outil puissant pour étudier les règles d'introgression et de transmission au sein des métagénomes. Nous avons suggéré des analyses qu'il serait intéressant d'effectuer à partir des microbiomes de lézards. Une perspective de ce travail est désormais d'appliquer ces analyses théoriques à notre jeu de données.

Nous avons aussi défini le concept d'externalisation et classé les gènes du microbiome en trois groupes : des gènes résidents (i.e. ne se retrouvant pas dans des EGM), des gènes potentiellement externalisés (se trouvant dans un seul type d'EGM) et des gènes potentiellement très externalisés (se trouvant dans plusieurs types d'EGM).

5.1.4. Les réseaux de similarités de reads

Les réseaux de similarité de reads sont constitués de nœuds (les reads) et d'arêtes représentant la similarité entre deux reads. Deux nœuds sont reliés par une arête si les reads qu'ils représentent ont une couverture supérieure ou égale à 80%, une E-value inférieure à 10^{-5} et un pourcentage d'identité supérieur ou égal à 90% (Figure 29).

Couverture = **80%** de la longueur totale

Séquence 1 AATGAGTTCGCAATGGAGCA

Séquence 2 TTTGAGTTCGCAATGGAGGC

Identité = **100%** de l'alignement

Dans ce premier exemple, le pourcentage de couverture et le pourcentage d'identité permettent bien de considérer la séquence 1 et la séquence 2 comme similaires.

Couverture = **100%** de la longueur totale

Séquence 1 AATGAGTTCGCAATGGAGCA

Séquence 3 AAAGAGTTCGGAATGGAGCA

Identité = **90%** de l'alignement

De la même façon, les séquences 1 et 3 présentent un pourcentage de couverture et de similarité suffisants pour considérer que ces deux séquences sont similaires.

Couverture = **70%** de la longueur totale

Séquence 2 TTTGAGTTCGCAATGGAGGC

Séquence 3 AAAGAGTTCGGAATGGAGCA

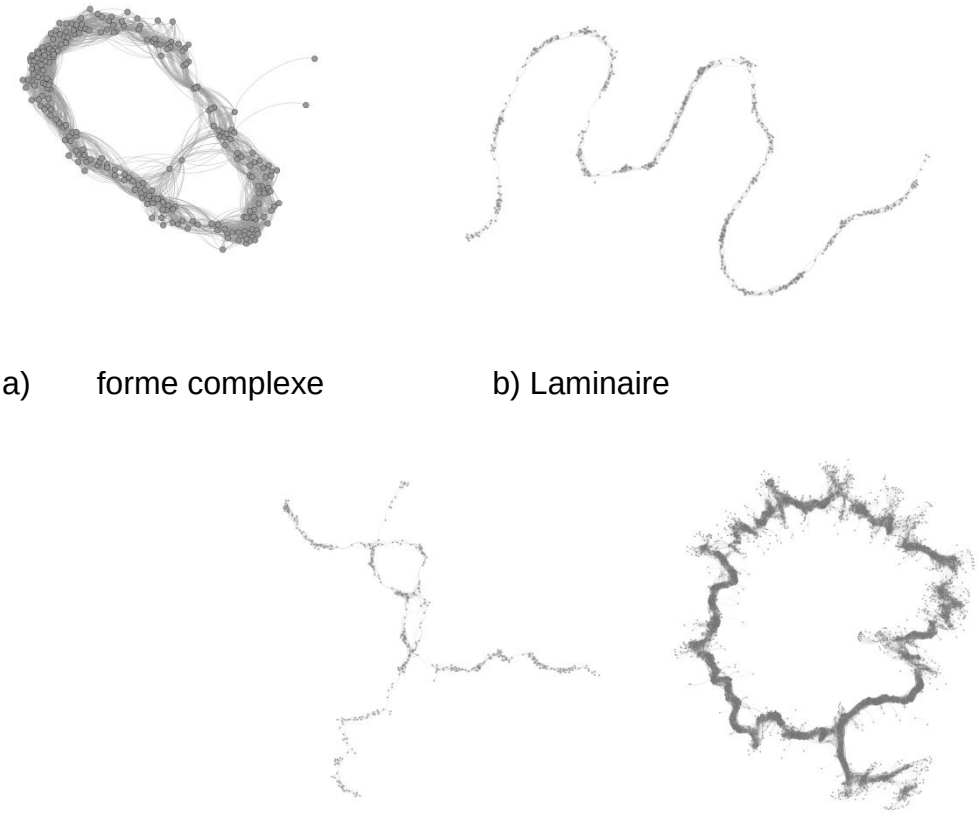
Identité = **93%** de l'alignement

En revanche, les séquences 2 et 3 ont un pourcentage de couverture trop faible (couverture de 70%) pour être considérées comme similaires. Les nœuds qui les représentent ne sont donc pas reliés entre eux. Le réseau de similarité complet contenant ces 3 reads serait alors le réseau suivant :

Figure 29 : Construction d'un réseau de reads.

Cette méthode permet d'analyser les différents contextes génomiques d'un métagénome comme nous le verrons par la suite. Cependant elle comporte un inconvénient : dans la mesure où le nombre de reads est bien supérieur au nombre d'ORFs, ce type de réseau est coûteux en temps de calcul et en espace de stockage.

L'assemblage des reads réduit nécessairement la diversité du jeu de données d'origine. En effet, l'assemblage produit des contigs (i.e. concaténation de reads en morceau de séquence linéaire plus long) alors qu'avec les réseaux de reads on obtient une grande variabilité de formes, dont certaines sont présentées ci-dessous (Figure 30) :



a) forme complexe b) Laminaire

c) forme complexe d) Boucle

Figure 30 : Quelques exemples de composantes connexes provenant de réseaux de similarités entre reads de microbiome intestinaux de Podarcis sicula.

Représentation graphique obtenue avec le logiciel Gephi (Bastian, Heymann, and Jacomy 2009)*, spatialisation : multi-niveaux de Yifan Hu* (Hu 2005)*. a) c) et d) sont des cas de composantes connexes présentant des boucles. b) est une composante connexe de type k-laminaire.*

Notre choix de spatialisation s'est porté sur l'algorithme multi-niveaux de Yifan Hu parce que c'est celui qui permettait de visualiser au mieux nos composantes connexes et parce qu'il est adapté aux réseaux de taille importante.

Nous détaillerons par la suite (en 5.3) nos interprétations des différentes formes de composantes connexes. Nous avons obtenu un réseau de similarité de reads pour chaque microbiome intestinal de lézard. Chacun de ces graphes contient plusieurs centaines de milliers de composantes connexes. Par exemple, le microbiome intestinal du lézard insectivore dont l'identifiant est PSK21MDI contient 544 838 composantes connexes.

5.2 Les k-laminaires

Puisque les reads sont des fragments d'ADN séquencés à des endroits aléatoires du métagénome, on s'attend à ce que le RSS reconstruit à partir de ces fragments ait une topologie particulière. En effet, plusieurs reads provenant d'un même contexte génomique peuvent se chevaucher et les nœuds correspondants du RSS vont alors former une sorte de chaîne plus ou moins épaisse comme représenté à gauche de la Figure 31.

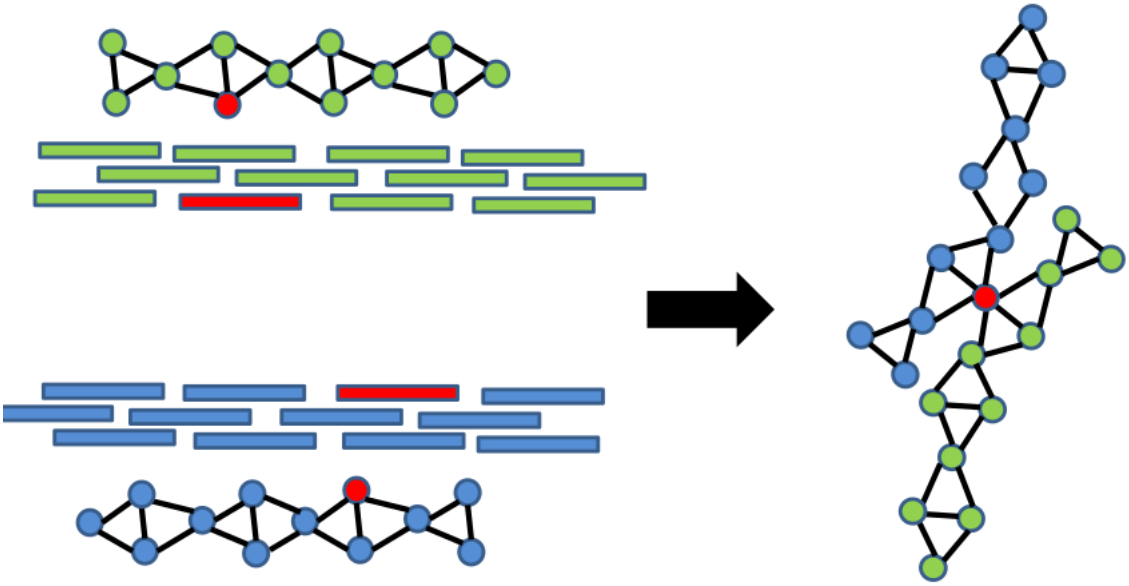


Figure 31 : définition des laminaires et des points de jonction.

A gauche, deux graphes laminaires (vert en haut et bleu en bas) sont représentés. Puisque les reads sont séquencés aléatoirement le long d'un ou plusieurs génomes, ceux qui sont présents dans le même contexte génomique et qui se chevauchent, sont reliés par des arêtes, ce qui finit par former les chaînes observées. Le read représenté en rouge sur les deux laminaires est un read qui appartient à ces deux contextes génomiques (celui du laminaire bleu et celui du laminaire vert). Dans ce cas de figure, les deux laminaires s'assemblent en une forme plus complexe, présentant un point de jonction (le nœud rouge). Dans ce cas de figure, le read rouge est donc une séquence présente dans deux contextes génomiques différents.

Ces graphes à la topologie particulière s'appellent des laminaires, de par leur ressemblance aux algues laminaires (Figure 32). Ce type de graphe a été défini et caractérisé par Michel Habib et Finn Volkel, suite à notre collaboration.

Figure 32 : Algue brune marine, la laminaire. Illustration provenant de l'encyclopédie Larousse.

Plus formellement, un k-laminaire (voir Figure 30 b)) est une composante connexe dont chaque nœud se situe à une distance inférieure ou égale à *k* arêtes du chemin diamétral, c'est-à-dire à *k* arêtes du plus long des plus courts chemins entre deux nœuds dans le graphe. Cette distance k au chemin diamétral est intéressante, car plus elle est grande, plus il y a de reads suffisamment semblables, mais non identiques, pour être reliés par une arête. Cette topologie suggère une diversité génétique dans la communauté microbienne puisque les mêmes gènes sont présents sous des formes variables. De ce fait, plus *k* est grand, plus il y a de variants rattachés à une même région d'ADN, et donc, plus la diversité génétique est importante. La valeur de *k* est une nouvelle manière de quantifier la diversité génétique dans un métagénome, même si dans nos jeux de données *k* est généralement faible (environ 3, M. Habib comm. pers.). Ceci est probablement lié à la faible couverture de séquençage de nos jeux de données. Pour une définition plus mathématique des k-laminaires, on se référera à l'article (Völkel et al. 2016).

5.3 Détection de laminaire et découpage des composantes connexes (article n°2)

Finn Völkel et Michel Habib ont développé un algorithme permettant non seulement la détection des laminaires mais aussi la décomposition des composantes connexes en sous graphes laminaires et non-laminaires. Ce programme détecte automatiquement les topologies « boucles », et « laminaire » présentées Figure 30 b) et 5.4 a) et d) dans les composantes connexes des réseaux, et en donne le nombre. Ce travail est exposé dans l'article suivant, intitulé « Read networks and k-laminar graphs » et déposé sur ArXiv.

Read networks and k-laminar graphs

Finn Völkel * Eric Bapteste † Michel Habib * ‡ Philippe Lopez†
Chloe Vigliotti†§

March 4, 2016

Abstract

In this paper we introduce k-laminar graphs a new class of graphs which extends the idea of Asteroidal triple free graphs. Indeed a graph is k-laminar if it admits a diametral path that is k-dominating. This bio-inspired class of graphs was motivated by a biological application dealing with sequence similarity networks of reads (called hereafter read networks for short). We briefly develop the context of the biological application in which this graph class appeared and then we consider the relationships of this new graph class among known graph classes and then we study its recognition problem. For the recognition of k-laminar graphs, we develop polynomial algorithms when k is fixed. For k=1, our algorithm improves a Deogun and Krastch's algorithm (1999). We finish by an NP-completeness result when k is unbounded.

Keywords: diameter, asteroidal triple, diametral path graphs, k-dominating paths, k-laminar graphs, (meta)genomic sequences, reads, read networks.

1 Introduction and biological motivation

Roughly speaking a k-laminar graph has a spine and all others vertices are closed to the spine (a more formal definition will be given in the next section). The definition of this graph class was motivated by its appearance in reads similarity networks of genomics or metagenomics data [4] see Figure 1. In sequence similarity networks, vertices are biological sequences (either DNA or protein sequences) and two vertices are adjacent if the corresponding sequences are similar, meaning that the pair shows a high enough BLAST score [2] and matches over more than 90% of the longest sequence. Here, sequences come from a metagenomic project and are usually called reads. Basically, reads are raw sequences that come off a sequencing

*IRIF, CNRS and Univ. Paris Diderot, France
†Team Adaptation, Integration, Reticulation, Evolution Lab. CNRS and IBPS, Univ. Pierre et Marie Curie, Paris, France
‡GANG Project, Inria Paris, France
§MECADEV, CNRS and Museum National d'Histoire Naturelle, Paris

machine, they are random DNA fragments, roughly 300 characters long, coming from the various microbial genomes that are found in a given environment. In our multidisciplinary approach [3] we wonder if two species of lizards can be distinguished by the analysis of read networks sequenced from the microbial DNA (microbiome) present in their gastro enteric tract. These networks are useful to biologists because, in addition to allowing the visualization of the genetic diversity that is found in the microbes of a given environment, they offer an alternative to more classical approaches, like the building of contigs[1], where one tries to rebuild the original genomic sequences of each organism out of the fragments, after a step of binning. The step of binning is a process which clusters contigs or reads, generally based on their composition, and tries to assign them to Operational Taxonomic Units (OTUs - which is the most commonly used microbial diversity unit)[16]. Sequence similarity networks are indeed a relaxation of de Bruijn graphs [7], which are commonly used to build contigs, since they are undirected and, more importantly, since two sequences are adjacent if they show a high enough, but not necessarily exact, similarity. In particular, they allow for the quantification of the genetic diversity of an ensemble of reads. For example of such networks, see Figure 1.

When a subset of reads covers a contiguous part of a genome (or parts of the genomes which have the same origin (common ancestor) also called homologous parts), they assemble into a k-laminar graph in sequence similarity networks, thus defining a singular genomic context (e.g. a suite of genes) on which biologists can measure the genetic diversity of the community. However, some genetic sequences, like repeats or transposases[2], can be found in more than one genomic context (i.e. when copies of the same transposase are inserted in multiple distinct locations of a genome), effectively linking together k-laminar graphs in sequence similarity networks. Building contigs out of sequences from such connected components is an especially difficult task.

To sum it up, sequence similarity networks of reads are thus composed of k-laminar parts, corresponding to singular regions of the genomes of a given environment, joined together by groups of vertices corresponding to repetitions in the genomes of a given environment. Identifying k-laminar parts in such networks, and eventually achieving a k-laminar decomposition, is thus of major interest to biologists.

2 k-laminar graphs

The graphs considered here are finite, loopless and undirected. For a connected graph G, with vertex set $V(G)$ and edge set $E(G)$, we denote by $d(x, y)$ for $x, y \in V(G)$ the distance between two vertices, i.e., the length of a shortest path joining x to y in G. We will use also the notion of eccentricity of a vertex $x \in V(G)$: $ecc(x) = max_{y \neq x, y \in V(G)} d(x, y)$ and so the diameter is $diam(G) = max_{x \in V(G)} ecc(x)$, similarly the radius is defined as $radius(G) = min_{x \in V(G)} ecc(x)$. Furthermore let us denote by $MaxEcc(G)$ the set of all

[1] a contig is a simple path in the approach based on the de Bruijn graph for assembling reads [15].

[2] transposase is a self-replicating enzyme that can insert itself in various part of genome, and is thus found in a variety of genomic contexts.

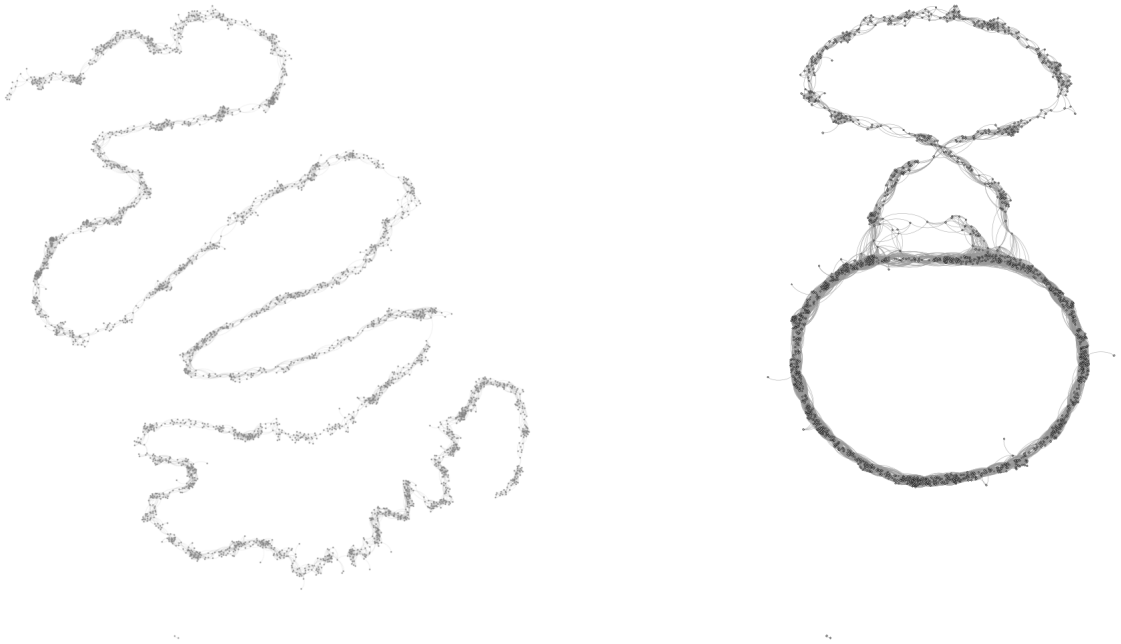


Figure 1: Two read graphs: the left one is a 4-laminar graph, the right one contains big cycles but can be decomposed into 4-laminars parts. Data used here comes from our project [3].

vertices of maximum eccentricity. When there is no ambiguity for a graph G we will denote by n, m respectively $|V(G)|, |E(G)|$.

We extend this notion to the distance of vertex to a path, namely $d(x, \mu)$, for some path μ, is the smallest distance from x to some vertex on μ. $N(x)$ will be the standard neighborhood of a vertex and we use also the notation $N[x] = N(x) \cup \{x\}$ for the closed neighborhood.

Similarly $N^k(x)$ the k-neighborhood, i.e. the vertices with distance equal to k from x or more formally $N^k(x) := \{y \mid d(x, y) = k\}$. We denote by $N^k[x]$ all vertices with distance less or equal to k, i.e. $N^k[x] := \{y \mid d(x, y) \leq k\}$ called the closed k-neighborhood. When G is connected, the maximum length of a path is called the diameter of G and denoted by $diam(G)$.

In this section we recall some standard definitions on graphs and introduce the notion of laminar graphs and the practical motivations of such a definition.

Definition 1. *An asteroidal triple (AT) is a triple of vertices such that each pair is joined by a path that avoids the neighborhood of the third.*

An AT-free graph is a graph that does not contain any AT. Intuitively if a graph does not contain any AT, then it cannot "expand" in more than 2 directions. The following definition of laminar graphs introduced here, is to generalize this intuitive notion of linearity.

Definition 2. *A path μ of a graph G is a diametral path if the length of μ is equal to $diam(G)$.*

Furthermore for every fixed integer k, a path μ in a graph G is called a k-dominating path if $\forall x \in V(G)$ we have $d(x, \mu) \leq k$.

Definition 3. *A graph G is called k-laminar (resp. strongly k-laminar) if G has a k-dominating diametral path (resp. if every diametral path is k-dominating).*

Proposition 4. *[8] AT-free graphs are 1-laminar.*

Proof. Corneil, Olariu and Stewart proved that AT-free graphs contain a dominating pair that achieves the diameter. Hence, AT-free graphs are 1-laminar. □

Definition 5. *A comparability graph is a graph G whose edge set $E(G)$ can be transitively oriented. A cocomparability graph is simply the complement of a comparability graph.*

It is well known that cocomparability graphs are AT-free [9]. Therefore also cocomparability graphs and interval graphs are 1-laminar.

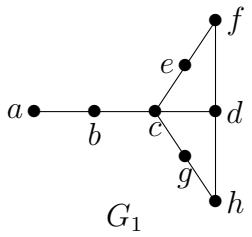


Figure 2: $\mu = [a, b, c, d, h]$ is a dominating diametral path of G_1, and (a, f, h) is an AT.

As shown by the graph G_1 of Figure 2, not all 1-laminar graphs are AT-free graphs. Thus AT-free graphs $\subsetneq$ 1-laminar. Furthermore AT-free are not always strongly 1-laminar as can be seen with the graph G_2 of Figure 3. To complete the picture, the class of AT-free graphs overlaps the class of 1-strongly laminar as can be seen with the graph G_3 of Figure 4.

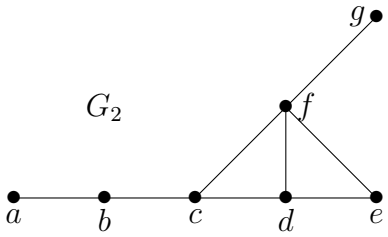


Figure 3: G_2 is AT-free but $\mu = [a, b, c, d, e]$ is a non dominating diametral path of G_2.

The smallest k such that a graph is k-laminar is called the laminar index of G and denoted by $Laminar(G)$. This invariant is well defined since obviously $Laminar(G) \leq diam(G)$ and

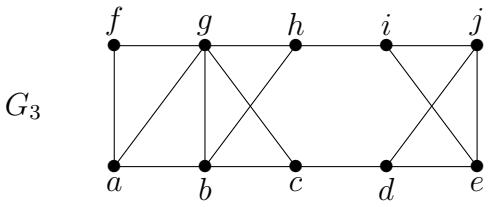


Figure 4: G_3 is not AT-free, (g, i, d) is an AT, but G_3 is 1-strongly laminar.

furthermore if a center of the graph belongs to a diametral path: $Laminar(G) \leq radius(G)$. This paper is devoted to the study of (strongly) k-laminar graphs, their structure but also the existence of polynomial recognition algorithms. Since it is well known that a graph may have an exponential number of diametral paths (see for example the graph in Figure 4), at first glance we can only state that the recognition problem of strongly k-laminar paths is in **co-NP**. In [10], Deogun and Kratsch introduced a very similar graph class, namely the diametral path graphs.

Definition 6. *[10] A graph G is called a diametral path graph if every connected induced subgraph H of G has a dominating diametral path or equivalently H is a 1-laminar graph using definition 3.*

It is not hard to see that all diametral path graphs are 1-laminar. But 1-laminar graphs strictly contain diametral path graphs, as can be seen with the graph $G_1 \setminus \{d\}$ which is no 1-laminar. Therefore $G_1 \in$ 1-laminar graphs $\setminus$ diametral path graphs.

Using a polynomial time algorithm [10] for testing if a graph has a dominating diametral path, we know that the recognition of 1-laminar graphs is polynomial.

Moreover Deogun and Kratsch were able to prove that diametral path graphs that are trees or chordal graphs have simple forbidden subgraphs (polytime-recognizable). But to our knowledge it is still an open question whether diametral path graphs can be recognized in polynomial time.

The remaining part of the paper is organized as follows:

In section 3 we show that the recognition of strongly laminar graphs is polynomial, such as the recognition of k-laminar graphs when k is fixed. To this aim we improve an algorithm from [10] to recognize 1-laminar graphs and we generalize it for every fixed k.

In section 4 we present strong evidence that it is intractable to find the laminar index. In fact we present a reduction which proves that recognizing if a graph G is k-laminar is NP-complete, for a given range of k values in $O(\sqrt{|V(G)|})$.

3 Polynomial algorithms

The main contribution of this section is that we present a polynomial recognition algorithms for any fixed k for k-laminar (resp. strongly k-laminar) graphs. Let us begin with the strongly case.

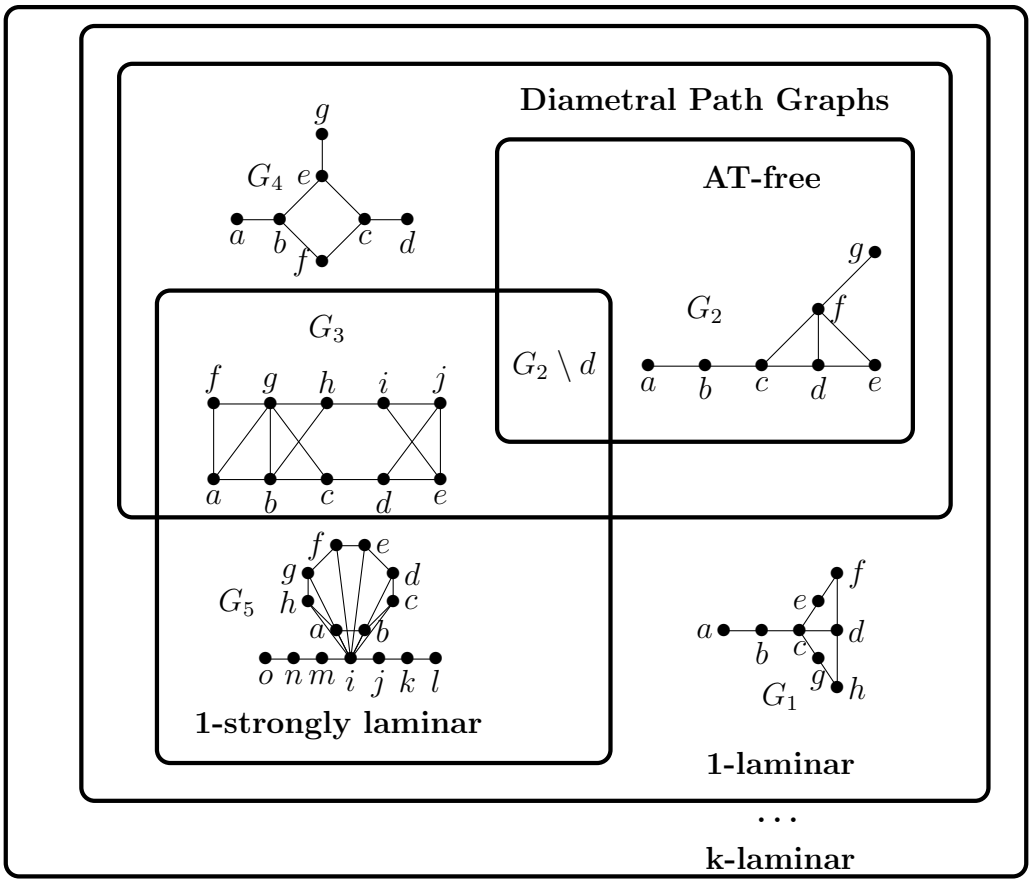


Figure 5: Relationships among the main graph classes considered so far, including that **AT-free** $\subset$ **Diametral Path Graphs** as first noticed in [13].

3.1 Strongly k-laminar graphs

First we need an easy lemma.

Lemma 7. *Let $x \in MaxEcc(G)$ and H be an induced subgraph of G containing x. If $ecc_H(x) = diam(G)$ then there exists a shortest path $\mu = [x, y]$ in H and G of size $diam(G)$.*

Proof. We notice that $ecc_G(x) \leq ecc_H(x)$. In case of equality it yields that there exists a path $\mu = [x, y]$ in H of length $diam(G)$. μ is still a shortest path with no shortcut in G, unless $x \in MaxEcc(G)$. □

Theorem 8. *The recognition of strongly k-laminar graphs can be done in $O(|MaxEcc(G)|nm)$ bounded by $O(n^2m)$ for every fixed k.*

Proof. If a graph is not strongly k-laminar then there exists some diametral path that does not pass through the k-neighborhood of some vertex x. It suffices therefore to verify that every diametral path passes through $N^k[x]$ $\forall x \in V$. This can easily be done by recalculating the distance matrix in $G \setminus N^k[x]$ for every x. We know that $diam(G \setminus N^k[x]) \geq diam(G)$.

If for some vertex x $d_{G\setminus N^k[x]}(a,b) = d_G(a,b) = diam(G)$, using lemma 7 we know there exsit some path μ in G which is a diametral path that does not pass through $N^k[x]$ and therefore the strongly laminar condition is not satisfied.

We need for every vertex x to compute $G' = G \setminus N^k[x]$ which can be done $O(|V(G| + |E(G)|)$ using a BFS. But then we must compute the eccentricity of all $MaxEcc(G)$ vertices in G' which can be done in a naive way by processing $|MaxEcc(G)|$ BFS's in $O(|V(G')|.|E(G')|)$.

Therefore for each k this can be done in $O(|MaxEcc(G)|nm)$, i.e., in $O(n^2m)$. □

As an immediate consequence:

Corollary 9. *The computation of the smallest k for which a graph G is k-strongly laminar is polynomial.*

Proof. Since $1 \leq k \leq n-1$ and we can use a dichotomic process of the above algorithm, which yields a complexity of $O(log(n)n^2m)$. □

3.2 1-laminar graphs

Let us now describe an improved variation of the $O(n^3m)$ Deogun and Kratsch's algorithm [10], searching for the existence of a dominating diametral path in $O(n^2m)$.

As a preprocessing, we can compute $ecc(x)$ for every vertex x of G. Afterwards $\forall s \in MaxEcc(G)$, we process a BFS and let us denote by T_s the associated BFS-tree. L_i represent the different layers of the BFS-tree, i.e. by convention $L_0 = \{s\}$ and L_i is equal to the i-th neighborhood of s. Then $\forall v \in V(G)$, let us denote by $Level_s(v)$ its level in T_s. We can also preprocess in linear time : $\forall v \in V(G)$ and $\forall i$ such that $Level_s(v)$-1 $\leq i \leq Level_s(v)+1$ we compute $N_i(v) = N(v) \cap L_i$.

Then we can use for every vertex $s \in MaxEcc(G)$ the following modified BFS, which is in fact a partial BFS since only the vertices that can be part of a dominating diametral path are explored.

Theorem 10. *Algorithm Dominating-Diameter(G,s) computes if a graph G admits a dominating diametral path starting from s in $O(nm)$.*

Proof. Any diametral path must go sequentially through the all the layers of H_i. Furthermore using the BFS-tree structure any edge $xy \in V(G)$ satisfies $|Level_s(x) - Level_s(y)| \leq 1$.

In order to prove the modified BFS algorithm we need to prove that for every vertex s of maximal eccentricity it is enough to check that the following easy invariant s:

Invariant 11. *For all i, $1 \leq i \leq Diam(G)$, and for every $v \in L_i(s)$ If $v \in Queue$, then there exists a path from s to v in G, that dominates the first $i-1$ layers. Moreover all these dominating paths reach v with an edge marked FEASIBLE.*

Complexity analysis: The preprocessing time, i.e., computing all eccentricities can be done in a naive way by processing n Breadth First searches (BFS) in $O(nm)$.

Dominating-Diameter(G,s):
Data: a graph $G = (V, E)$ and a start vertex $s \in MaxEcc(G)$;
Result: YES / NO G has a dominating diametral path starting at s;
Mark FEASIBLE all edges adjacent to s. Initialize *Queue* to $N(s)$;
while $Queue \neq \emptyset$ **do**
  dequeue v from beginning of *Queue*;
  $h \leftarrow Level_s(v)$;
  for $\forall u \in N_{h-1}(v)$ *with* uv *marked FEASIBLE* **do**
    $A(v) \leftarrow N_h(v) \cup N_h(u)$;
    if $h = diam(G)$ **then**
      if $L_h = A(v)$ **then**
        YES a dominating diametral path from s to v has been found **STOP**
      end
    end
    for $\forall w \in N_{h+1}(v)$ **do**
      if L_h=$A(v) \cup N_h(w)$ **then**
        Mark vw as FEASIBLE;
        if w *is not already in Queue* **then**
          enqueue w to the end of *Queue*
        end
      end
    end
  end
end
NO G has no dominating diametral path starting at s;

Algorithm 1: A modified Breadth First Search

Let us consider a BFS search starting at some $s \in MaxEcc(G)$ and its BFS numbering τ (the visiting ordering of the vertices during the BFS), one can easily sort all the neighborhood lists of all the vertices according to τ in linear time. Then for every vertex $x \in L_h$, $N_{h-1}(x), N_h(x)$ and $N_{h+1}(x)$ can be extracted from $N(x)$ in $O(1)$. Therefore for each BFS before using the modified BFS, the preprocessing requires $O(n + m)$. The structure of the modified BFS, i.e., the while loop, is a partial BFS visiting only vertices that can still belong to a dominating path. Let us now consider the inside instructions.

For every edge uv the test L_h=$A(v) \cup N_h(w)$ can be done by computing $A(v) \cap N_h(w)$ in $O(|A(v)| + |N_h(w)|)$ since they are encoded as sorted lists and then comparing the sizes $|A(v) \cap N_h(w)|$ and $|L_h|$ in $O(1)$.

For every vertex $v \in L_h$, in the whole : $N_h(v)$ is used at most $|N_{h-1}(v)|+|N_{h+1}(v)|$ times.

Therefore for all v it is bounded by $\Sigma_v|N_h(v)|(|N_{h-1}(v)|+|N_{h+1}(v)|)$. Bounding $|N_{h-1}(v)|+|N_{h+1}(v)|$ by n we obtain: $n.\Sigma_v d(v) \in O(n.m)$ Therefore the overall time complexity of this algorithm is $O(nm)$. □

Corollary 12. *The recognition of 1-laminar graphs can be done in* $O(|MaxEcc(G)|.nm)$ *bounded by* $O(n^2m)$.

Proof. To recognize if a graph is 1-laminar, it is enough to process for every $s \in MaxEcc(G)$ the algorithm Dominating-Diameter(G,s). Including the preprocessing and the computation of all eccentricities in G in $O(nm)$, the overall time complexity is $O(|MaxEcc(G)|.nm)$ bounded by $O(n^2m)$. □

This algorithm can be easily adapted to compute a 1-dominating diametral path and generalized for every fixed integer k, and this yields :

Theorem 13. *The recognition of k-laminar graphs can be done in* $O(n^{2k+1})$.

For a proof the reader is referred to the Appendix.

4 NP-completeness

In this section we give a reduction from 3SAT to the recognition of k-laminar graphs. It is therefore NP-hard to compute $Laminar(G)$. The reader is encouraged to look at figure 4 for an better understanding of the reduction. In this section n denotes the number of variables in a satisfiability formula and N the number of vertices in a graph. Capital letters are used to denote vertices and small letters to denote variables.

Given a 3SAT formula ϕ made up with m clauses C_j, $1 \leq j \leq m$, on n boolean variables x_i $1 \leq i \leq n$.

We construct a graph $G(\phi)$ and we will prove that: $G(\phi)$ is $(\frac{n}{2}+1)$-laminar iff ϕ is satisfiable.

Let us first detail the construction of $G(\phi)$. For each literal x_i (resp. $\overline{x_i}$) we associate a vertex X_i (resp. $\overline{X_i}$). We put an edge between a variable and its negation. Moreover we connect the vertices X_i and $\overline{X_i}$ with $X_{i-1}, \overline{X_{i-1}}, X_{i+1}, \overline{X_{i+1}}$ if existent. We add a pending chain $V_1, \dots V_n$ to X_1 and $\overline{X_1}$. The same is done symmetrically with a pending chain $V_{n+1}, \dots V_{2n}$ attached to X_n and $\overline{X_n}$. Up to now we have 2^n shortest paths of length $3n+2$ going from V_1 to V_{2n}. Now for every clause C_i, $1 \leq j \leq m$, we add a vertex C_i. Every C_i is connected by a chain of length $\frac{n}{2}+1$ to every vertex associated to a litteral that appears in the clause C_i. Note that here for sake of simplicity n is supposed to be even, otherwise we would add a dummy variable.

Suppose for now that the diametral path starts and ends from the end vertices of the two chains respectively (V_1 and V_{2n}). Such a diametral path will never pass through X_i and $\overline{X_i}$ because it would either need to use an edge $X_i\overline{X_i}$ or do some detour which would mean that the length of the path is greater than the diameter $3n+2$.

The graph $G(\phi)$ contains exactly $4n+m_\phi(\frac{n}{2}+1) = |V(G)|$ vertices, where m_ϕ is the total number of variables in the clauses C_j.

Lemma 14. $diam(G(\phi)) = 3n+2$

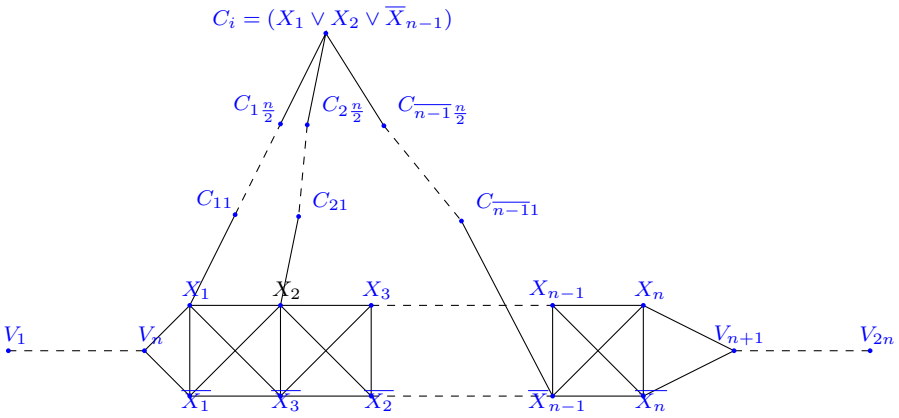


Figure 6: An example of graph $G(\phi)$

Proof. For any pair of clauses : $C_j, C_{j'}$, $d(C_j, C_{j'}) \leq 2(\frac{n}{2}+1)+n \leq 3n$

Furthermore: Let p (resp. q) be the minimum (resp. maximum) index of a literal in C_j.

Then $ecc(C_j) = max\{n-p+n+1, n+1+q\} = max\{2n-p+1, n+q+1\} \leq 2n+1$.

We already have seen that : $ecc(V_1) \leq 3n+2$, using a path going only through the X_i's up to V_{2n}. Moreover no C_j can provide a shortcut to this path. Thus $ecc(V_1) = 3n+2 = ecc(V_{2n})$ □

Theorem 15. *$G(\phi)$ is a $(\frac{n}{2}+1)$-laminar graph iff ϕ is satisfiable.*

Proof. Suppose ϕ is satisfiable and let $\mathbb{A}$ be some satisfying truth assignment of the variables. Consider a path μ from V_1 to V_{2n} forced to visit the vertex X_i if the variable is set to true in $\mathbb{A}$ and $\overline{X_i}$ otherwise.

μ is obviously a diametral path. Since $\mathbb{A}$ is a truth assigment every clause C_j of ϕ has a true literal which belongs to μ and therefore $d(C_j, \mu) \leq \frac{n}{2}+1$. All other vertices either belongs to μ are at distance 1. Therefore μ is $(\frac{n}{2}+1)$-dominating diametral path.

Conversely, suppose $G(\phi)$ is $(\frac{n}{2}+1)$-laminar, hence using Lemma 14 there exists a diametral path μ of length $3n+2$ such that a every vertex is at distance $\frac{n}{2}+1$ from μ. As explained above X_i and $\overline{X_i}$ can not be both on a diametral path. We set the variable x_i to be true if μ passes through the vertex X_i to false otherwise. Every clause C_j must be satisfied because there is at least one variable vertex X_j at distance $\frac{n}{2}+1$ from it. Therefore this $(\frac{n}{2}+1)$-dominating diametral path provides a truth assignment for ϕ. □

It is obvious that the transformation can be computed in polynomial time. Let us consider the following decision problem :

Name: Laminarity

Data: A graph G and k an integer such that $k \in [\frac{\sqrt{|V(G)|}}{4}, \frac{\sqrt{|V(G)|}}{2}]$

Question: Is G k-laminar ?

Corollary 16. *Laminarity is an NP-complete problem.*

Proof. If we consider the 3SAT NP-complete variant in which every variable occurs at most 3 times [14]. The relationship between the number of variables n of such an instance and its number of clauses m is :

$2n \leq m_\phi \leq 3n$ where m_ϕ denotes the total number of occurences of variables in clauses. This inequalities just say that each variable has 2 or 3 occurences in the clauses, since we can get rid of the cases where a variable occurs only in one clause.

Considering the first inequalities we deduce:

$4n + 2n(\frac{n}{2} + 1) \leq |V(G)| \leq 4n + 3n(\frac{n}{2} + 1)$, which gives : $n^2 + 6n \leq |V(G)| \leq 3\frac{n^2}{2} + 7n$.

Replacing n by $2k - 1$ we obtain : $4k^2 + 8k - 5 \leq |V(G)| \leq 12k^2 + 2K - 4$.

Therefore : $4k^2 \leq |V(G)| \leq 16k^2$

If we consider the range $[\frac{\sqrt{|V(G)|}}{4}, \frac{\sqrt{|V(G)|}}{2}]$ for k, using the construction described above we can encode all instances of a NP-complete variant of 3SAT. □

5 Conclusion and perspectives

It would be interesting to improve the running time for the recognition of k-laminar graphs (especially for 1-laminar ones). But it should be noticed that for graphs having a constant number of extremal vertices (i.e., $|MaxEcc(G)|$ is bounded by this constant) then the complexity of the algorithms proposed here in theorems 8,10 goes down to $O(nm)$ which could be optimal, see [5, 1]. In particular when dealing with read networks their laminar parts seem to have a bounded number of extremal vertices.

One of the few theoretical results on clustering for restricted graph classes is presented in [12] and proposes an approximation algorithm for 1-laminar graphs. Therefore we think that these bio-inspired k-laminar graphs are worth to be studied further. As for example, searching for diameter computations in linear time using a constant number of BFSs as in [6] and may have other applications not only in bioinformatics.

Perhaps the k-laminar class of graphs is too large to capture all properties of read networks. The good notion could be k-diametral path graphs with its recursive definition for all induced subgraphs. Unfortunately there is no polynomial recognition algorithm for this class. A good algorithmic compromise would be to add some connectivity requirements, i.e., k-laminar and h-connected. It would be interesting to develop a robust decomposition method of read networks into their k-laminar parts. In other words we want to find a skeleton of a read network that captures most of its biological properties. Such a decomposition could provide an interesting alternative process to analyze the biodiversity of read networks.

Acknowledgements: The authors wish to thank Anthony Herrel for many discussions on the project and for having selected the lizards on which this study is based.

References

[1] A. Abdoud, V.V. Williams, J. Wang, *Approximation and Fixed parameter subquadratic algorithms Radius and Diameter*, in Proceedings of the twenty-seventh annual ACM-SIAM sym-

posium on Discrete Algorithms, p. 377-391, SIAM, 2016.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Basic local alignment search tool*, Journal of molecular biology, vol. 3, p. 403-410, 2015.

[3] E. Bapteste, M. Habib, A. Herrel, P. Lopez, C. Vigliotti, *Projet Evolézards, Défi ENVIROMICS, Mission Interdisciplinarité CNRS*, 2014.

[4] E. Boon, S. Halary, E. Bapteste, M. Hijri, *Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm,* Genome Biol Evol. Jan 7;7(2): 505-521, 2015.

[5] M. Borassi, P. Crescenzi, M. Habib, *Into the Square - On the Complexity of Quadratic-Time Solvable Problems*, CoRR, abs/1407.4972, 2014.

[6] M. Borassi, P. Crescenzi, M. Habib, W. A. Kosters, A. Marino, F. W. Takes, *Fast diameter and radius BFS-based computation in (weakly connected) real-world graphs: With an application to the six degrees of separation games, Theor. Comput. Sci.*, Vol 586, p. 59-80, 2015.

[7] Phillip E C Compeau, Pavel A Pevzner , Glenn Tesler, *How to apply de Bruijn graphs to genome assembly,* Nature Biotechnology 29, p. 987-991, 2011

[8] D.G. Corneil, S. Olariu, L. Stewart, *Asteroidal triple-free graphs*, SIAM J. Discrete Math., Vol 10, No. 3, p. 399-430, 1997.

[9] M.C. Golumbic, C.L. Monma, W.T. Trotter, *Tolerance graphs*, Disc. Applied Math. 9, p. 157-170, 1997.

[10] J.S. Deogun, D. Kratsch, *Diametral Path Graphs*, Graph Theoretic Concepts in CS, p. 344-357, 1999.

[11] J.S. Deogun, D. Kratsch, *Dominating Pair Graphs*, SIAM J. Discrete Math., Vol. 15, No. 3, p. 353-366, 2002.

[12] J.S. Deogun, D. Kratsch, G. Steiner, *An approximation algorithm for clustering graphs with dominating diametral path*, Information Processing Leters, 61, p. 121-127,1997.

[13] D. Kratsch, *The structure of graphs and the design of efficient algorithms*, Habilitation Thesis, Friedrich-Schiller-Universität, Jena, 1995.

[14] C. H. Papadimitriou, *Computational Complexity*, 1994, Addison-Welsey.

[15] Y. Peng, Henry C. Leung, S.M. Yu, F.Y. L. Chin, *IBDA-UD a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth*, Bioinformatics, Volume 28, Issue 11, p. 1420-1428, 2012.

[16] J. H. Saw, A. Spang, K. Zaremba-Niedzwiedzka, L. Juzokaite, J. A. Dodsworth, S. K. Murugapiran, D. R. Colman, C. Takacs-Vesbach, B. P. Hedlund, L. Guy and T. J. Ettema, *Exploring microbial dark matter to resolve the deep archaeal ancestry of eukaryotes.* Philos Trans R Soc Lond B Biol Sci 370(1678): 20140328 (2015).

6 Appendix

6.1 k-laminar recognition algorithm

We first notice that every graph G is trivially $diam(G)$-laminar, and let us now generalize the previous recognition algorithm 1 to any fixed integer k such that : $k < diam(G)$.

Theorem 17. *For every fixed $k \geq 2$ such that $k < diam(G)$, the algorithm k-Dominating-Diameter(G,s) finds a k-dominating diametral path starting form s if some exists in $O(n^{2k})$.*

Proof. To generalize Dominating-Diameter(G,s) algorithm, we will proceed similarly from a given vertex $s \in MaxEcc(G)$, by considering all the paths of length k starting at s and then make them grow layer by layer keeping only those which are potential extendable to a k-dominating diametral path.

We keep the same preprocessing as for the recognition of 1-laminar graphs, namely: we can compute $ecc(x)$ for every vertex x of G. Afterwards $\forall s \in MaxEcc(G)$, we process a BFS and let us denote by T_s the associated BFS-tree. L_i represent the different layers of the BFS-tree, i.e. by convention $L_0 = \{s\}$ and L_i is equal to the i-th neighborhood of s. Then $\forall v \in V(G)$, let us denote by $Level_s(v)$ its level in T_s. We can also preprocess in the same time : $\forall v \in V(G)$ and $\forall i$ such that $Level_s(v)$-$k \leq i \leq Level_s(v) + k$ we compute $N_i^k(v) = N^k(v) \cap L_i$. Since k is fixed, the sets $N_i^k(v)$ can be computed in $O(nm)$ also.

Invariant: If the pair (v, P) with $P = [s, \dots v]$ belongs to Queue, and if $p = Level_s(v)$ then P a path k-dominating the first $p - k + 1$ layers.

This invariant is clearly satisfied with the initializations of Queue. Then During the While loop a new pair (v, P) is only inserted if it satisfies this property.

Complexity Analysis: The initialisation step may costs $O(n^k)$ since we could have $\prod_{i=1}^{h} |L_i|$ different pairs (v, P). The queue data structure forces the vertices to be visited in a breadth first way, giving an $O(n + m)$ to the managment of the while loop. During this while loop:

For every vertex v the set $A(v)$ is used at most $|N_{p+1}(v)|$ times, so in the whole it is bounded by $O(nm)$. But to compute the sets $A(v)$ we have to maintain paths of length 2k. Unfortunately there could be n^{2k} such paths. This yields a polynomial algorithm in $O(n^{2k})$. □

Corollary 18. *k-laminar graphs can be recognized in $O(|MaxEcc(G)|.n^{2k})$ or $O(n^{2k+1})$.*

Proof. First we have to compute all eccentricites in G in $O(nm)$ and then it is enough to repeat this Algorithm 2 for every $x \in MaxEcc(G)$, this provides an algorithm running $O(|MaxEcc(G)|.n^{2k})$.

□

k-Dominating-Diameter(G,s):
Data: a graph $G = (V, E)$ and a start vertex $s \in MaxEcc(G)$, an integer $k \geq 2$;
Result: YES / NO G has a k-dominating diametral path starting at s;
Initialisations:
Initialize a queue *Queue* with all different pairs (v, P) such that P is a path of length k starting at s in the BFS-tree, and v its last vertex. This list is supposed to be lexicographically ordered accordingly to the layer orderings.
while $Queue \neq \emptyset$ **do**
    dequeue (v, P) from beginning of *Queue*;
    $p \leftarrow Level_s(v)$;
    $q \leftarrow p - k + 2$;
    $A(v) \leftarrow \bigcup_{u \in P, q-k+1 \leq Level_s(u) \leq p} N_q^k(u)$;
    if $p = diam(G)$ **then**
        if $\forall i,\ q \leq i \leq p,\ L_i = \bigcup_{u \in P, i-k+1 \leq Level_s(u) \leq i+k} N_i^k(u)$ **then**
            YES "a k-dominating diametral path from s to v has been found"**STOP**
        end
    end
    for $\forall x \in N_{p+1}(v)$ **do**
        if L_q=$A(v) \cup N_q^k(x)$ **then**
            $P' \leftarrow P + x$;
            enqueue (x, P') to the end of *Queue* ;
        end
    end
end
NO "G has no k-dominating diametral path starting at s";

Algorithm 2:

5.4 Les boucles et points de jonction

Biologiquement, les reads composant un même laminaire sont donc retrouvés dans le même ordre dans la population des génomes de notre jeu de données. Par conséquent, un laminaire représente un fragment conservé de génome(s) microbien(s). Les RSS de reads sont donc de très bons outils pour déterminer quels sont les contextes génomiques stables dans un métagénome. Cependant, l'intérêt biologique des autres formes de composantes connexes (Figure 30) n'a pas encore été explicité. Les formes 30 a) c) et d) traduisent la présence de reads dans plusieurs contextes génomiques différents. En effet, une boucle au sein d'une structure (Figure 30 c) d)) peut être interprétée comme un variant architectural de la communauté caractérisé par une insertion. La présence de plusieurs boucles au sein d'une composante connexe (cf Figure 30 a)) peut être interprétée comme un ensemble de répétitions et de variations autour de régions génomiques conservées (Figure 33).

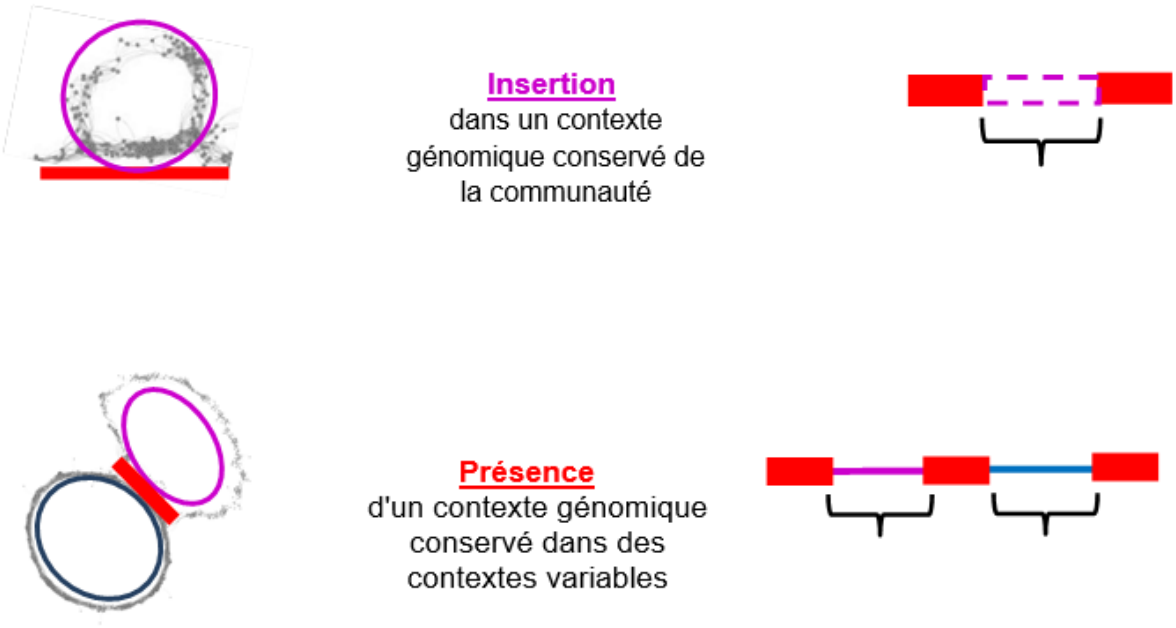


Figure 33 : Interprétation biologique potentielle des boucles dans les composantes connexes.

Ci-dessous (Figure 34), un autre exemple de composante connexe illustre le cas des points de jonction déterminés par la méthode de décomposition de Habib et Volkel (évoquée ci-dessus et illustré Figure 30) :

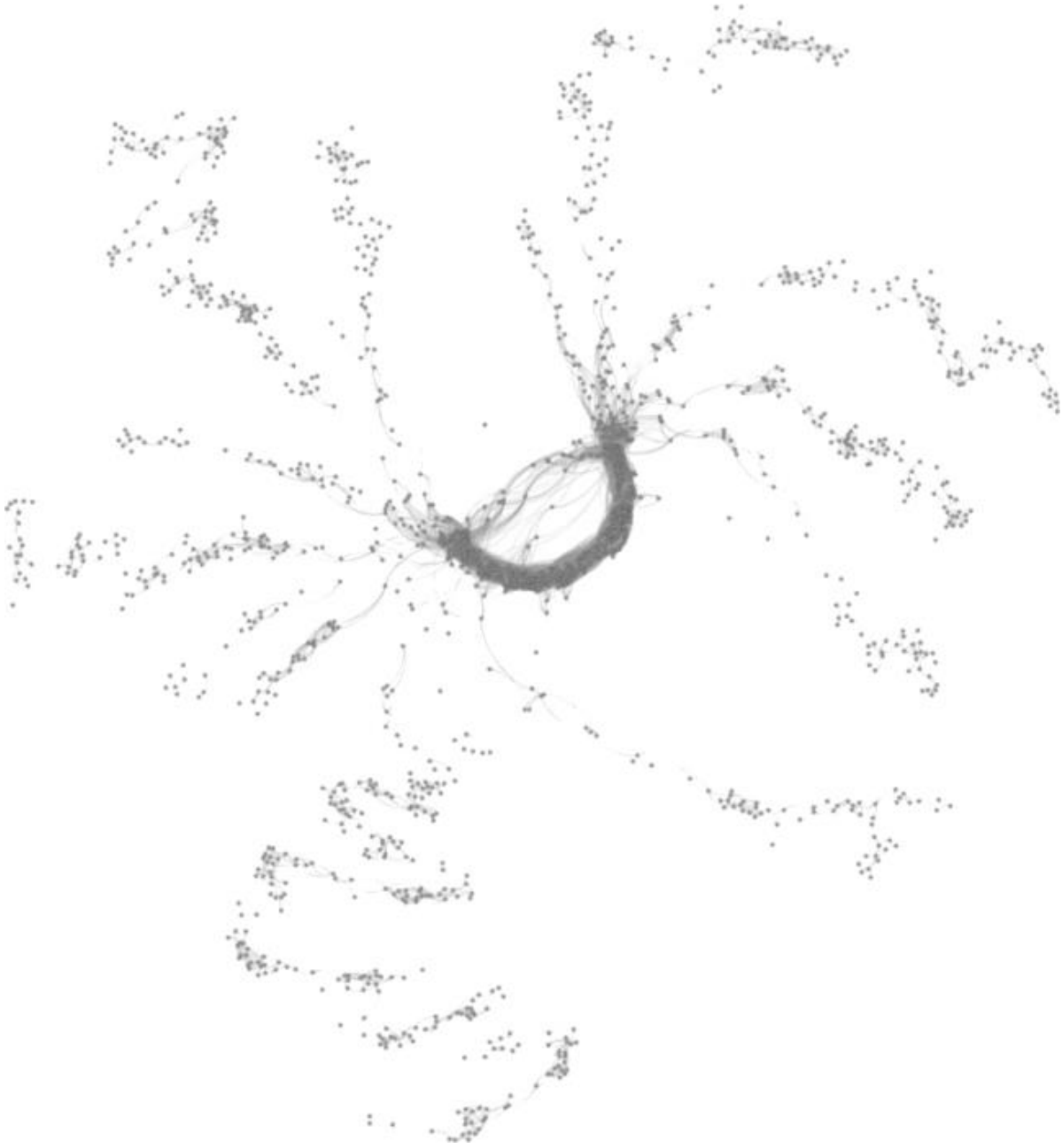


Figure 34 : Composante connexe (1843 nœuds, 25 022 arêtes) identifiée par la méthode d'Habib et Volkel parmi les 544 808 composantes connexes RSS du microbiome du lézard PSK21MDI.

(5 565 118 noeuds, 329 467 945 arêtes, comparaison BLAST avec pid ≥ 90%, cover≥ 80%, E-value ≤ 1e-5).

La boucle centrale présente deux points de jonction, auxquels sont raccrochés des laminaires. Cette situation peut être interprétée comme des variations (les

laminaires) de contexte génomique autour d'un contexte génomique conservé. Effectivement, en annotant taxonomiquement et fonctionnellement (à l'aide de BLAST) la composante connexe de la Figure 34, on constate que la boucle centrale est une transposase (Figure 35).

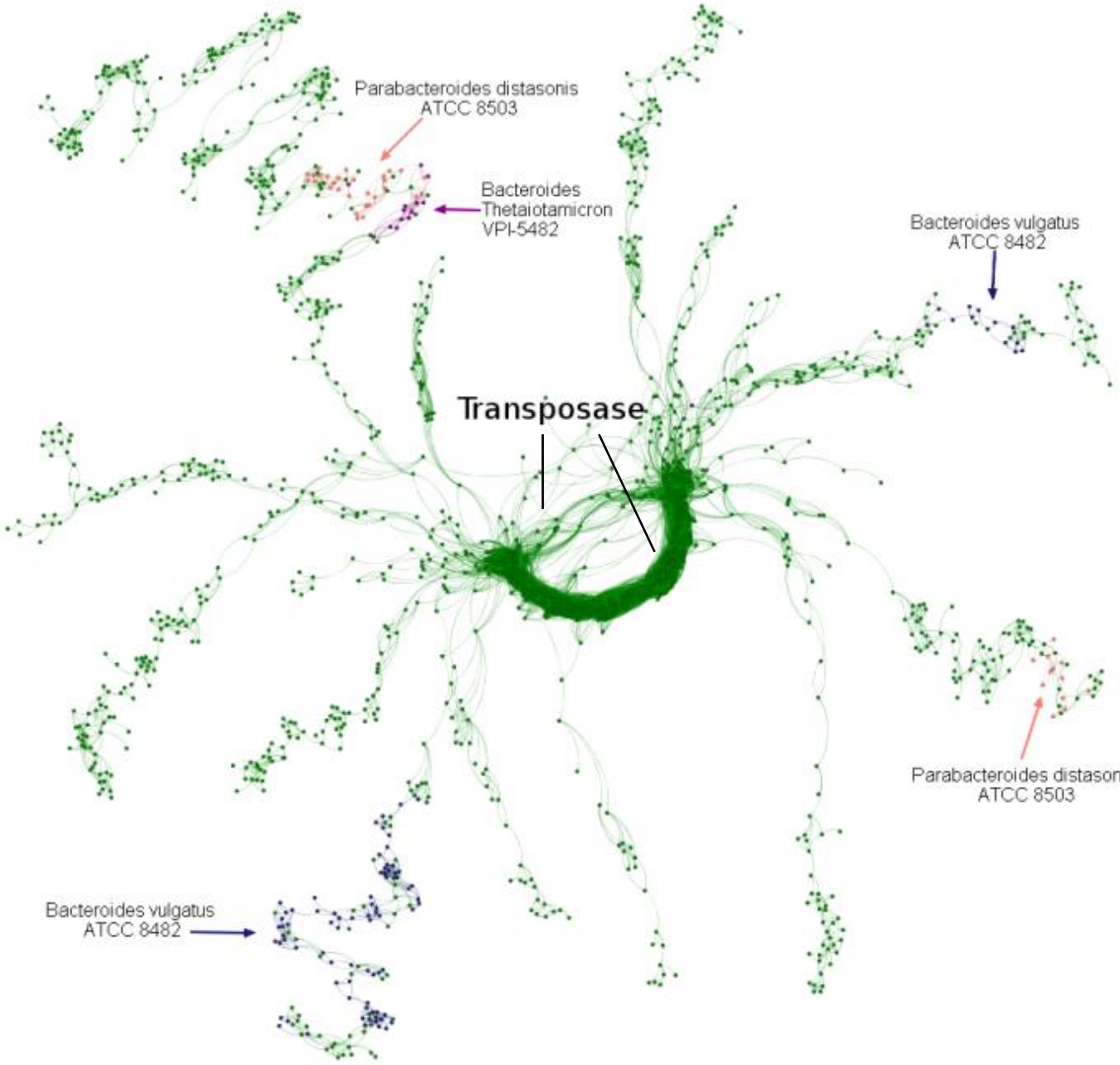


Figure 35 : Annotation taxonomique de la composante connexe présentée Figure 34.

Les nœuds en vert sont ceux qui n'ont pas été annotés. Outil de visualisation : Gephi (spatialisation : multi-niveaux de Yifan Hu)

Les transposases sont des protéines permettant l'insertion d'éléments transposables dans les génomes (Heffron et al. 1979). Si l'on regarde les quelques annotations taxonomiques que l'on a obtenu sur les laminaires, on constate que cette transposase se retrouve dans des contextes génomiques différents dans le métagénome étudié, puisqu'on la trouve au moins dans une souche de *Bacteroides vulgatus*, de *Parabacteroides distasonis*, et de *Parabacteroides thetaiotamicron*. Cependant, le RSS montre que le contexte génomique en amont et en aval de cette transposase est différent selon les génomes considérés. Une étude plus systématique de ce type complexe de composante connexe pourrait donc nous renseigner sur la diversité architecturale des génomes présents dans un environnement donné.

5.5 Création d'indices quantifiables afin de pouvoir analyser statistiquement la diversité

Après avoir établi l'intérêt biologique de ces réseaux et les différents contextes de diversité génétique qu'ils représentent, nous nous sommes intéressés au fait de pouvoir quantifier cette diversité à l'aide d'indices. Michel Habib et son doctorant Léo Planche ont alors utilisé deux indices, la laminarité et la delta-hyperbolicité. L'indice de laminarité est k, défini précédemment. La delta hyperbolicité est un indice permettant d'appréhender le diamètre des composantes connexes. Cet indice a été défini de la façon suivante par M. Gromov : pour 4 points quelconques du graphe u,v,w,x les deux plus grande valeurs des sommes d(u,v) + d(w,x), d(u,w) + d(v,x) et d(u,x) + d(v,w) diffèrent au plus de 2 x delta (Chepoi et al. 2008; Gromov 1987). Plus delta est petit pour un graphe, plus ce graphe est proche d'un simple arbre (i.e. un graphe acyclique).

Cette collaboration avec Michel Habib nous a permis d'écrire un premier article sur les k-laminaires, qui apparaissent comme une nouvelle classe de graphe. Cette méthode offre une alternative pour étudier la diversité génétique présente dans les microbiomes.

6. Conclusion

6.1 De la diversité des méthodes à la standardisation des analyses

L'un des objectifs de cette thèse a été de déterminer si les études de microbiome et de microbiote sont engagées sur des sentiers de dépendance. Nous proposons que c'est le cas pour les analyses de microbiotes. En effet, il existe des méthodes très standards d'analyser le microbiote (filtrage des reads en fonction de leur qualité, création des OTUs, analyses d'alpha et beta diversité, tables d'abondances relatives au niveau du phylum et du genre, LefSe, ... *cf* chapitre 3). En revanche, selon nous, les analyses de microbiome ne sont pas encore engagées sur un sentier de dépendance, avec une standardisation des analyses. Cela s'illustre par la possibilité d'étudier les microbiomes aussi bien à l'échelle du read (Blasco et al. 2017; Le Chatelier et al. 2013) qu'à l'échelle de l'ORF (HMP 2012). Cela s'illustre également par la quantité de logiciels d'assemblage (Boisvert et al. 2012; Ghurye, Cepeda-Espinoza, and Pop 2016; Namiki et al. 2012; Peng et al. 2011, 2012; Treangen et al. 2011; Zerbino 2010; Zerbino and Birney 2008) permettant de générer les contigs (sur lesquels on cherche les ORFs), ainsi que la diversité des valeurs possibles pour chaque paramètre des assembleurs. Les microbiomes peuvent être étudiés sous l'angle fonctionnel en considérant les catégories COGs (Turnbaugh et al. 2008), ou les voies métaboliques Kegg (Huttenhower and Human Microbiome Project Consortium 2012; Yatsunenko et al. 2012). Et il existe évidemment d'autres façons d'étudier les microbiomes, en particulier des perspectives évolutionnistes.

6.2 Le changement de régime alimentaire de Podarcis sicula est associé à des changements ciblés dans le microbiote

L'un des objectifs majeurs de cette thèse était de déterminer si une modification du microbiote est associée au changement de régime alimentaire. Nous avons pu établir que l'alpha diversité des microbiotes de lézards omnivores est plus importante que celle des microbiotes de lézards insectivores, ce qui est en adéquation avec les résultats de l'étude de R.E. Ley sur l'évolution des microbiomes intestinaux de mammifères (Ley et al. 2008).

Nous n'avons pas trouvé de changement significatif en terme de beta diversité entre les microbiotes de lézards insectivores et omnivores, ni de changements corrélés à l'insularité, à la géographie, à la saison ou à l'année de séquençage. Par ailleurs, nous n'avons trouvé aucune différence significative entre les microbiotes de lézards mâles et femelles. Cela n'était pas évident dans la mesure où les microbiotes intestinaux des mâles et des femelles de souris, par exemple, présentent des différences (Fransen et al. 2017; Org et al. 2016).

Après avoir constaté qu'il y a une plus grande diversité d'espèces dans les microbiotes de lézards omnivores que dans les microbiotes de lézards insectivores, nous avons cherché à identifier une structure de la population de *Podarcis sicula* étudiée, basée sur la composition du microbiote intestinal. Pour cela, on a choisi de chercher des entérotypes au niveau du phylum et au niveau du genre en suivant le protocole proposé par Arumugam et Raes (Arumugam et al. 2011). Nous avons trouvé 5 entérotypes au niveau du phylum, mais le pouvoir explicatif de ce modèle est faible (30.5%). Ainsi, il ne semble pas y avoir de structure claire de la population de *Podarcis sicula* en se basant sur la composition taxonomique du microbiote intestinal.

Cependant, nous avons pu observer que le microbiote intestinal des *Podarcis sicula* comporte une structure globale commune. En effet il contient trois phyla majoritaires : Bacteroidetes, Firmicutes et Protéobactéries. Ces phyla font partie des phyla les plus abondants chez les mammifères, et deux de ces trois phyla sont majoritaires (Bacteroidetes et Firmicutes) chez l'iguane.

Toujours afin de trouver une structure dans la population de *Podarcis sicula*, nous avons recherché le microbiote ubiquitaire (i.e. les OTUs présentes dans tous les microbiotes intestinaux de lézards). Nous avons trouvé un microbiote ubiquitaire composé de 158 OTUs sur les 32 000 OTUs présentes dans le jeu de données. Le fait que ce microbiote ubiquitaire soit petit peut être dû au sous échantillonnage. Nous avons aussi regardé l'abondance des OTUs, et trouvé que seules 25 OTUs pouvaient être considérées comme abondantes (i.e. représentant au moins 10% d'un microbiote).

Nous avons ensuite voulu comprendre quels taxons étaient à l'origine des différences d'alpha diversité observées entre insectivores et omnivores. Pour cela, nous avons appliqué une analyse linéaire discriminante avec effet de la taille (LefSe).

Cela permet de trouver les taxons dont les abondances distinguent deux populations (ici, insectivores *vs* omnivores, île *vs* continent, années et saisons d'échantillonnage, Pod Kopište *vs* Pod Mrčaru *vs* continent, mâle *vs* femelle). Nous avons identifié 4 phyla discriminants (Spirochaetes, Euryarchaeota, Elusimicrobia, Planctomycètes). Ces phyla sont, selon d'autres études, impliqués dans la digestion des plantes (Herlemann, Geissinger, and Brune 2007; Mountfort, Asher, and Bauchop 1982; Piknová et al. 2006; Santana et al. 2015; Su et al. 2016; Wei et al. 2016). Notre travail a aussi permis de montrer l'importance des archées méthanogènes (dont certaines Euryarchaeota telles que *Methanobrevibacter*) dans la digestion de plantes chez les *Podarcis sicula* omnivores. Suite à cela, les différences d'abondance en archées et plus précisément entre méthanogènes ont été étudiées, entre les microbiotes de lézards insectivores et omnivores. Il a été constaté que les lézards omnivores ont un microbiote globalement enrichi en archées, et plus spécifiquement en méthanogènes par rapport aux lézards insectivores.

Enfin, après avoir recherché les taxons permettant de distinguer des groupes de lézards en fonction de différentes variables (régime alimentaire, genre, insularité,...) nous avons souhaité créer un modèle qui à l'aide des variables précédemment détaillées, permette d'expliquer la composition de la table d'abondance (à l'échelle du phylum, puis du genre) du jeu de données. Nous avons constaté que les variables utilisées dans notre étude ne permettent pas d'expliquer cette table d'abondance (le pouvoir explicatif du modèle, prenant en compte toutes les variables, n'explique que 18% de la composition au niveau du phylum et 15% au niveau du genre). Nous avons aussi essayé d'expliquer cette composition à l'aide des entérotypes trouvés au niveau du phylum, cependant cela n'expliquait que 5% de la table d'abondance.

6.3 Le changement de régime alimentaire de Podarcis sicula est associé à des changements ciblés dans le microbiome

Les études de microbiomes n'étant pas encore engagées sur un sentier de dépendance, certains de nos protocoles sont similaires à ce que l'on peut trouver dans d'autres études (abondance en reads de chaque catégorie COG) et d'autres le sont

moins (analyse linéaire discriminante sur les enzymes de chaque voie métabolique, analyse de la diversité des microbiomes à l'aide de réseaux de similarités de reads).

La comparaison de l'abondance relative des différentes catégories COG a permis d'observer que les microbiomes de lézards omnivores contiennent plus de prédictions de fonction générale que les microbiomes de lézards insectivores. En revanche, les microbiomes de lézards insectivores contiennent davantage de fonctions de production et conversion d'énergie et de fonctions associées au métabolisme et au transport des lipides que dans les microbiomes omnivores (*cf* chapitre 4). Les fonctions des microbiomes des lézards omnivores sont donc moins bien connus que ceux des lézards insectivores.

Néanmoins, l'analyse des voies métaboliques a permis de montrer qu'il existe peu de voies exclusives des microbiomes insectivores ou omnivores pour un échantillon de 12 individus. Sur le plan des gènes métaboliques, les différences entre ces deux groupes de lézards se situent plutôt en terme d'abondance des enzymes (et non en terme de présence). En effet, certaines voies métaboliques telles que la dégradation de l'Atrazine ou les phosphorylations oxydatives permettent de distinguer les lézards insectivores des lézards omnivores dans notre jeu de donnée de reads.

Il serait donc très intéressant d'étudier les voies métaboliques à partir des prédictions, non plus de reads, mais d'ORFs, réalisées par Guillaume Bernard sur un jeu de données plus large (62 individus au moment de terminer cette thèse). L'une des perspectives à ce travail est donc de déterminer les voies métaboliques présentant des différences significatives entre les microbiomes de lézards insectivores et omnivores.

6.4 Proposition de l'hypothèse des changements ciblés

Pris dans leur ensemble, et en dépit de leur nature préliminaire, nos travaux sur le microbiote et le microbiome convergent dans leurs observations. Le changement de régime alimentaire des lézards ne s'est pas accompagné d'un bouleversement majeur des taxas et des fonctions de leurs communautés microbiennes intestinales. Au contraire, les changements ont été ciblés, portant sur des taxas précis et des fonctions

précises dont l'abondance a varié. Nous proposons donc que dans le cas des lézards *P. sicula* les communautés intestinales se sont adaptées par petites touches à l'herbivorie. Cette conception nous parait compatible avec les théories qui attribuent au microbiome intestinal une diversité de fonctions. Un changement profond de microbiome et de microbiote aurait probablement eu un impact trop considérable sur la biologie de ces populations pour être supportés. Bien que la consommation de 80% de végétaux soit une variation susbtantielle du mode de vie de ces lézards, le microbiome et le microbiote, à l'instar de la génétique de l'hôte et de son environnement, n'ont changé que marginalement.

L'analyse plus aboutie des microbiomes d'un plus grand nombre d'individus, récoltés durant cette thèse, devrait permettre de tester cette hypothèse.

6.5 De la diversité des contextes génomiques dans les réseaux de similarité de reads

A l'aide des réseaux de similarité de reads, on a cherché à identifier les différents types de contextes génomiques des microbiomes de lézards, et de quantifier la diversité de ces contextes. Cette étude a permis de définir une nouvelle classe de graphe, les k-laminaires (Birmelé, de Montgolfier, and Planche 2016; Völkel et al. 2016) mais aussi de caractériser les différentes topologies des réseaux de reads et de les décomposer en formes plus simples. Une perspective de cette étude est de l'étendre à l'ensemble du jeu de données en construisant des réseaux de reads pour chaque lézard. Cela permettra de pouvoir effectuer des statistiques afin de comparer les microbiomes insectivores et omnivores, et ainsi déterminer s'ils présentent la même diversité génétique sur la base de ces nouvelles mesures.

6.6 Recherche des règles d'introgression et de transmission dans les microbiomes à l'aide de réseaux

Au cours de cette thèse l'aspect théorique de l'étude des phénomènes d'introgression et de transmission dans les microbiomes à l'aide de réseaux a été

abordé. Cela a donné lieu à l'écriture du chapitre « Tracking the rules of transmission and introgression with networks » dans l'ouvrage « Experimental and theoretical modes of transmission ». Une des perspectives naturelle à ce travail théorique est l'application de ces méthodes au jeu de données microbiome des lézards. Les reads des microbiomes ont déjà été assemblées et les ORFs ont été prédites. Il reste cependant à construire les réseaux de similarités de séquences, puis les graphes bipartis correspondant, qui représenteront quelles familles de gènes sont présentes dans quels microbiomes. L'utilisation des graphes bipartis permettra par exemple de déterminer quelles sont les familles de gènes exclusivement présentes dans les microbiomes de lézards insectivores ou omnivores. Une autre perspective consistera à analyser ces ORFs prédites et à quantifier leur mobilisation potentielle par des éléments génétiques mobiles, en reprenant le protocole décrit dans le chapitre intitulé « Tracking the rules of transmission and introgression with networks » (Vigliotti et al. 2017).

6.7 Perspective : Quantifier et identifier la matière noire.

Enfin, il demeure beaucoup de séquences non annotées taxonomiquement dans les microbiotes intestinaux de lézards (environ 50% des OTUs construites à partir de l'ARNr 16S sont non annotées au niveau du genre, et 75% des ORFs correspondent à de nouvelles lignées non identifiées, c'est-à-dire moins de 85% identiques à des ORFs de micro-organismes connus). De la même façon, 26.5% des ORFs prédites sont non annotées fonctionnellement. Il y a donc un peu plus d'un quart des fonctions du microbiome qui sont inconnues, ce qui suggère que les différents rôles du microbiome de *P. sicula* restent largement à préciser. Il serait donc particulièrement intéressant de porter l'emphase sur cette « matière microbienne noire », ou bien en identifiant quels micro-organismes sont présent dans ces communautés, ou tout du moins, quel rôle cette matière noire pourrait jouer dans la communauté. L'une des façons de procéder pour annoter davantage de cette matière noire, serait de construire les OTUs de reads de V4 à des seuils moins stringents au niveau du genre (94.5% d'identité) ou au niveau du phylum (75% d'identité), et ainsi

de repérer les séquences qui n'étaient pas annotées précédemment, au sein des nouvelles OTUs. Néanmoins, la faible quantité d'assignation taxonomique effectivement réalisée en utilisant des ORFs suggère que le constat que de nombreux microbes présents dans les intestins des lézards sont inconnus demeurera valide.

C'est pourquoi les nombreuses données accumulées pendant cette thèse ne sont probablement que le premier pas de la découverte des relations entre hôtes et microbes dans ce système non modèle d'holobionte.

Bibliographie

Abecia, L., A. I. Martín-García, G. Martínez, C. J. Newbold, and D. R. Yáñez-Ruiz. 2013. "Nutritional Intervention in Early Life to Manipulate Rumen Microbial Colonization and Methane Output by Kid Goats postweaning1." *Journal of Animal Science* 91:4832–40. Retrieved (http://dx.doi.org/10.2527/jas.2012-6142).

Arumugam, Manimozhiyan et al. 2011. "Enterotypes of the Human Gut Microbiome." *Nature* 473(7346):174–80. Retrieved (http://dx.doi.org/10.1038/nature09944).

Avila, Maria, David M. Ojcius, and Özlem Yilmaz. 2009. "The Oral Microbiota: Living with a Permanent Guest." *DNA and Cell Biology* 28(8):405–11. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2768665/).

Bäckhed, Fredrik, Ruth E. Ley, Justin L. Sonnenburg, Daniel A. Peterson, and Jeffrey I. Gordon. 2005. "Host-Bacterial Mutualism in the Human Intestine." *Science* 307(5717):1915 LP-1920. Retrieved (http://science.sciencemag.org/content/307/5717/1915.abstract).

Bapteste, E., C. Bicep, and P. Lopez. 2012. "Evolution of Genetic Diversity Using Networks: The Human Gut Microbiome as a Case Study." *Clinical Microbiology and Infection* 18:40–43. Retrieved (http://www.sciencedirect.com/science/article/pii/S1198743X14609678).

Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. 2009. "Gephi: An Open Source Software for Exploring and Manipulating Networks." Retrieved (http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154).

Belkaid, Yasmine and Timothy Hand. 2014. "Role of the Microbiota in Immunity and Inflammation." *Cell* 157(1):121–41. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4056765/).

Bennett, Darin C., Hein Min Tun, Ji Eun Kim, Frederick C. Leung, and Kimberly M. Cheng. 2013. "Characterization of Cecal Microbiota of the Emu (Dromaius Novaehollandiae)." *Veterinary Microbiology* 166(1):304–10. Retrieved (http://www.sciencedirect.com/science/article/pii/S0378113513003003).

Binnewies, Tim T. et al. 2006. "Ten Years of Bacterial Genome Sequencing: Comparative-Genomics-Based Discoveries." *Functional & Integrative Genomics* 6(3):165–85. Retrieved (https://doi.org/10.1007/s10142-006-0027-2).

Birmelé, Etienne, Fabien de Montgolfier, and Léo Planche. 2016. "Minimum Eccentricity Shortest Path Problem: An Approximation Algorithm and Relation with the K-Laminarity Problem." *CoRR* abs/1609.04593. Retrieved (http://arxiv.org/abs/1609.04593).

Blasco, Gerard et al. 2017. "The Gut Metagenome Changes in Parallel to Waist Circumference, Brain Iron Deposition, and Cognitive Function." *The Journal of Clinical Endocrinology & Metabolism* 102(8):2962–73. Retrieved (+).

BLAST. n.d. "NCBI BLAST Web site[http://blast.ncbi.nlm.nih.gov/Blast.cgi]." Retrieved

(http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Boisvert, Sébastien, Frédéric Raymond, Élénie Godzaridis, François Laviolette, and Jacques Corbeil. 2012. “Ray Meta: Scalable de Novo Metagenome Assembly and Profiling.” *Genome Biology* 13(12):R122. Retrieved (http://dx.doi.org/10.1186/gb-2012-13-12-r122).

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30(15):2114–20. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103590/).

Boon, Eva, Sébastien Halary, Eric Bapteste, and Mohamed Hijri. 2015. “Studying Genome Heterogeneity within the Arbuscular Mycorrhizal Fungal Cytoplasm.” *Genome Biology and Evolution* 7(2):505–21. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350173/).

Borcard, Daniel, François Gillet, and Pierre Legendre. 2011. *Numerical Ecology With R*.

Bordenstein, Seth R. and Kevin R. Theis. 2015. “Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes.” *PLoS Biology* 13(8).

Bradnam, Keith R. et al. 2013. “Assemblathon 2: Evaluating de Novo Methods of Genome Assembly in Three Vertebrate Species.” *GigaScience* 2:10. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3844414/).

Bray, J.Roger and J. T. Curtis. 1957. “An Ordination of the Upland Forest Communities of Southern Wisconsin.” *Ecological Monographs* 27(4):325–49. Retrieved (http://dx.doi.org/10.2307/1942268).

Burke, Catherine, Staffan Kjelleberg, and Torsten Thomas. 2009. “Selective Extraction of Bacterial DNA from the Surfaces of Macroalgae.” *Applied and Environmental Microbiology* 75(1):252–56. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612226/).

Buttigieg, Pier Luigi and Alban Ramette. 2014. “A Guide to Statistical Analysis in Microbial Ecology: A Community-Focused, Living Review of Multivariate Data Analyses.” *FEMS Microbiology Ecology* 90(3):543–50. Retrieved (+).

Caliński, T. and J. Harabasz. 1974. “A Dendrite Method for Cluster Analysis.” *Communications in Statistics* 3(1):1–27. Retrieved (http://www.tandfonline.com/doi/abs/10.1080/03610927408827101).

Camacho, Christiam et al. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10(1):421. Retrieved (http://dx.doi.org/10.1186/1471-2105-10-421).

Caporaso, J.Gregory et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7(5):335–36. Retrieved (http://www.nature.com/naturemethods/.%5Cnhttp://dx.doi.org/10.1038/nmeth.f.303).

Carberry, Ciara A., David A. Kenny, Sukkyan Han, Matthew S. McCabe, and Sinead M. Waters. 2012. “Effect of Phenotypic Residual Feed Intake and Dietary Forage Content on the Rumen Microbial Community of Beef Cattle.” *Applied and Environmental Microbiology* 78(14):4949–58. Retrieved

(http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3416373/).

Le Chatelier, E. et al. 2013. “Richness of Human Gut Microbiome Correlates with Metabolic Markers.” *Nature* 500(7464):541–46. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/23985870).

Chepoi, Victor, Feodor Dragan, Bertrand Estellon, Michel Habib, and Yann Vaxès. 2008. “Diameters, Centers, and Approximating Trees of Delta-Hyperbolicgeodesic Spaces and Graphs.” Pp. 59–68 in *Proceedings of the Twenty-fourth Annual Symposium on Computational Geometry*, *SCG ’08*. New York, NY, USA: ACM. Retrieved (http://doi.acm.org/10.1145/1377676.1377687).

Ciccarelli, F. D. et al. 2006. “Toward Automatic Reconstruction of a Highly Resolved Tree of Life.” *Science* 311 (5765):1283-7. Retrieved (http://dx.doi.org/10.1126/science.1123061).

CLARKE, K. R. 1993. “Non-Parametric Multivariate Analyses of Changes in Community Structure.” *Australian Journal of Ecology* 18(1):117–43. Retrieved (http://dx.doi.org/10.1111/j.1442-9993.1993.tb00438.x).

Cock, Peter J. A., John M. Chilton, Björn Grüning, James E. Johnson, and Nicola Soranzo. 2015. “NCBI BLAST+ Integrated into Galaxy.” *GigaScience* 4(1):1. Retrieved (+).

Colwell , Robert K., Coddington, Jonanthan A. 1994. “Biodiversity: Measurement and Estimation - Estimating Terrestrial Biodiversity through Extrapolation.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 345(1311):101–18. Retrieved (http://rstb.royalsocietypublishing.org/content/345/1311/101).

Cooper Jr, William E. and Laurie J. Vitt. 2002. “Distribution, Extent, and Evolution of Plant Consumption by Lizards.” *Journal of Zoology* 257(4):487–517. Retrieved (https://www.cambridge.org/core/article/distribution-extent-and-evolution-of-plant-consumption-by-lizards/14D15142F2E2B67A6FF3DA076938900D).

Corel, Eduardo, Philippe Lopez, Raphaël Méheust, and Eric Bapteste. 2016. “Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution.” *Trends in Microbiology* 24(3):224–37.

Costello, Elizabeth K., Jeffrey I. Gordon, Stephen M. Secor, and Rob Knight. 2010. “Postprandial Remodeling of the Gut Microbiota in Burmese Pythons.” *The ISME Journal* 4(11):1375–85. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3923499/).

David, Lawrence A. et al. 2014. “Diet Rapidly and Reproducibly Alters the Human Gut Microbiome.” *Nature* 505(7484):559–63. Retrieved (http://dx.doi.org/10.1038/nature12820).

Dearing, M.Denise. 1993. “An Alimentary Specialization for Herbivory in the Tropical Whiptail Lizard Cnemidophorus Murinus.” *Journal of Herpetology* 27(1):111–14. Retrieved (http://www.jstor.org/stable/1564920).

Delmont, Tom O., Patrick Robe, Ian Clark, Pascal Simonet, and Timothy M. Vogel. 2011. “Metagenomic Comparison of Direct and Indirect Soil DNA Extraction Approaches.”

Journal of Microbiological Methods 86(3):397–400. Retrieved (http://www.sciencedirect.com/science/article/pii/S0167701211002351).

DeSantis, T. Z. et al. 2006. “Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB.” *Applied and Environmental Microbiology* 72(7):5069–72. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1489311/).

Dewhirst, Floyd E. et al. 2010. “The Human Oral Microbiome.” *Journal of Bacteriology* 192(19):5002–17. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2944498/).

Doolittle, W.Ford. 1999. “Phylogenetic Classification and the Universal Tree.” *Science* 284(5423):2124–28. Retrieved (http://science.sciencemag.org/content/284/5423/2124).

Eberl, G. 2010. “A New Vision of Immunity : Homeostasis of the Superorganism.” *Mucosal Immunology* 3(5):450–60. Retrieved (http://dx.doi.org/10.1038/mi.2010.20).

Edgar, Robert C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics* 26(19):2460–61. Retrieved (+).

Endres, D. M. and J. E. Schindelin. 2003. “A New Metric for Probability Distributions.” *IEEE Transactions on Information Theory* 49(7):1858–60.

Enright, Anton J., Ioannis Iliopoulos, Nikos C. Kyrpides, and Christos A. Ouzounis. 1999. “Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events.” *Nature* 402(6757):86–90. Retrieved (http://dx.doi.org/10.1038/47056).

Escobar-zepeda, Alejandra, Arturo Vera-ponce De León, and Alejandro Sanchez-flores. 2015. “The Road to Metagenomics : From Microbiology to DNA Sequencing Technologies and Bioinformatics.” 6(December):1–15.

Everard, Amandine et al. 2013. “Cross-Talk between Akkermansia Muciniphila and Intestinal Epithelium Controls Diet-Induced Obesity.” *Proceedings of the National Academy of Sciences* 110(22):9066–71. Retrieved (http://www.pnas.org/content/110/22/9066.abstract).

De Filippo, Carlotta et al. 2010. “Impact of Diet in Shaping Gut Microbiota Revealed by a Comparative Study in Children from Europe and Rural Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 107(33):14691–96. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2930426/).

FastQC n.d. “FastQC. [Http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].” Retrieved (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Forster, Dominik et al. 2015. “Testing Ecological Theories with Sequence Similarity Networks: Marine Ciliates Exhibit Similar Geographic Dispersal Patterns as Multicellular Organisms.” *BMC Biology* 13(16):1–16.

Fransen, Floris et al. 2017. “The Impact of Gut Microbiota on Gender-Specific Differences in Immunity.” *Frontiers in Immunology* 8:754. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5491612/).

Fu, Limin, Beifang Niu, Zhengwei Zhu, Sitao Wu, and Weizhong Li. 2012. “CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data.” *Bioinformatics*

28(23):3150–52. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3516142/).

Ghurye, Jay S., Victoria Cepeda-Espinoza, and Mihai Pop. 2016. “Metagenomic Assembly: Overview, Challenges and Applications.” *The Yale Journal of Biology and Medicine* 89(3):353–62. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045144/).

Gill, Steven R. et al. 2006. “Metagenomic Analysis of the Human Distal Gut Microbiome.” *Science* 312(5778):1355–59.

Gomez, Andres et al. 2015. “Gut Microbiome Composition and Metabolomic Profiles of Wild Western Lowland Gorillas (Gorilla Gorilla Gorilla) Reflect Host Ecology.” *Molecular Ecology* 24(10):2551–65.

Gotelli, Nicolas J. and Robert K. Colwell. n.d. “Estimating Species Richness.” in: *Biological Diversity: Frontiers In Measurement And Assessment*. A.E. Magurran and B.J. McGill (eds.). Oxford University Press, Oxford. 345: 39-54

Gromov, M. 1987. “Hyperbolic Groups.” Pp. 75–263 in *Essays in Group Theory*, edited by S. M. Gersten. New York, NY: Springer New York. Retrieved (https://doi.org/10.1007/978-1-4613-9586-7_3).

Guerrero, Ricardo, Lynn Margulis, and Mercedes Berlanga. 2013. “Symbiogenesis: The Holobiont as a Unit of Evolution.” *International Microbiology* 16(3):133–43.

Hamady, Micah, Catherine Lozupone, and Rob Knight. 2010. “Fast UniFrac: Facilitating High-Throughput Phylogenetic Analyses of Microbial Communities Including Analysis of Pyrosequencing and PhyloChip Data.” *The ISME Journal* 4(1):17–27. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2797552/).

Hartman, Kyle, Marcel G. A. van der Heijden, Valexia Roussely-Provent, Jean-Claude Walser, and Klaus Schlaeppi. 2017. “Deciphering Composition and Function of the Root Microbiome of a Legume Plant.” *Microbiome* 5:2. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5240445/).

Heffron, Fred, Brian J. McCarthy, Hisako Ohtsubo, and Eiichi Ohtsubo. 1979. “DNA Sequence Analysis of the Transposon Tn3: Three Genes and Three Sites Involved in Transposition of Tn3.” *Cell* 18(4):1153–63. Retrieved (http://www.sciencedirect.com/science/article/pii/0092867479902289).

Herlemann, Daniel P. R., Oliver Geissinger, and Andreas Brune. 2007. “The Termite Group I Phylum Is Highly Diverse and Widespread in the Environment.” *Applied and Environmental Microbiology* 73(20):6682–85. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2075069/).

Herrel, A., B. Vanhooydonck, and R. Van Damme. 2004. “Omnivory in Lacertid Lizards: Adaptive Evolution or Constraint?” *Journal of Evolutionary Biology* 17(5):974–84. Retrieved (http://dx.doi.org/10.1111/j.1420-9101.2004.00758.x).

Herrel, Anthony. 2007. “Herbivory and Foraging Mode in Lizards.” *Lizard Ecology: The Evolutionary Consequences of Foraging Mode* 209-236.

Herrel, Anthony et al. 2008. “Rapid Large-Scale Evolutionary Divergence in Morphology and Performance Associated with Exploitation of a Different Dietary Resource.” *Proceedings*

of the National Academy of Sciences 105(12):4792–95. Retrieved (http://www.pnas.org/content/105/12/4792.abstract).

Hildebrandt, Marie A. et al. 2009. “High Fat Diet Determines the Composition of the Murine Gut Microbiome Independently of Obesity.” *Gastroenterology* 137(5):1712–16. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2770164/).

HMP Consortium. 2012. “A Framework for Human Microbiome Research.” *Nature* 486. Retrieved (http://dx.doi.org/10.1038/nature11209).

Holscher, Hannah D. 2017. “Dietary Fiber and Prebiotics and the Gastrointestinal Microbiota.” *Gut Microbes* 8(2):172–84. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5390821/).

Hong, Pei-Ying, Emily Wheeler, Isaac K. O. Cann, and Roderick I. Mackie. 2011. “Phylogenetic Analysis of the Fecal Microbial Community in Herbivorous Land and Marine Iguanas of the Galápagos Islands Using 16S rRNA-Based Pyrosequencing.” *The ISME Journal* 5(9):1461–70. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3160690/).

Hu, Yifan. 2005. “Efficient, High-Quality Force-Directed Graph Drawing.” *Mathematica Journal* 10(1):37–71.

Huerta-Cepas, Jaime et al. 2016. “eggNOG 4.5: A Hierarchical Orthology Framework with Improved Functional Annotations for Eukaryotic, Prokaryotic and Viral Sequences.” *Nucleic Acids Research* 44(D1):D286. Retrieved (+).

Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. “MEGAN Analysis of Metagenomic Data.” *Genome Research* 17(3):377–86. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1800929/).

Huson, Daniel H. and David Bryant. 2006. “Application of Phylogenetic Networks in Evolutionary Studies.” *Molecular Biology and Evolution* 23(2):254–67. Retrieved (+).

Huson, Daniel H. and Tobias H. Kloepper. 2005. “Computing Recombination Networks from Binary Sequences.” *Bioinformatics* 21(suppl_2):ii159-ii165. Retrieved (+).

Huttenhower, Curtis and Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature* 486(7402):207–14. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564958&tool=pmcentrez&rendertype=abstract).

Ismail, Wazim Mohammed, Yuzhen Ye, and Haixu Tang. 2014. “Gene Finding in Metatranscriptomic Sequences.” *BMC Bioinformatics* 15(9):S8. Retrieved (https://doi.org/10.1186/1471-2105-15-S9-S8).

Iverson, J. B. 1980. “Anatomical Modifications of the Colon in Lizards of the Subfamily Iguaninae.” *Journal of Morphology* 163(1):79–93.

Jeffery, Ian B., Marcus J. Claesson, Paul W. O’Toole, and Fergus Shanahan. 2012. “Categorization of the Gut Microbiota: Enterotypes or Gradients?” *Nature Reviews Microbiology* 10(9):591–92. Retrieved (http://dx.doi.org/10.1038/nrmicro2859).

Jeffery, Ian B., Marcus J. Claesson, Paul W. O'Toole, and Fergus Shanahan. 2012. "Categorization of the Gut Microbiota: Enterotypes or Gradients?" *Nat Rev Micro* 10(9):591–92. Retrieved (http://dx.doi.org/10.1038/nrmicro2859).

Johnson, M. et al. 2008. "NCBI BLAST: A Better Web Interface." *Nucleic Acids Res* 36: W5–W9. Retrieved (http://dx.doi.org/10.1093/nar/gkn201).

Kageyama, Shinya et al. 2017. "Relative Abundance of Total Subgingival Plaque-Specific Bacteria in Salivary Microbiota Reflects the Overall Periodontal Condition in Patients with Periodontitis" edited by M. Komaki. *PLoS ONE* 12(4):e0174782. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5378373/).

Kanehisa, Minoru, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. "KEGG: New Perspectives on Genomes, Pathways, Diseases and Drugs." *Nucleic Acids Research* 45(D1):D353. Retrieved (+).

Kanehisa, Minoru and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28(1):27–30. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102409/).

Kanehisa, Minoru, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2016. "KEGG as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research* 44(D1):D457. Retrieved (+).

Kaufman, Leonard and Peter Rousseeuw. 1987. *Clustering by Means of Medoids*. North-Holland.

Kent, W. 2002. "BLAT--the BLAST-like Alignment Tool." *Genome Res* 12. Retrieved (http://dx.doi.org/10.1101/gr.229202.).

Kim, Tae Kyung et al. 2009. "Heterogeneity of Vaginal Microbial Communities within Individuals ." *Journal of Clinical Microbiology* 47(4):1181–89. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2668325/).

Kinross, James M., Ara W. Darzi, and Jeremy K. Nicholson. 2011. "Gut Microbiome-Host Interactions in Health and Disease." *Genome Medicine* 3(3):14. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3092099/).

Knights, Dan et al. 2014. "Rethinking 'Enterotypes.'" *Cell Host & Microbe* 16(4):433–37. Retrieved (http://www.sciencedirect.com/science/article/pii/S1931312814003461).

Knights, Dan et al. 2017. "Rethinking "Enterotypes"" *Cell Host & Microbe* 16(4):433–37. Retrieved (http://dx.doi.org/10.1016/j.chom.2014.09.013).

Kohl, Kevin D., Tawnya L. Cary, William H. Karasov, and M.Denise Dearing. 2013. "Restructuring of the Amphibian Gut Microbiota through Metamorphosis." *Environmental Microbiology Reports* 5(6):899–903.

Konstantinidis, Konstantinos T. and James M. Tiedje. 2005. "Genomic Insights That Advance the Species Definition for Prokaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 102(7):2567–72. Retrieved (http://www.pnas.org/content/102/7/2567%5Cnhttp://files/385/Konstantinidis and Tiedje - 2005 - Genomic insights that advance the species

definiti.pdf%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/15701695%5Cnhttp://files/416/2567.html).

Krohs, Ulrich. 2012. “Convenience Experimentation.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43(1):52–57. Retrieved (http://www.sciencedirect.com/science/article/pii/S1369848611000811).

Kruskal, J. 1964. “Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis.” *Psychometrika* 29(1):1–27. Retrieved (https://econpapers.repec.org/RePEc:spr:psycho:v:29:y:1964:i:1:p:1-27).

Kuczynski, Justin, Jesse Stombaugh, William Anton Walters, Antonio González, J.Gregory Caporaso, et al. 2012. “Using QIIME to Analyze 16s rRNA Gene Sequences from Microbial Communities.” *Current Protocols in Microbiology* (SUPPL.27):Chapter 10:Unit 10.7.

Kuczynski, Justin, Jesse Stombaugh, William Anton Walters, Antonio González, J.Gregory Caporaso, et al. 2012. “Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities.” *Current Protocols in Bioinformatics* 36:10.7:10.7.1–10.7.20. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3249058/).

Lab, Hannon. n.d. “FASTX Toolkit.” *Http://hannonlab.cshl.edu/fastx_toolkit/index.html*. Retrieved (http://hannonlab.cshl.edu/fastx_toolkit/).

Lau, Jennifer T. et al. 2016. “Capturing the Diversity of the Human Gut Microbiota through Culture-Enriched Molecular Profiling.” *Genome Medicine* 8:72. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4929786/).

Ley, Ruth E. et al. 2008. “Evolution of Mammals and Their Gut Microbes.” *Science* 320(5883):1647–51. Retrieved (http://science.sciencemag.org/content/320/5883/1647).

Ley, Ruth E., Rob Knight, and Jeffrey I. Gordon. 2007. “The Human Microbiome: Eliminating the Biomedical/environmental Dichotomy in Microbial Ecology.” *Environmental Microbiology* 9(1):3–4. Retrieved (http://dx.doi.org/10.1111/j.1462-2920.2006.01222_3.x).

Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. “MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph.” *Bioinformatics* 31(10):1674–76. Retrieved (+).

Li, Jilian et al. 2015. “Two Gut Community Enterotypes Recur in Diverse Bumblebee Species.” *Current Biology* 25(15):R652–53. Retrieved (http://www.sciencedirect.com/science/article/pii/S0960982215007253).

Li, Jilian et al. 2017. “Two Gut Community Enterotypes Recur in Diverse Bumblebee Species.” *Current Biology* 25(15):R652–53. Retrieved (http://dx.doi.org/10.1016/j.cub.2015.06.031).

Li, Ying et al. 2015. “The Evolution of the Gut Microbiota in the Giant and the Red Pandas.” *Scientific Reports* 5:10185. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4434948/).

Lim, Mi Young et al. 2014. “Stability of Gut Enterotypes in Korean Monozygotic Twins and Their Association with Biomarkers and Diet.” *Scientific Reports* 4:7348. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/25482875%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4258686).

Liu, Jianzheng, Jie Li, Weifeng Li, and Jiansheng Wu. 2016. “Rethinking Big Data: A Review on the Data Quality and Usage Issues.” *ISPRS Journal of Photogrammetry and Remote Sensing* 115:134–42. Retrieved (http://www.sciencedirect.com/science/article/pii/S0924271615002567).

Liu, Lin et al. 2012. “Comparison of Next-Generation Sequencing Systems.” *Journal of Biomedicine and Biotechnology* 2012:251364. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3398667/).

Lloyd-Price, Jason, Galeb Abu-Ali, and Curtis Huttenhower. 2016. “The Healthy Human Microbiome.” *Genome Medicine* 8:51. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4848870/).

Lozupone, C. A., J. I. Stombaugh, J. I. Gordon, J. K. Jansson, and R. Knight. 2012. “Diversity, Stability and Resilience of the Human Gut Microbiota.” *Nature* 489(7415):220–30. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/22972295).

Ma, Bing, Larry J. Forney, and Jacques Ravel. 2012. “Vaginal Microbiome: Rethinking Health and Disease.” *Annual Review of Microbiology* 66:371–89. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/22746335%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3780402).

Margulis, L. and R. Fester. 1991. *Symbiosis as a Source of Evolutionary Innovation: Speciation and Morphogenesis*. MIT Press. Retrieved (citeulike-article-id:11500951).

Martinson, Vincent G. et al. 2011. “A Simple and Distinctive Microbiota Associated with Honey Bees and Bumble Bees.” *Molecular Ecology* 20(3):619–28.

McCann, Joshua C., Tryon A. Wickersham, and Juan J. Loor. 2014. “High-Throughput Methods Redefine the Rumen Microbiome and Its Relationship with Nutrition and Metabolism.” *Bioinformatics and Biology Insights* 8:109–25.

Medina-Colorado, Audrie A. et al. 2017 “Vaginal Ecosystem Modeling of Growth Patterns of Anaerobic Bacteria in Microaerophilic Conditions.” *Anaerobe*. Retrieved (http://www.sciencedirect.com/science/article/pii/S1075996417300811).

Moeller, Andrew H. et al. 2012. “Chimpanzees and Humans Harbor Compositionally Similar Gut Enterotypes.” *Nature Communications* 3:1179. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3520023/).

Moeller, Andrew H. et al. 2015. “Stability of the Gorilla Microbiome despite Simian Immunodeficiency Virus Infection.” *Molecular Ecology* 24(3):690–97.

Mokane Bouzeghoub, Rémy Mosseri. 2017. *Les Big Data À Découvert*. CNRS EDITI. edited by Elsa Godet. Paris.

Morgavi, D. P., W. J. Kelly, P. H. Janssen, and G. T. Attwood. 2013. “Rumen Microbial (Meta)genomics and Its Application to Ruminant Production.” *Animal* 7(s1):184–201.

Mountfort, D. O., R. A. Asher, and T. Bauchop. 1982. “Fermentation of Cellulose to Methane and Carbon Dioxide by a Rumen Anaerobic Fungus in a Triculture with Methanobrevibacter Sp. Strain RA1 and Methanosarcina Barkeri.” *Applied and Environmental Microbiology* 44(1):128–34.

Nagaraj, Veena, Lucy Skillman, Goen Ho, Dan Li, and Alexander Gofton. 2017. “Characterisation and Comparison of Bacterial Communities on Reverse Osmosis Membranes of a Full-Scale Desalination Plant by Bacterial 16S rRNA Gene Metabarcoding.” *NPJ Biofilms and Microbiomes* 3:13. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476683/).

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. “MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads.” *Nucleic Acids Research* 40(20):e155–e155. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3488206/).

NEISH, ANDREW S. 2009. “Microbes in Gastrointestinal Health and Disease.” *Gastroenterology* 136(1):65–80. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2892787/).

Nevo, Eviatar et al. 1972. “Competitive Exclusion between Insular Lacerta Species (Sauria, Lacertidae).” *Oecologia* 10(2):183–90. Retrieved (http://dx.doi.org/10.1007/BF00347990).

Noguchi, Hideki, Takeaki Taniguchi, and Takehiko Itoh. 2008. “Meta Gene Annotator: Detecting Species-Specific Patterns of Ribosomal Binding Site for Precise Gene Prediction in Anonymous Prokaryotic and Phage Genomes.” *DNA Research* 15(6):387–96.

Org, Elin et al. 2016. “Sex Differences and Hormonal Effects on Gut Microbiota Composition in Mice.” *Gut Microbes* 7(4):313–22. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4988450/).

Peng, Y., H. C. M. Leung, S. M. Yiu, and F. Y. L. Chin. 2011. “Meta-IDBA: A de Novo Assembler for Metagenomic Data.” *Bioinformatics* 27. Retrieved (http://dx.doi.org/10.1093/bioinformatics/btr216).

Peng, Yu, Henry C. M. Leung, S. M. Yiu, and Francis Y. L. Chin. 2012. “IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth.” *Bioinformatics* 28(11):1420. Retrieved (+).

Pesant, Stéphane et al. 2015. “Open Science Resources for the Discovery and Analysis of Tara Oceans Data.” *Scientific Data* 2(Lmd):150023. Retrieved (http://www.nature.com/articles/sdata201523).

Piknová, M. et al. 2006. “New Species of Rumen Treponemes.” Pp. 303–5 in *Folia Microbiologica*, vol. 51.

Prado-Irwin, Sofia R., Alicia K. Bird, Andrew G. Zink, and Vance T. Vredenburg. 2017. “Intraspecific Variation in the Skin-Associated Microbiome of a Terrestrial Salamander.” *Microbial Ecology* 1–12. Retrieved (http://dx.doi.org/10.1007/s00248-017-0986-y).

Rho, M., H. Tang, and Y. Ye. 2010. “FragGeneScan: Predicting Genes in Short and Error-Prone Reads.” *Nucleic Acids Res* 38(20):191. Retrieved (http://dx.doi.org/10.1093/nar/gkq747).

Ritchie Kim B. 2006. “Regulation of Microbial Populations by Coral Surface Mucus and Mucus-Associated Bacteria .” *Marine Ecology Progress Series* 322:1–14. Retrieved (http://www.int-res.com/abstracts/meps/v322/p1-14/).

Roberts, Adam P. and Peter Mullany. 2010. “Oral Biofilms: A Reservoir of Transferable, Bacterial, Antimicrobial Resistance.” *Expert Review of Anti-Infective Therapy* 8(12):1441–50. Retrieved (http://dx.doi.org/10.1586/eri.10.106).

Robles Alonso, V. and F. Guarner. 2013. “Linking the Gut Microbiota to Human Health.” *Br J Nutr.* 109. Retrieved (https://doi.org/10.1017/S0007114512005235).

Rosenthal, Mariana, Deborah Goldberg, Allison Aiello, Elaine Larson, and Betsy Foxman. 2011. “Skin Microbiota: Microbial Community Structure and Its Potential Association with Health and Disease.” *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 11(5):839–48. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3114449/).

Santana, Renata Henrique et al. 2015. “The Gut Microbiota of Workers of the Litter-Feeding Termite Syntermes Wheeleri (Termitidae: Syntermitinae): Archaeal, Bacterial, and Fungal Communities.” *Microbial Ecology* 70(2):545–56. Retrieved (https://doi.org/10.1007/s00248-015-0581-z).

Schloss, P. D. et al. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Appl Environ Microbiol* 75(23):7537-41. Retrieved (http://dx.doi.org/10.1128/AEM.01541-09).

Schueller, Katharina, Alessandra Riva, Stefanie Pfeiffer, David Berry, and Veronika Somoza. 2017. “Members of the Oral Microbiota Are Associated with IL-8 Release by Gingival Epithelial Cells in Healthy Individuals.” *Frontiers in Microbiology* 8:416. Retrieved (http://journal.frontiersin.org/article/10.3389/fmicb.2017.00416).

Segata, Nicola et al. 2011. “Metagenomic Biomarker Discovery and Explanation.” *Genome Biology* 12(6):R60–R60. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218848/).

Selosse, Marc André, Alain Bessis, and María J. Pozo. 2014. “Microbial Priming of Plant and Animal Immunity: Symbionts as Developmental Signals.” *Trends in Microbiology* 22(11):607–13.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. “Revised Estimates for the Number of Human and Bacteria Cells in the Body.” *PLOS Biology* 14(8):e1002533. Retrieved (https://doi.org/10.1371/journal.pbio.1002533).

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27(3):379–423. Retrieved (http://dx.doi.org/10.1002/j.1538-

7305.1948.tb01338.x).

Simpson, E. H. 1949. “Measurement of Diversity.” *Nature* 163(688):688-688.

Sneath, P. H. and R. R. Sokal. 1962. “Numerical Taxonomy.” *Nature* 193:855–60.

Solden, Lindsey M. et al. 2017. “New Roles in Hemicellulosic Sugar Fermentation for the Uncultivated Bacteroidetes Family BS11.” *The ISME Journal* 11(3):691–703. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5322302/).

Sonnenburg, Erica D. et al. 2016. “Diet-Induced Extinctions in the Gut Microbiota Compound over Generations.” *Nature* 529(7585):212–15. Retrieved (http://dx.doi.org/10.1038/nature16504).

Su, LiJuan et al. 2016. “Comparative Gut Microbiomes of Four Species Representing the Higher and the Lower Termites.” *Journal of Insect Science* 16(1):97. Retrieved (http://jinsectscience.oxfordjournals.org/lookup/doi/10.1093/jisesa/iew081).

Suen, Garret et al. 2010. “An Insect Herbivore Microbiome with High Plant Biomass-Degrading Capacity.” *PLOS Genetics* 6(9):1–14. Retrieved (https://doi.org/10.1371/journal.pgen.1001129).

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d\textquoterightOvidio, et al. 2015. “Structure and Function of the Global Ocean Microbiome” edited by E. Boss et al. *Science* 348(6237): 1261359
. Retrieved (http://science.sciencemag.org/content/348/6237/1261359).

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco D’Ovidio, et al. 2015. “Structure and Function of the Global Ocean Microbiome.” *Science* 348(6237):1261359. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/25999513).

Sydow, Jörg, Georg Schreyögg, and Jochen Koch. 2005. “Organizational Paths: Path Dependency and Beyond.”

Tang, Zheng-Zheng, Guanhua Chen, and Alexander V Alekseyenko. 2016. “PERMANOVA-S: Association Test for Microbial Community Composition That Accommodates Confounders and Multiple Distances.” *Bioinformatics* 32(17):2618–25. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5013911/).

Tarca, Adi L., Vincent J. Carey, Xue-wen Chen, Roberto Romero, and Sorin Drăghici. 2007. “Machine Learning and Its Applications to Biology” edited by F. Lewitter. *PLoS Computational Biology* 3(6):e116. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1904382/).

Tatusov, R. L., E. V Koonin, and D. J. Lipman. 1997. “A Genomic Perspective on Protein Families.” *Science* 278(5338):631–37.

Tatusov, Roman L. et al. 2003. “The COG Database: An Updated Version Includes

Eukaryotes." *BMC Bioinformatics* 4:41. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC222959/).

Tatusov, Roman L., Michael Y. Galperin, Darren A. Natale, and Eugene V Koonin. 2000. "The COG Database: A Tool for Genome-Scale Analysis of Protein Functions and Evolution." *Nucleic Acids Research* 28(1):33–36. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC102395/).

Thomas, Torsten, Jack Gilbert, and Folker Meyer. 2012. "Metagenomics - a Guide from Sampling to Data Analysis." *Microbial Informatics and Experimentation* 2:3. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3351745/).

Thursby, Elizabeth and Nathalie Juge. 2017. "Introduction to the Human Gut Microbiota." *Biochemical Journal* 474(11):1823–36. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5433529/).

Treangen, Todd J. et al. 2011. "MetAMOS: A Metagenomic Assembly and Analysis Pipeline for AMOS." *Genome Biology* 12(1):P25. Retrieved (http://dx.doi.org/10.1186/gb-2011-12-s1-p25).

Treangen, Todd J. et al. 2013. "MetAMOS: A Modular and Open Source Metagenomic Assembly and Analysis Pipeline." *Genome Biology* 14(1):R2–R2. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4053804/).

Turnbaugh, P. J. et al. 2007. "The Human Microbiome Project." *Nature* 449:804-810. Retrieved (http://dx.doi.org/10.1038/nature06244).

Turnbaugh, Peter J., Micah Hamady, et al. 2009. "A Core Gut Microbiome in Obese and Lean Twins." *Nature* 457(7228):480–84. Retrieved (http://dx.doi.org/10.1038/nature07540).

Turnbaugh, Peter J., Vanessa K. Ridaura, et al. 2009. "The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice." *Science Translational Medicine* 1(6):6-14. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2894525&tool=pmcentrez&rendertype=abstract).

Turnbaugh, Peter J., Fredrik Backhed, Lucinda Fulton, and Jeffrey I. Gordon. 2008. "Marked Alterations in the Distal Gut Microbiome Linked to Diet-Induced Obesity." *Cell Host & Microbe* 3(4):213–23. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3687783/).

Turner, Thomas R., Euan K. James, and Philip S. Poole. 2013. "The Plant Microbiome." *Genome Biology* 14(6):209. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3706808/).

Ursell, Luke K., Jessica L. Metcalf, Laura Wegener Parfrey, and Rob Knight. 2012. "Defining the Human Microbiome." *Nutrition Reviews* 70(Suppl 1):S38–44. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426293/).

Vandenkoornhuyse, Philippe, Achim Quaiser, Marie Duhamel, Amandine Le Van, and Alexis Dufresne. 2015. "The Importance of the Microbiome of the Plant Holobiont." *New*

Phytologist 206(4):1196–1206. Retrieved (http://dx.doi.org/10.1111/nph.13312).

de Vargas, Colomban et al. 2015. “Eukaryotic Plankton Diversity in the Sunlit Ocean.” *Science* 348(6237):1261605–1261605. Retrieved (http://www.sciencemag.org.proxy.libraries.rutgers.edu/content/348/6237/1261605.full%5Cnhttp://www.sciencemag.org/cgi/doi/10.1126/science.1261605).

Venter, J. C. et al. 2004. “Environmental Genome Shotgun Sequencing of the Sargasso Sea.” *Science* 304 (5667):66-74. Retrieved (http://dx.doi.org/10.1126/science.1093857).

Vigliotti, Chloe, Cedric Bicep, Eric Bapteste, Philippe Lopez, and Eduardo Corel. 2017. “TRACKING THE RULES OF TRANSMISSION AND INTROGRESSION WITH NETWORKS.” in *« Experimental and theoretical modes of transmission ».* edited by F. Baquero & T. Coque. Paris.

Völkel, Finn, Eric Bapteste, Michel Habib, Philippe Lopez, and Chloe Vigliotti. 2016. “Read Networks and K-Laminar Graphs.” *arXiv* 1–14. Retrieved (http://arxiv.org/abs/1603.01179).

Walter, Jens and Ruth Ley. 2011. “The Human Gut Microbiome: Ecology and Recent Evolutionary Changes.” *Annual Review of Microbiology* 65(1):411–29.

Walters, William et al. 2016. “Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys” edited by H. Bik. *mSystems* 1(1):e00009-15. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5069754/).

Watanabe, Hidemi and Jinya Otsuka. 1995. “A Comprehensive Representation of Extensive Similarity Linkage between Large Numbers of Proteins.” *Bioinformatics* 11(2):159–66. Retrieved (+).

Watson, Andrew K. et al. 2017. “The Methodology Behind Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution.” in *Evolutionary genomics: statistical and computational methods*, edited by A. Ed.

Watson, L., W. T. Williams, and G. N. Lance. 1966. “Angiosperm Taxonomy: A Comparative Study of Some Novel Numerical Techniques.” *Journal of the Linnean Society of London, Botany* 59(380):491–501. Retrieved (http://dx.doi.org/10.1111/j.1095-8339.1966.tb00075.x).

Wei, Fuwen, Yibo Hu, et al. 2015. “Giant Pandas Are Not an Evolutionary Cul-de-Sac: Evidence from Multidisciplinary Research.” *Molecular Biology and Evolution* 32(1):4–12. Retrieved (+).

Wei, Fuwen, Xiao Wang, and Qi Wu. 2015. “The Giant Panda Gut Microbiome.” *Trends in Microbiology* 23(8):450–52.

Wei, Ya Qin et al. 2016. “Fiber Degradation Potential of Natural Co-Cultures of Neocallimastix Frontalis and Methanobrevibacter Ruminantium Isolated from Yaks (Bos Grunniens) Grazing on the Qinghai Tibetan Plateau.” *Anaerobe* 39:158–64.

Whittaker, R. H. 1960. “Vegetation of the Siskiyou Mountains, Oregon and California.” *Ecological Monographs* 30(3):279–338. Retrieved (http://www.jstor.org/stable/1943563).

Whittaker, R. H. 1972. “Evolution and Measurement of Species Diversity.” *Taxon* 21(2/3):213–51. Retrieved (http://www.jstor.org/stable/1218190).

Wiener, Norbert. 1948. *Cybernetics; or Control and Communication in the Animal and the Machine*. MIT Press. edited by J. Wiley. Oxford, UK.

Wilke, Andreas et al. 2016. “The MG-RAST Metagenomics Database and Portal in 2015.” *Nucleic Acids Research* 44(Database issue):D590–94. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702923/).

van den Wollenberg, Arnold L. 1977. “Redundancy Analysis an Alternative for Canonical Correlation Analysis.” *Psychometrika* 42(2):207–19. Retrieved (https://doi.org/10.1007/BF02294050).

Wu, Gary D. et al. 2011. “Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes.” *Science* 334(6052):105 LP-108. Retrieved (http://science.sciencemag.org/content/334/6052/105.abstract).

Wu, Hsin-Jung and Eric Wu. 2012. “The Role of Gut Microbiota in Immune Homeostasis and Autoimmunity.” *Gut Microbes* 3(1):4–14. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3337124/).

Wu, Wen Ming, Yun Sheng Yang, and Li Hua Peng. 2014. “Microbiota in the Stomach: New Insights.” *Journal of Digestive Diseases* 15(2):54–61. Retrieved (http://dx.doi.org/10.1111/1751-2980.12116).

Xu, Jinyu et al. 2016. “The Impact of Dietary Energy Intake Early in Life on the Colonic Microbiota of Adult Mice.” *Scientific Reports* 6(November 2015):19083. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4705468&tool=pmcentrez&rendertype=abstract).

Xue, Zhengsheng et al. 2015. “The Bamboo-Eating Giant Panda Harbors a Carnivore-Like Gut Microbiota, with Excessive Seasonal Variations” edited by J. Zhou. *mBio* 6(3):e00022-15. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4442137/).

Yáñez-Ruiz, David R., Leticia Abecia, and Charles J. Newbold. 2015. “Manipulating Rumen Microbiome and Fermentation through Interventions during Early Life: A Review.” *Frontiers in Microbiology* 6(OCT):1133.

Yang, Xin et al. 2017. “The Normal Vaginal and Uterine Bacterial Microbiome in Giant Pandas (*Ailuropoda Melanoleuca*).” *Microbiological Research* 199(Supplement C):1–9. Retrieved (http://www.sciencedirect.com/science/article/pii/S0944501316305845).

Yarza, Pablo et al. 2014. “Uniting the Classification of Cultured and Uncultured Bacteria and Archaea Using 16S rRNA Gene Sequences.” *Nat Rev Micro* 12(9):635–45. Retrieved (http://dx.doi.org/10.1038/nrmicro3330).

Yatsunenko, T. et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486(7402):222–27. Retrieved (http://www.ncbi.nlm.nih.gov/pubmed/22699611%5Cnhttp://www.nature.com/nature/journal/v486/n7402/pdf/nature11053.pdf).

Ye, Yuzhen. 2011. “Identification and Quantification of Abundant Species from

Pyrosequences of 16S rRNA by Consensus Alignment." *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine* 2010:153–57. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3217275/).

Yeoman, Carl J. et al. 2011. "Towards an Evolutionary Model of Animal-Associated Microbiomes." *Entropy* 13(3): 570-594.

Zeng, Bo et al. 2015. "The Bacterial Communities Associated with Fecal Types and Body Weight of Rex Rabbits." *Scientific Reports* 5:9342. Retrieved (http://www.nature.com/srep/2015/150320/srep09342/full/srep09342.html).

Zerbino, D. R. 2010. "Using the Velvet de Novo Assembler for Short-Read Sequencing Technologies."

Zerbino, D. R. and E. Birney. 2008. "Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs." *Genome Res* 18. Retrieved (http://dx.doi.org/10.1101/gr.074492.107).

Zhang, Chenhong et al. 2012. "Structural Resilience of the Gut Microbiota in Adult Mice under High-Fat Dietary Perturbations." *The ISME Journal* 6(10):1848–57. Retrieved (http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3446802/).

Zhaxybayeva, Olga and W.Ford Doolittle. 2011. "Lateral Gene Transfer." *Current Biology* 21(7):R242–46. Retrieved (http://www.sciencedirect.com/science/article/pii/S0960982211001011).

Zheng, Jia et al. 2016. "The Programming Effects of Nutrition-Induced Catch-up Growth on Gut Microbiota and Metabolic Diseases in Adult Mice." *MicrobiologyOpen* 5(2):296–306.

Zhu, Lifeng, Qi Wu, Jiayin Dai, Shanning Zhang, and Fuwen Wei. 2011. "Evidence of Cellulose Metabolism by the Giant Panda Gut Microbiome." *Proceedings of the National Academy of Sciences of the United States of America* 108(43):17714–19. Retrieved (http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3203778&tool=pmcentrez&rendertype=abstract).

Zhu, W., A. Lomsadze, and M. Borodovsky. 2010. "Ab Initio Gene Identification in Metagenomic Sequences." *Nucleic Acids Research* 38:132. Retrieved (http://dx.doi.org/10.1093/nar/gkq275).

Zilber-Rosenberg, Ilana and Eugene Rosenberg. 2008. "Role of Microorganisms in the Evolution of Animals and Plants: The Hologenome Theory of Evolution." *FEMS Microbiology Reviews* 32(5):723–35. Retrieved (+).

Zilber-Rosenberg, Ilana and Eugene Rosenberg. 2013. "The Hologenome Concept: Human, Animal and Plant Microbiota." *The Quarterly Review of Biology* 90(2):230. Retrieved (https://doi.org/10.1086/681496).

Annexes

Table SI-1: description of the data set

The first column contains the identification of the sample (*Podarcis sicula*). The second column is the year of sampling (2014, 2015, 2016), the third is the season of sampling (spring or summer). The fourth is the geographic origin of the sample (the islands: Pod Kopište and Pod Mrčaru; the continent: Split and Zagreb). The fifth column is the sex of the lizard, the sixth is the diet (insectivorous or omnivorous), the seventh is the existence of an ARN 16S sample for these lizards, the eighth column is the presence of a microbiome for these lizards. The ninth column is the 16S sequencing method used, the tenth is the shotgun sequencing method. The eleventh column is the RNAse zap. The twelfth column is the method of conservation, the thirteenth column is the number of ARN 16S reads from the sample, and the last column is the number of 16S reads after filtering.

Lizard ID	Year	Season	Geographic origin	Sex	Diet	16S	16S sequencing method	RNAse zap	method of conservation	Number of 16S reads	Number of 16S reads after filtering
PSK34MDI	2014	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	82292	71261
PSK35MDI	2014	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	115802	100593
PSK36MDI	2014	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	76087	64805
PSKF20MDI	2014	summer	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	75555	65613
PSKF21MDI	2014	summer	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	327138	286173
PSKF23MDI	2014	summer	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	109680	95720
PSM30MDI	2014	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	82098	71194
PSM31MDI	2014	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	63917	55741
PSM32MDI	2014	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	76577	66727
PSM33DIC	2014	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	no	frozen in LN2 and kept in the -80C freezer	49535	43144
PSMF18MDI	2014	summer	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	70557	61513
PSMF19MDI	2014	summer	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	204118	178897
PSMF20MDI	2014	summer	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	72357	62193
PSS10MDI	2014	spring	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	64236	55907
PSS20MDI	2014	summer	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	73081	63846
PSS21MDI	2014	summer	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	45590	39756
PSS22MDI	2014	summer	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	80597	70077
PSSF10MDI	2014	summer	continent (split)	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	67905	59242
PSSF8MDI	2014	summer	continent (split)	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	73510	64041
PSSF9MDI	2014	summer	continent (split)	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	73795	64422
PSZ21MDI	2014	spring	Continent (Zagreb)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	76965	67121
PSZ23MDI	2014	spring	Continent (Zagreb)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	86334	76463
PSZ31MDI	2014	summer	Continent (Zagreb)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	97452	85774
PSZ32MDI	2014	summer	Continent (Zagreb)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	67486	58477
PSZ33MDI	2014	summer	Continent (Zagreb)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	84270	74294
PSK43DIGC	2015	spring	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	no	frozen in LN2 and kept in the -80C freezer	75144	65960
PSKF30DIGC	2015	spring	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	no	frozen in LN2 and kept in the -80C freezer	85123	75678
PSM41DIGC	2015	spring	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	no	frozen in LN2 and kept in the -80C freezer	77076	67178
PSMF27DIGC	2015	spring	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	77598	68332
PSS30MDI	2015	spring	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	68292	60369
PSS31MDI	2015	spring	continent (split)	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	frozen in LN2 and kept in the -80C freezer	63639	55258

Lizard ID	Year	Season	eographic orig	Sex	Diet	16S	16S sequencing method	RNAse zap	hod of conserv	nber of 16S re	Number of 16S reads after filtering
PSS32MDI	2015	spring	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	61548	52855
PSS33MDI	2015	spring	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	66899	58175
PSSF18MDI	2015	spring	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	57138	47406
PSSF19MDI	2015	spring	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	63156	55486
PSSF20MDI	2015	spring	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	74531	65217
PSSF21MDI	2015	spring	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	78282	68820
PSZ41MDI	2015	spring	ontinent (Zagre	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	78694	67876
PSZ42MDI	2015	spring	ontinent (Zagre	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	75449	66126
PSZ43MDI	2015	spring	ontinent (Zagre	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	87348	76299
PSZ44MDI	2015	spring	ontinent (Zagre	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	99792	87787
PSZF15MDI	2015	spring	ontinent (Zagre	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	65065	57116
PSZF16MDI	2015	spring	ontinent (Zagre	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	56587	49712
PSZF17MDI	2015	spring	ontinent (Zagre	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	58228	51429
PSK47DIGC	2016	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	118899	102668
PSK48DIGC	2016	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	116828	101854
PSK49DIGC	2016	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	139566	121308
PSK50DIGC	2016	summer	Pod Kopiste	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	124834	107135
PSKF36DIGC	2016	summer	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	120588	104785
PSKF37DIGC	2016	summer	Pod Kopiste	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	128650	111401
PSM50DIGC	2016	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	85818	76067
PSM51DIGC	2016	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	82953	73191
PSM52DIGC	2016	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	69428	61125
PSM53DIGC	2016	summer	Pod Mcraru	M	Omnivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	66773	58628
PSMF32DIGC	2016	summer	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	115462	100215
PSMF33DIGC	2016	summer	Pod Mcraru	F	Omnivore	oui	2x250bp PE illumina 150-300	no	and kept in the	112451	96985
PSS34DIGC	2016	summer	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	90556	78958
PSS35DIGC	2016	summer	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	86487	75377
PSS36DIGC	2016	summer	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	111591	96913
PSS37DIGC	2016	summer	continent (split	M	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	118094	102342
PSSF22DIGC	2016	summer	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	38320	33520
PSSF23DIGC	2016	summer	continent (split	F	Insectivore	oui	2x250bp PE illumina 150-300	yes	and kept in the	99336	86898

Table SI-2: Number of OTUs per lizard

The first column is the lizard ID and the second column is the number of OTUs per lizard.

Lizard ID	Number of OTUs
PSMF27DIGC	4631
PSKF20MDI	3687
PSS37DIGC	5341
PSZ41MDI	3759
PSSF18MDI	2938
PSS32MDI	3270
PSM52DIGC	3775
PSS20MDI	3865
PSMF33DIGC	5016
PSSF8MDI	3757
PSMF32DIGC	5107
PSS21MDI	2378
PSSF21MDI	4215
PSS10MDI	2738
PSZF15MDI	2210
PSZ23MDI	2529
PSS34DIGC	3467
PSM50DIGC	1364
PSSF20MDI	3489
PSZ44MDI	3676
PSK34MDI	4287
PSKF30DIGC	3050
PSKF23MDI	4952
PSM53DIGC	3846
PSK43DIGC	4363
PSM41DIGC	4628
PSMF18MDI	4411
PSK35MDI	3780
PSM51DIGC	3171
PSS31MDI	2726
PSSF10MDI	2980
PSM31MDI	4126
PSK36MDI	3795
PSKF21MDI	9012
PSS22MDI	3824
PSK49DIGC	3967
PSZ42MDI	3388
PSZ21MDI	3874
PSK48DIGC	4625
PSZ31MDI	3721
PSMF19MDI	7415
PSS33MDI	3825
PSZ33MDI	2826
PSSF9MDI	3457
PSS30MDI	2327
PSM32MDI	4311
PSM33DIC	3297
PSKF37DIGC	5234
PSS35DIGC	4411
PSKF36DIGC	4365
PSZ43MDI	3446
PSZF17MDI	2869
PSK47DIGC	5510
PSMF20MDI	4242
PSK50DIGC	4850
PSSF22DIGC	2501
PSM30MDI	4987
PSZF16MDI	2569
PSSF23DIGC	3759
PSZ32MDI	3359
PSS36DIGC	3462
PSSF19MDI	2800

Figure SI-3: Saturation curves

This curve is obtained by multiple rarefaction on alpha diversity. It has been performed on QIIME. As input, we give the OTU abundance table. The chosen upper limit of rarefaction depth is 100. On the x-axis is represented the number of sequences per sample and on the y-axis is represented the Chao1 index, which is the measure used for the rarefaction.

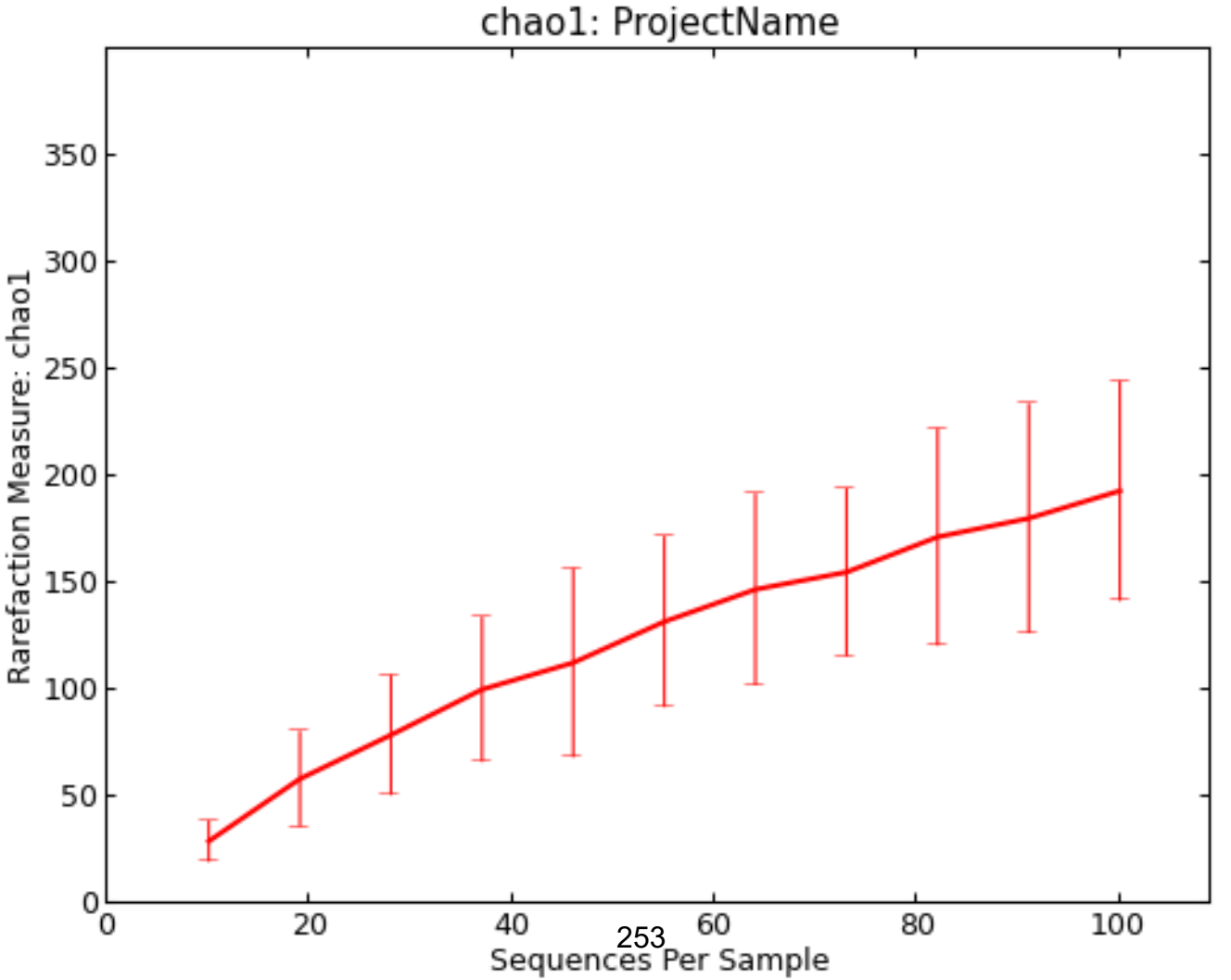

chao1: ProjectName
Rarefaction Measure: chao1
Sequences Per Sample
0
50
100
150
200
250
300
350
0
20
40
60
80
100

Table SI-4: Results of alpha diversity with Shannon, Simpson, and Chao1 indices.

The first column contains the variables tested. The second column contains the index chosen (Shannon, Simpson, or Chao1) and the group of lizards whose index is calculated. The third column is the result of the alpha diversity analyses pooling all samples from a group (for example all insular lizards together and all continental lizards together). The fourth column is the average alpha diversity within a group, the fifth column is the standard deviation of the alpha diversity within a group. The last column is the significance of the result (*P*-value from a Mann-Whitney U test.

Test	Type of alpha diversity (Lizards category)	Samples pooled	mean	standard deviation	significativity
Insularity	**apha diversité Shannon (PK+PM)**	5.06506914372	**7.145225**	**+/- 1.334243**	*
	apha diversité Shannon(C)	5.30530339508	**6.762955**	**+/- 0.8159249**	**P-value= 0,006**
	apha diversité Simpson (PK+PM)	0.988915617408	0.9368664	0.08867474	
	apha diversité Simpson (C)	0.985813689844	0.9427192	0.04337025	P-value= 0.06
	apha diversité Chao1 (PK+PM)	31513.4145316	**9157.002**	**2168.897**	**
	apha diversité Chao1(C)	28360.3012	**7302.994**	**1493.371**	**P-value = 3.206e-05**
Diet	**alpha diversite Shannon (PM)**	4.99461678936	**7.38132**	**1.463731**	*
	alpha diversite Shannon (C+PK)	5.23316489939	**6.801474**	**0.9276718**	**P-value = 0.001**
	alpha diversite Simpson (PM)	0.986024011498	**0.944925**	**0.09336283**	*
	alpha diversite Simpson (C+PK)	0.98825297453	**0.9384039**	**0.05867629**	**P-value = 0.03**
	alpha diversite Chao1 (PM)	27026.3545611	**9048.728**	**2440.345**	*
	alpha diversite Chao1 (C+PK)	31646.2	**7889.808**	**1850.329**	**P-value = 0.01**
Insularity within insectivorous	alpha diversite Shannon (PK)	4.7998429788	6.892266	1.180567	
	alpha diversite Shannon (C)	5.30530339508	6.762955	0.8159249	Pvalue 0.4
	alpha diversite Simpson (PK)	0.98456253895	0.9282321	0.0859872	
	alpha diversite Simpson (C)	0.985813689844	0.9427192	0.04337025	Pvalue 0.6
	alpha diversite Chao1 (PK)	28824.7385621	**9273.01**	**1921.076**	*
	alpha diversite Chao1 (C)	28360.3012	**7302.994**	**1493.371**	**P-value = 0.0006**
Diet within islands	alpha diversite Shannon (PK)	4.7998429788	6.892266	1.180567	
	alpha diversite Shannon (PM)	4.99461678936	7.38132	1.463731	Pvalue 0.06
	alpha diversite Simpson (PK)	0.98456253895	0.9282321	0.0859872	
	alpha diversite Simpson (PM)	0.986024011498	0.944925	0.09336283	P-value = 0.3
	alpha diversite Chao1 (PK)	28824.7385621	9273.01	1921.076	
	alpha diversite Chao1 (PM)	27026.3545611	9048.728	2440.345	P-value = 0.6

Test	Type of alpha diversity (Lizards category)	Samples pooled	mean	standard deviation	significativity
Gender	alpha diversite Shannon (Male)	5.30544839576	6.876123	1.192499	
	alpha diversite Shannon (Female)	5.18813198744	7.045682	0.9409948	P-value = 0.7
	alpha diversite Simpson (Male)	0.989094644526	0.9345026	0.0777904	
	alpha diversite Simpson (Female)	0.989114475208	0.9486567	0.04847373	P-value = 0.8
	alpha diversite Chao1 (Male)	32577.0695569	7884.6	1748.152	
	alpha diversite Chao1 (Female)	31687.3172182	8622.378	2423.232	P-value = 0.3
'ear of samplin	alpha diversite Shannon (2014)	5.10005004798	7.182537	0.8942495	
	alpha diversite Shannon (2015)	5.36307571082	6.891238	1.042066	
	alpha diversite Shannon (2016)	5.00324948572	6.660673	1.366536	P-value = 0.3
	alpha diversite Simpson (2014)	0.988219275281	0.9557946	0.04332116	
	alpha diversite Simpson (2015)	0.98640458914	0.9427418	0.05467383	
	alpha diversite Simpson (2016)	0.985420733131	0.9151056	0.09837137	P-value = 0.3
	alpha diversite Chao1 (2014)	30291.0270565	8692.172	2173.04	
	alpha diversite Chao1 (2015)	28924.698324	7349.97	1673.362	
	alpha diversite Chao1 (2016)	27351.8038564	8311.008	2064.45	P-value = 0.07
Season	alpha diversite Shannon (spring)	5.40880688996	6.822995	1.022443	
	alpha diversite Shannon (summer)	5.18257205513	7.007079	1.143214	P-value = 0.3
	alpha diversite Simpson (spring)	0.986834754404	0.9422183	0.05147338	
	alpha diversite Simpson (summer)	0.990269411482	0.9387514	0.07596889	P-value = 0.6
	alpha diversite Chao1 (spring)	29031.7534497	7245.264	1714.209	
	alpha diversite Chao1 (summer)	32409.8311911	8678.902	2058.48	P-value = 0.004

Table - SI5: Results of the RDA

The left column describes the variables that the model takes into account. At the right is indicated the explanatory power of the model.

Host characteristics in the model	R^2 (explanatory power of the model)
Diet + sex + geography + insularity + year + season	17.1%
Sex	0 %
Season (of sampling)	0.984 %
Diet	4.53%
Year (of sampling)	6.74%
Insularity	7.86%
Geography	9.57 %

Figure - SI6: Results of the linear discriminant analysis at the phylum level

a) LDA to distinguish omnivorous from insectivorous lizards. In green are represented phyla which positively discriminate omnivorous lizards from insectivorous ones based on their gut microbial composition.

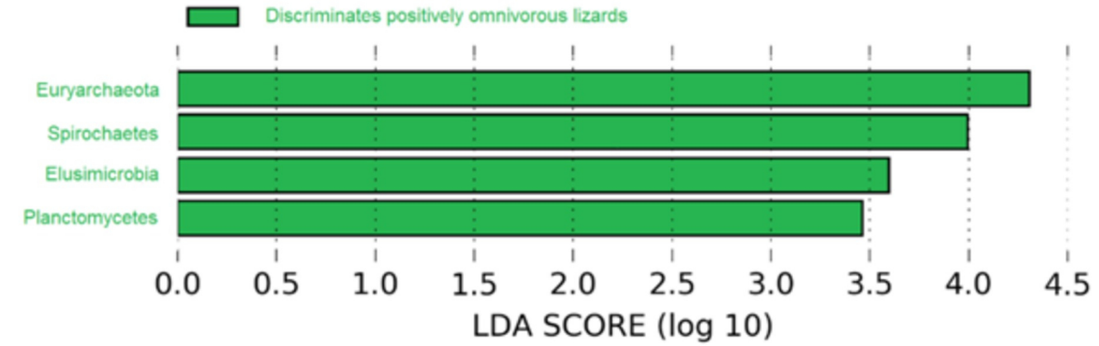
b) LDA to distinguish lizards sampled in spring from lizards sampled in summer. In green are represented phyla which discriminate positively summer lizards from spring ones based on their gut microbial composition. In yellow are represented phyla which negatively discriminate spring lizards from summer ones.

c) LDA to distinguish continental lizards from insular lizards from Pod Kopište and insular lizards from Pod Mrčaru. In green are represented phyla which discriminate positively lizards from Pod Mrčaru, in brown are represented phyla which discriminate positively lizards Pod Kopište, and in pink, phyla which discriminate positively lizards from the continent based on their gut microbial composition.

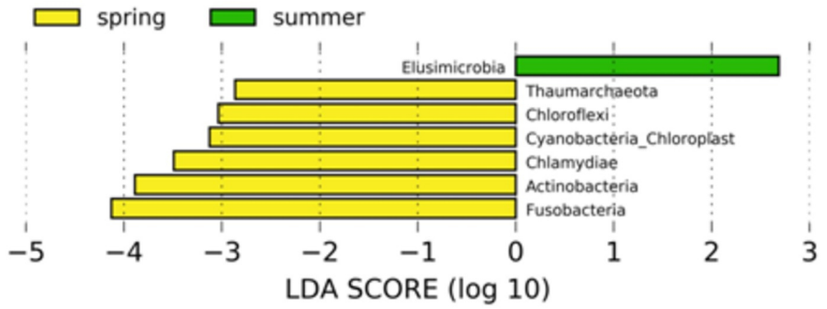
d) LDA performed to distinguish lizards sampled in 2014 from lizards sampled in 2015 and 2016. In red are represented phyla which discriminate positively lizards sampled in 2016, in blue are represented phyla which discriminate positively lizards sampled in 2015, in purple, are represented phyla which discriminate positively lizards sampled in 2016 based on their gut microbial composition.

e) LDA performed to distinguish continental lizards from insular lizards. In orange are represented phyla which discriminate positively insular lizards from continental ones based on their gut microbial composition. In pink are represented phyla which discriminate negatively continental lizards from insular ones.

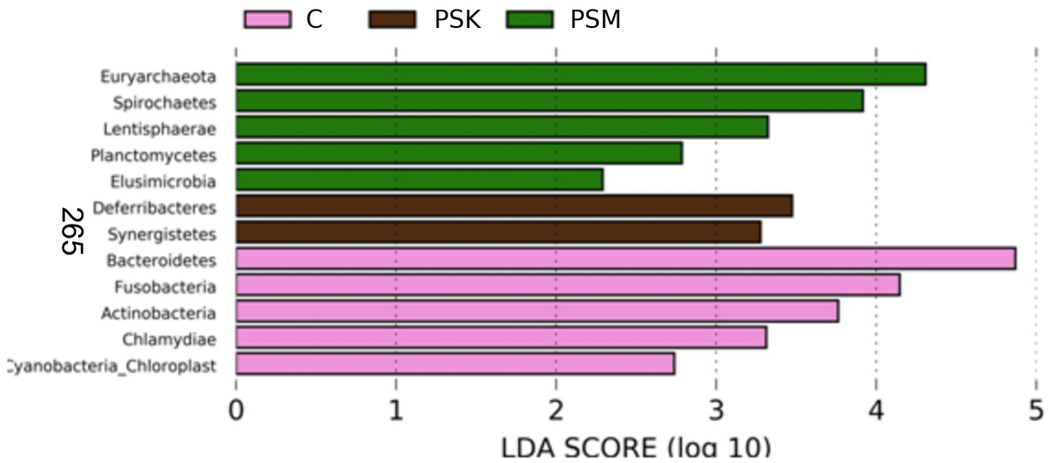
f) LDA performed to distinguish continental lizards from insular and insectivorous lizards. In brown are represented phyla which discriminate positively insular lizards from continental ones based on their gut microbial composition. In pink are represented phyla which discriminate negatively continental lizards from insular ones


Discriminates positively omnivorous lizards
Euryarchaeota
Spirochaetes
Elusimicrobia
Planctomycetes
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5
LDA SCORE (log 10)

a)

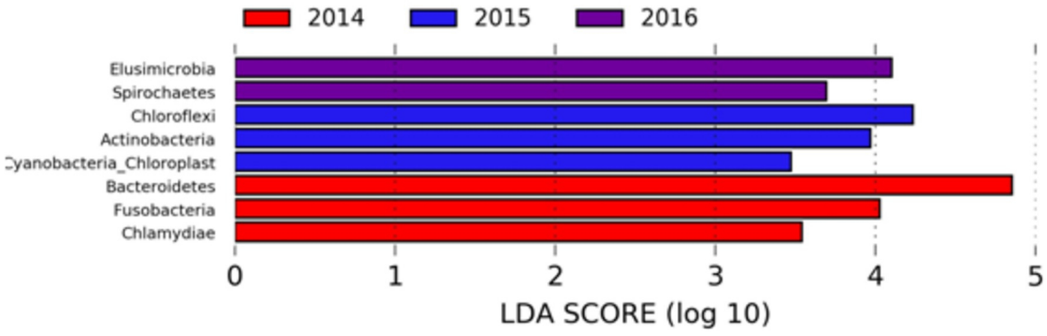

spring
summer
Elusimicrobia
Thaumarchaeota
Chloroflexi
Cyanobacteria_Chloroplast
Chlamydiae
Actinobacteria
Fusobacteria
−5 −4 −3 −2 −1 0 1 2 3
LDA SCORE (log 10)

b)

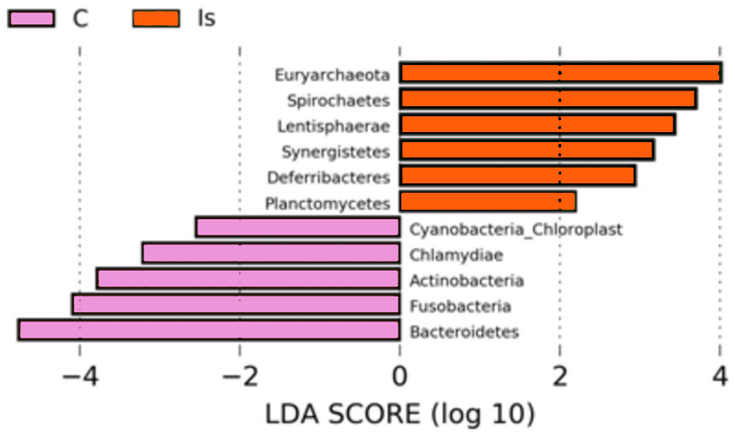

C
PSK
PSM
Euryarchaeota
Spirochaetes
Lentisphaerae
Planctomycetes
Elusimicrobia
Deferribacteres
Synergistetes
Bacteroidetes
Fusobacteria
Actinobacteria
Chlamydiae
Cyanobacteria_Chloroplast
0 1 2 3 4 5
LDA SCORE (log 10)

265

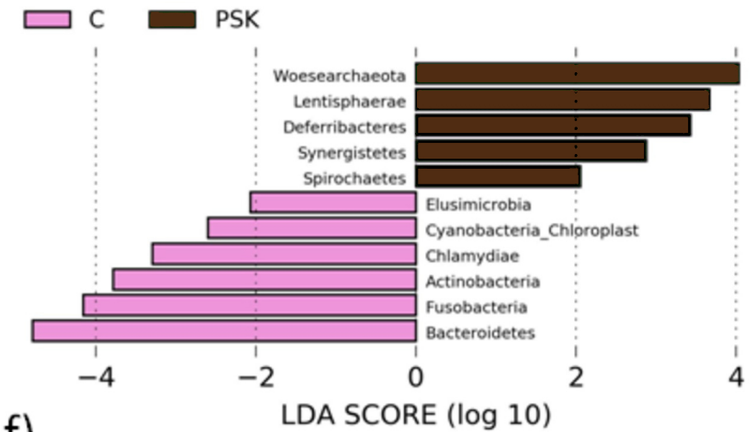
c)


2014
2015
2016
Elusimicrobia
Spirochaetes
Chloroflexi
Actinobacteria
Cyanobacteria_Chloroplast
Bacteroidetes
Fusobacteria
Chlamydiae
0 1 2 3 4 5
LDA SCORE (log 10)

d)


C
Is
Euryarchaeota
Spirochaetes
Lentisphaerae
Synergistetes
Deferribacteres
Planctomycetes
Cyanobacteria_Chloroplast
Chlamydiae
Actinobacteria
Fusobacteria
Bacteroidetes
−4 −2 0 2 4
LDA SCORE (log 10)

e)


C
PSK
Woesearchaeota
Lentisphaerae
Deferribacteres
Synergistetes
Spirochaetes
Elusimicrobia
Cyanobacteria_Chloroplast
Chlamydiae
Actinobacteria
Fusobacteria
Bacteroidetes
−4 −2 0 2 4
LDA SCORE (log 10)

f)

Résumé

Nous avons collecté et comparé les microbiotes et les microbiomes intestinaux de plusieurs dizaines de lézards de l'espèce *Podarcis sicula*, vivant dans des populations continentales et insulaires croates. L'une de ces populations présentait la particularité d'avoir subi un changement de régime alimentaire récent, une transition d'un régime insectivore vers un régime omnivore (à 80% herbivores) sur une période de 46 ans. Les analyses de diversité menées sur la région V4 de l'ARN ribosomique 16S de ces communautés microbiennes ont révélé que la diversité spécifique (diversité alpha) des microbiotes de lézards omnivores (enrichis en archées méthanogènes) excède celle des microbiotes de lézards insectivores. Les communautés microbiennes des lézards apparaissent en outre faiblement structurée : 5 entérotypes peuvent être identifiés au niveau du phylum, et 3 phyla majoraires (les Bactéroidètes, les Firmicutes et les Protéobactéries) sont présents dans cette espèce. Cependant, ni le régime alimentaire, l'origine spatiale ou temporelle, et le sexe des lézards ne se traduisent par des différences significatives et majeures dans les microbiotes. Des analyses linéaires discriminantes avec effet de la taille des OTUs et des reads des microbiomes fonctionnellement annotés indiquent plutôt que le changement de régime alimentaire de *Podarcis sicula* est associé à des changements ciblés dans l'abondance des certains composants du microbiote et du microbiome de ces lézards, nous conduisant à formuler l'hypothèse de changements ciblés des communautés microbiennes dans cet holobionte non-modèle, par opposition à des transformations plus radicales. Sur un plan plus théorique, cette thèse propose également des modèles de réseaux (réseaux de similarité de reads et graphes bipartis) susceptibles d'aider à approfondir les analyses des microbiomes.

Abstract

We collected and compared intestinal microbiota and microbiomes from several *Podarcis sicula* lizards, which live in Croatian continental and insular populations. One of these populations has recently changed its diet over an 46 years timespan, switching from an insectivorous diet to an omnivorous one (up to 80% herbivorous). Diversity analyses of these microbial communities, based on the V4 region of their 16S rRNA, showed that the microbiota taxonomic diversity (or alpha diversity) is higher in omnivorous lizards (enrichment in methanogenic archaea) than in insectivorous ones. Besides, microbial communities seem weakly structured: 5 enterotypes are detected at the phylum level, and 3 major phyla (Bacteroidetes, Firmicutes and Proteobacteria) are present. However, neither diet, spatial or temporal origin, nor lizard gender correlate with significant differences in microbiota. Linear discriminant analyses with size effect, based on OTUs and functionally annotated reads from the microbiomes, suggest that *Podarcis sicula* diet change is associated to targeted changes of the abundance of some enzymes in the microbiomes. Such a result leads us to propose a hypothesis of targeted changes in the microbial communities of this non-model holobiont, instead of more radical transformations. On a more theoretical level, this thesis also proposes network models (Reads similarity networks and bipartite graphs) that can help improving microbiome analyses.