Analyse statistique de réseaux d'associations entre espèces microbiennes à partir de données métagénomiques

Arnaud Cougoul

Université Clermont Auvergne
École Doctorale des Sciences de la Vie, Santé, Agronomie, Environnement

Thèse
présentée pour obtenir le titre de
Docteur d'Université

Spécialité : Biostatistiques

Analyse statistique de réseaux d'associations entre espèces microbiennes à partir de données métagénomiques

---

Arnaud Cougoul

Directrice de thèse : Gwenaël Vourc'h
Co-encadrants : Patrick Gasqui, Xavier Bailly

Soutenue le 29 Novembre 2019 devant le jury composé de :

François Enault	Maître de Conférence, Université Clermont Auvergne, LMGE	Président
Julien Chiquet	Chargé de Recherche, INRA MIA Paris	Rapporteur
Nathalie Peyrard	Directeur de Recherche, INRA MIA Toulouse	Rapportrice
Béatrice Laroche	Directeur de Recherche, INRA MaIAGE Jouy-en-Josas	Examinatrice
Gwenaël Vourc'h	Directeur de Recherche, INRA EPIA Clermont-Ferrand	Directrice
Xavier Bailly	Ingénieur de Recherche, INRA EPIA Clermont-Ferrand	Encadrant

Unité Mixte de Recherche d'épidémiologie des maladies animales et zoonotiques
INRA Centre de Clermont-Theix, 63122 Saint-Gènes-Champanelle

Remerciements

J'aimerais ici remercier toutes les personnes qui m'ont soutenu au cours de ces trois années. Un travail de thèse ne repose pas sur un seul doctorant. J'ai bénéficié de l'expérience et de la sagesse de nombreuses personnes.

Merci à ma directrice de thèse Gwenaël Vourc'h et à mes encadrants Xavier Bailly et Patrick Gasqui qui m'ont soutenu avec bienveillance et enthousiasme. Chers encadrants, je n'oublierais pas nos longues discussions passionnées de travail dans le bureau de Patrick. Vous avez toujours eu les mots justes pour me remettre en selle.

J'adresse toute ma gratitude aux membres du jury de thèse. Je remercie particulièrement Julien Chiquet et Nathalie Peyrard qui me font l'honneur d'être les rapporteurs de ce travail de thèse. Je remercie également François Enault et Béatrice Laroche pour leur engagement. Je tiens aussi à remercier les membres de mes comités de thèse Corine Vacher, David Biron et Julien Chiquet pour avoir suivi mes travaux tout au long de ma thèse.

Ce travail n'aurait pas été possible sans le soutien financier du métaprogramme INRA MEM Méta-omiques des Ecosystèmes Microbiens et du métaprogramme INRA GISA Gestion Intégrée de la Santé des Animaux.

Un grand merci à tous les membres de l'unité d'Epidémiologie des maladies animales et zoonotiques pour les échanges scientifiques et l'environnement amical qu'ils apportent. Mes chers « co-bureaux » et « co-bureautes », Sylvain Coly, Tiphaine Le Bris, Cécile Adam et Christian Ducrot, c'est un honneur de vous avoir rencontré. Je remercie également David Abrial, Jocelyn De Goër, Émilie Bard et Myriam Garrido pour leur disponibilité, toujours prêts à échanger et répondre à mes questionnements.

Merci à Ernst Wit de m'avoir accueilli au sein de son laboratoire au Bernoulli Institute de Groningen aux Pays-Bas. C'était une aventure très enrichissante et je suis

ravi que cette mobilité ait donné lieu à une collaboration. Merci à Agreenium et son parcours doctoral EIRA, merci à l'INRA, au métaprogramme INRA MEM et au réseau européen COSTNET d'avoir financé deux séjours de recherche.

Merci à Iris Eouzan et David Biron du Laboratoire LMGE pour nos échanges et collaborations. Merci encore à Julien Chiquet, et les doctorants de l'UMR Mia-Paris avec qui j'ai pu passer une semaine de recherche dans leur laboratoire. Merci à Céline Delbès de l'UMRF de m'avoir accueilli à Aurillac.

À mes amis qui me sont chers et que j'ai quelque peu délaissés pour achever cette thèse. À mes parents et à ma compagne Sara pour leur soutien pendant ces années. À ma grand-mère, à ma fille Olivia.

Analyse statistique de réseaux d'associations entre espèces microbiennes à partir de données métagénomiques

Le séquençage haut débit révèle une nouvelle écologie des microorganismes. Ils sont présents partout et leurs fonctions sont primordiales pour leurs écosystèmes hôtes, organismes ou environnements. La métagénomique permet notamment d'estimer la composition et l'abondance des espèces microbiennes d'un ensemble d'échantillons de même type de communautés microbiennes. Lors d'études cherchant à comprendre la diversité et la structure de telles communautés, des approches réseaux permettent d'identifier des associations statistiques entre microbes, en faisant l'hypothèse que ces associations statistiques reflètent les interactions biologiques. Dans ce contexte, le sujet de ma thèse était de mieux cerner le potentiel des approches réseaux dans la détection d'associations entre OTUs au sein de données métagénomiques et de développer les outils nécessaires pour améliorer l'analyse des jeux de données. Dans un premier temps, j'ai étudié les pratiques et les outils d'analyse utilisables pour inférer des réseaux d'associations au sein de métagénomes. Compte-tenu des propriétés des données métagénomiques, j'ai déterminé leur efficacité et leurs limites. Ces travaux m'ont permis de déterminer des pistes pour améliorer l'étude des associations microbiennes. Sur la base des connaissances accumulées, j'ai développé un package d'analyse des associations entre OTUs (nommé MAGMA) visant à inférer les associations pertinentes au sein de métagénomes. MAGMA prend en compte les spécificités des données métagénomiques et offre la possibilité de prendre en compte l'effet d'un facteur structurant sur la distribution des OTUs avant de rechercher les associations entre microbes. Par le biais de participations dans différents projets de métagénomique, j'ai confirmé la pertinence de l'outil développé et identifié des pistes d'améliorations permettant de faire face aux problématiques biologiques actuelles.

Statistical analysis of networks of associations between microbial species from metagenomic data

High throughput sequencing reveals a new ecology of microorganisms. They are everywhere and their functions are essential for their host ecosystems, organisms or environments. Metagenomics makes it possible to estimate the composition and abundance of microbial species from a set of samples of the same type of microbial communities. In the studies that seek to understand the diversity and structure of such communities, network approaches can identify statistical associations between microbes, assuming that these statistical associations reflect biological interactions. In this context, the subject of my thesis was to better understand the potential of network approaches in the detection of associations between OTUs within metagenomic data and to develop the necessary tools to improve the analysis of datasets. As a first step, I studied the practices and analysis tools that can be used to infer association networks within metagenomes. Given the properties of metagenomic data, I determined their effectiveness and their limits. This work allowed me to identify ways to improve the study of microbial associations. Based on the accumulated knowledge, I developed an association analysis package between OTUs (named MAGMA) to infer relevant associations within metagenomes. MAGMA takes into account the specificities of metagenomic data and offers the possibility to take into account the effect of a structuring factor on the distribution of OTUs before looking for associations between microbes. Through participations in different metagenomics projects, I confirmed the relevance of the tool developed and identified ways of improving the current biological issues.

Table des matières

Table des figures

Chapitre 1

Introduction

Sommaire

Depuis plusieurs années, l'utilisation d'outils puissants de séquençage nous a révélé l'existence et l'importance de communautés de micro-organismes associés à des environnements vivants : les microbiotes. Les fonctions fournies par les microbiotes sont diverses, allant par exemple du rôle bénéfique de la microflore intestinale dans le processus de digestion à l'effet délétère en cas de perturbation de cette flore. Afin de mieux

appréhender les écosystèmes que représentent les microbiotes et leur environnement, il est nécessaire d'identifier les microorganismes qui sont liés aux fonctions fournies par les microbiotes et connaître les conditions où ils expriment ces fonctions.

La métagénomique permet de décrire les communautés microbiennes. Elle s'est développée ces dernières années avec l'essor du séquençage à haut débit. Elle permet notamment d'estimer à moindre coût la composition microbienne d'un échantillon. Les informations sur la composition et la diversité des microbiotes sont essentielles mais ne permettent pas de décrire complètement les microbiotes car des interactions biologiques structurent les microbiotes : (i) les micro-organismes interagissent entre eux, ils peuvent par exemple s'échanger des nutriments ou au contraire rentrer en compétition pour les ressources ou l'espace, (ii) des facteurs biotiques ou abiotiques peuvent également structurer les microbiotes, comme l'âge de l'organisme hôte du microbiote ou les conditions environnementales.

À partir d'un ensemble d'échantillons de microbiotes, mon travail de thèse consiste à prendre en compte les spécificités des données issues de la métagénomique ainsi que les facteurs structurant les microbiotes, dans le but d'identifier des associations pertinentes entre espèces microbiennes et d'inférer le réseau formé par ces associations.

1.1 La métagénomique pour la description du microbiote

Dans cette section, nous allons d'abord donner une définition du microbiote et des interactions qui s'y opère. Nous décrirons ensuite les approches métagénomiques permettant de caractériser la composition des communautés microbiennes. Nous aborderons les problèmes statistiques liés à l'analyse des communautés de microbes présents dans les échantillons étudiés. Les données métagénomiques sont singulières et leur analyse nécessite de bien intégrer leurs spécificités.

1.1.1 Les microbiotes

Les microorganismes, contractés en « microbes », sont présents presque partout : dans l'eau, dans les sols, en suspension dans l'air. Les microorganismes associés à chacun de ces environnements forment des communautés que l'on nomme microbiotes. Ces microbiotes rendent des fonctions importantes aux écosystèmes. Les microbiotes des sols et des océans sont par exemple indispensables aux processus des cycles biogéochimiques nécessaires au renouvellement et au recyclage des ressources dans les écosystèmes (Konopka, 2009). En dehors des fonctions jouées par les microbiotes dans les cycles de l'azote ou du carbone, certaines espèces de bactéries sont capables de décontaminer les métaux lourds (Burkhardt et al., 1993) ou de dégrader le plastique (Shah et al., 2008).

Les microorganismes sont aussi hébergés par les eucaryotes : les champignons, les plantes, et les animaux, dont l'homme. La composition de ces microbiotes est diverse et leur biologie est liée avec celle de leur environnement hôte, écosystème ou organisme (K. R. Foster et al., 2017). Un seul être humain est habité par environ 10^{14} bactéries et on estime à 10^{30} le nombre de bactéries et d'archae sur terre. Le nombre d'espèces microbiennes est lui estimé à 10^{12} (Whitman et al., 1998 ; Locey et Lennon, 2016).

Les communautés de microorganismes associés à un hôte et leurs génomes forment le microbiome (Klassen, 2018). Les organismes évoluent avec leur microbiome. C'est notamment le cas de l'homme et de son microbiome qui sont en coévolution depuis des millions d'années (Dominguez-Bello et al., 2019). Au sein du microbiome humain, on peut distinguer clairement le microbiote de la peau de celui de la plaque dentaire, de la salive, de l'intestin ou du vagin (Cho et Blaser, 2012). Ces microbiotes contribuent à de nombreux processus physiologiques : meilleure efficacité digestive, activité biochimique et métabolique, activité de synthèse (vitamines), système de défense contre l'invasion d'éventuelles bactéries pathogènes (Kho et Lal, 2018 ; Mohajeri et al., 2018). Ainsi, ces relations symbiotiques confèrent au « supraorganisme » composé d'un organisme hôte et de son microbiome un avantage adaptatif. Cet assemblage de différentes espèces forme une nouvelle unité écologique, l'holobionte (Simon et al.,

2019).

Les microbiotes ont donc des rôles positifs pour leur environnement mais ils peuvent aussi être délétères pour celui-ci. Des perturbations du microbiote intestinal ont été associées à des maladies chroniques chez l'homme comme le diabète, l'obésité ou encore des maladies comportementales (Johnson et K. R. Foster, 2018). Jusqu'à récemment, le concept d'une maladie, un microbe était la norme. En effet, une infection par un microbe peut être responsable de maladies aigües chez l'homme, les animaux et les plantes. Ces microbes peuvent-être des bactéries, des champignons ou des parasites. Cependant, ces infections sont multiples. Le multiparasitisme est fréquent et il est de plus en plus reconnu (Vaumourin et al., 2015). Les agents pathogènes interagissent entre eux et également avec les autres microorganismes. Ainsi, l'étude de la contamination d'un hôte par un agent pathogène pourrait gagner à intégrer les caractéristiques du microbiome et de sa biologie. Cet ensemble forme le pathobiome (Vayssier-Taussat, Albina et al., 2014 ; Vayssier-Taussat, Kazimirova et al., 2015). L'installation d'un agent pathogène peut être favorisée par un facteur environnemental, une caractéristique de l'hôte ou par la présence de microorganismes la facilitant (Bordes et Morand, 2011). Les nombreuses interactions qui s'opèrent dans les microbiotes peuvent donc jouer sur l'installation d'un agent pathogène.

L'étude des microbiotes est donc essentielle dans la compréhension du vivant. Cela commence par la compréhension de l'assemblage de ces communautés en tant que processus spatio-temporel. Comment les bactéries se dispersent, colonisent l'espace et se maintiennent sont autant de questions auxquelles l'écologie microbienne tente de répondre.

Quatre grands processus permettent de synthétiser les différents mécanismes qui régissent l'assemblage des communautés : diversification, dispersion, sélection et dérive (Vellend, 2010 ; Nemergut et al., 2013). Encore peu de choses sont comprises tant ces processus sont difficiles à étudier empiriquement. (A) La diversification est la génération de nouvelles variations génétiques. Celle-ci a lieu notamment au cours de mutations ou de transferts horizontaux de gènes entre bactéries. (B) La disper-

sion est le mouvement des bactéries dans l'espace. Par exemple, la taille peut avoir un effet sur la dispersion. (C) La dérive correspond aux variations stochastiques de l'abondance des espèces. Les espèces de faibles abondances sont plus vulnérables aux effets de la dérive étant donné qu'une fluctuation négative de leurs abondances peut impliquer leur extinction. Cependant, cette conclusion est discutable étant donné que les microorganismes peuvent se mettre en état de dormance (Nemergut et al., 2013). (D) La sélection est une force importante qui façonne les communautés microbiennes. La sélection correspond aux changements dans la communauté dus aux capacités des organismes à s'adapter aux conditions environnementales qu'elles soient d'origine biotique ou abiotique.

De nombreuses études montrent le rôle déterminant des facteurs abiotiques physicochimiques sur la structure, l'assemblage et la diversité des microbiotes. Le pH, la température ou l'hygrométrie sont des facteurs environnementaux connus pour structurer la distribution des microbes du sol (Mandakovic et al., 2018).

La relation entre l'hôte et son microbiote est un facteur biotique qui influence le microbiote. Pour K. R. Foster et al., l'hôte « maintient en laisse » ses microbiotes afin qu'ils accomplissent leurs fonctions avec un coût négligeable (K. R. Foster et al., 2017). Il impose ainsi des pressions de sélections qui vont agir sur les microbiotes. Par exemple, l'âge, la génétique, l'environnement et l'alimentation sont des facteurs qui affectent le microbiote intestinal humain (C. A. Lozupone, Stombaugh et al., 2012).

Bien que moins étudiées, les interactions microbes-microbes sont aussi un des facteurs biotiques qui doit influencer la distribution de l'abondance des espèces au sein d'un microbiote. Les microorganismes doivent non seulement s'adapter à leur milieu mais également survivre aux autres microbes. Les populations de microbes d'espèces différentes constituent un réseau de relations. Les relations entre microbes peuvent impliquer plus de deux espèces (Bairey et al., 2016 ; Pacheco et Segrè, 2019) mais les relations par paires sont un premier maillon essentiel dans la description de ces communautés microbiennes. Les interactions entre deux espèces peuvent être classées en différents types selon les impacts positif, négatif ou neutre que les deux espèces ont

l'une envers l'autre (LIDICKER, 1979). (Figure 1.1) :

— Le mutualisme et la symbiose sont des interactions où les deux acteurs s'entraident et tirent un bénéfice réciproque (HOEK et al., 2016).

— A l'opposé, deux entités sont considérées en compétition lorsque la présence de l'une nuit à la présence de l'autre et réciproquement. Ce phénomène est notamment observé quand deux espèces partagent une même ressource limitante (GHOUL et MITRI, 2016).

— La prédation et le parasitisme sont des interactions où une espèce tire profit de la présence d'une autre espèce et cette relation a un effet délétère pour cette dernière.

— L'amensalisme est une interaction biologique dans laquelle l'interaction est négative pour l'un des partenaires alors qu'elle est neutre pour l'autre partenaire.

— Le commensalisme est une interaction où l'effet est positif pour une espèce et neutre pour l'autre.

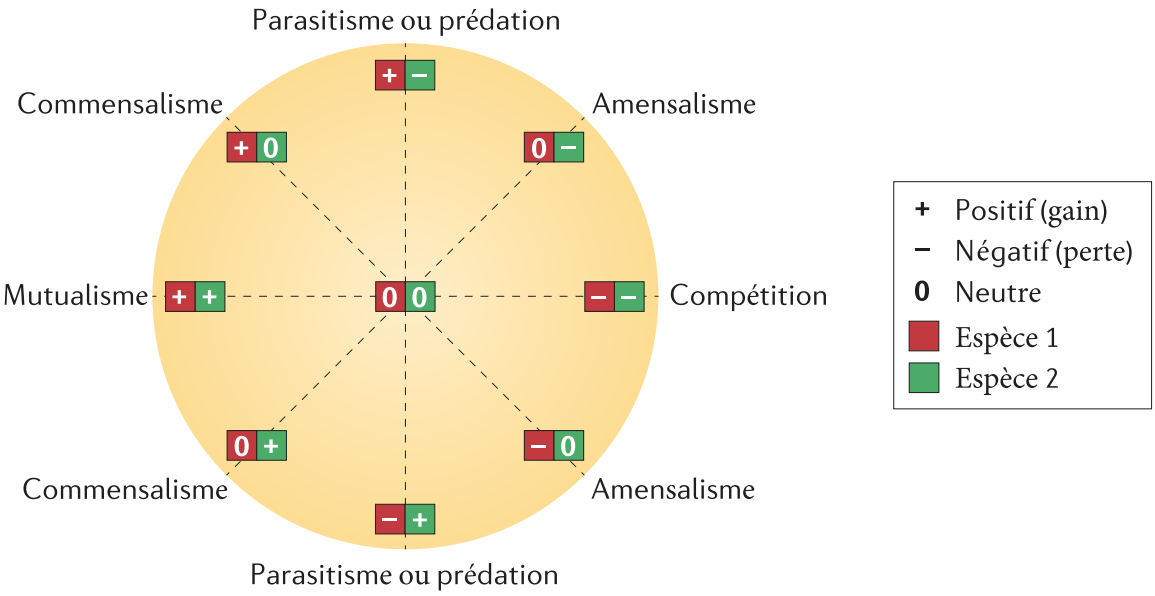


FIGURE 1.1 – Résumé des interactions écologiques entre membres de différentes espèces. La roue de Lidicker présentée a été adaptée par FAUST et RAES, 2012. Il existe trois effets possibles pour chaque partenaire d'interaction : positif (+), négatif (-) et neutre (0). Par exemple, dans le parasitisme, le parasite bénéficie de la relation (+), alors que l'hôte est désavantagé (-) ; cette relation est donc représentée par la paire de symboles + -.

Des études théoriques illustrent que la distribution des différentes interactions

entre microbes peuvent influencer les équilibres potentiels entre membres d'un microbiote (Gonze, Lahti et al., 2017 ; Gonze, Coyte et al., 2018), la stabilité de ces équilibres et, par conséquent, les propriétés des interactions entre hôtes et microbiotes (Coyte et al., 2015). Cela souligne l'importance d'étudier les interactions entre microbes au sein des microbiotes pour optimiser leur gestion.

1.1.2 Protocole d'étude

Pour étudier un microbiote donné, les biologistes récoltent des échantillons afin d'en étudier leur composition (Knight et al., 2018). En dehors du choix des techniques de métagénomique que nous exposerons dans le paragraphe suivant, différents plans d'étude sont fréquemment rencontrés :

(1) L'étude se concentre sur un groupe d'échantillons d'un seul type de microbiote étudié dans des conditions similaires pour limiter les facteurs confondants. L'objectif peut être d'étudier la composition, la structure ou les interactions microbes-microbes de ce microbiote.

(2) L'étude est faite sur des échantillons selon un gradient environnemental ou selon des caractéristiques de l'hôte afin d'étudier l'effet de ce facteur.

(3) L'étude est faite sur des échantillons récoltés sur un même individu au cours du temps afin d'étudier la dynamique de la composition du microbiote, les successions d'espèces ou les interactions entre espèces. Dans ce cas on parle d'étude longitudinale ou temporelle du microbiote.

Pour les cas (1) et (2) on parle d'étude transversale. Les études observationelles transversales représentent la grande majorité des études de microbiote. Dans le contexte de mon unité d'accueil, en épidémiologie animale, les microbiotes étudiés sont de fait prélevés sur des arthropodes vecteurs de maladies comme les tiques, capturés à l'affût. Ce modèle biologique rend les études longitudinales difficiles. Nous nous placerons donc dans le cas des études transversales.

1.1.3 La métagénomique

À partir des échantillons obtenus, les communautés microbiennes peuvent être caractérisées à l'aide de la métagénomique. En effet, 1% du nombre de bactéries de la plupart des environnements peut être cultivé (Steen et al., 2019 ; Martiny, 2019). Une grande proportion des espèces microbiennes reste donc non cultivables. Le séquençage haut-débit à la base des approches de métagénomique s'affranchit du besoin de culture et peut générer assez de séquences pour couvrir tous les organismes en présence dans plusieurs échantillons en un seul *run*. L'acide nucléique ADN ou ARN est extrait à cet effet de chaque échantillon. Les techniques de séquençage modernes de tout ou partie de ces acides nucléiques permettent ensuite d'obtenir à moindre coût la caractérisation de l'abondance des espèces microbiennes d'un environnement.

Il existe en effet deux grandes approches en métagénomique : la métagénomique ciblée et la métagénomique globale (Oulas et al., 2015 ; Breitwieser et al., 2017). Lorsqu'on étudie la composition des microbiotes, la métagénomique ciblée est préférée à la métagénomique globale, principalement pour des raisons de coût et d'analyse (Jovel et al., 2016 ; Knight et al., 2018).

La métagénomique globale « shotgun » est une approche métagénomique qui consiste à fragmenter et séquencer tous les acides nucléiques (ADN ou ARN) à l'aide d'un séquenceur à haut débit (Venter, 2004). Les séquences (reads) lues sont ensuite assemblées pour reconstruire les gènes microbiens présents dans les échantillons. La métagénomique globale va séquencer tout le matériel génétique présent dans un échantillon : par exemple, bactériens, viraux ou encore celui de l'hôte (Qin et al., 2010 ; Razzauti et al., 2015). Pour avoir ce niveau de description, il faut des efforts de séquençage importants qui peuvent être coûteux.

La métagénomique ciblée permet de décrire exclusivement des groupes spécifiques comme les communautés bactériennes et archées ou les champignons. Après extraction des ADNs, un gène d'identification est ciblé, amplifié par PCR et séquencé. En bactériologie, le gène de l'ARN 16S est ciblé (Pollock et al., 2018). Des régions du 16S sont recommandées selon le type de microbiote (Bukin et al., 2019). Les champignons

peuvent aussi être caractérisés par le séquençage du gène 18S ou ITS (Huffnagle et Noverr, 2013). Le nombre de lecture d'une même séquence 16S permet de donner une idée de l'abondance de cette bactérie. Cette technique est appelée metabarcoding ou métagénomique ciblée et c'est cette approche que nous étudierons plus particulièrement car elle reste encore la plus fréquente.

Une fois le séquençage réalisé, un traitement bioinformatique est opéré. L'objectif est d'assigner à chaque séquence le nom de la bactérie correspondante.

1.1.4 Traitement bioinformatique

Les différentes étapes d'analyses bioinformatiques sont regroupées au sein de pipelines dédiés comme mothur (Schloss, Westcott et al., 2009) ou QIIME (Caporaso et al., 2010). En premier lieu, plusieurs étapes de filtrage et de nettoyage des données se succèdent pour tenter de contrôler au mieux les nombreuses manipulations de l'échantillonage au séquençage (Brooks et al., 2015 ; Galan et al., 2016). Ensuite, il faut assigner à chaque lecture l'OTU correspondant. Un OTU (Unité Taxonomique Opérationnelle) est un groupement d'organismes sur la base de leur proximité phylogénétique. Un OTU peut être associé à une espèce microbienne mais peut aussi ne correspondre à aucune espèce identifiée. Le terme d'espèce pour les microbes est d'ailleurs controversé (Gevers et al., 2005 ; Achtman et Wagner, 2008).

Deux grandes stratégies d'assignation existent (Navas-Molina et al., 2013). Pour la première approche, les séquences lues sont assignées selon la plus proche taxonomie présente dans les bases de données. Greengene (DeSantis et al., 2006), Silva (Quast et al., 2013) et RDP (Cole et al., 2014) sont les bases de données d'ARN 16S les plus connues. Avec cette approche, les séquences absentes des bases de données sont annotées de manière partielle ce qui pose problème. La deuxième approche consiste à regrouper les séquences lues *de novo* par similarité, sans utilisation de base de données. Les séquences regroupées élisent alors une séquence consensus qui peut ensuite être annotée selon une base de données ou définir une espèce inconnue. Pour ces deux stratégies, le taux de similarité pour le regroupement des séquences est fixé à 97% depuis 1994 (Stackebrandt et Goebel, 1994). Un taux de 99% semblerait plus

adapté (Edgar, 2018).

Les données de lecture de séquence ainsi obtenues forment une table d'OTUs que l'on va chercher à décrire et à analyser. Cette table représente le nombre de lecture des OTUs en colonne, pour chaque échantillon en ligne. Ces quantités permettent d'approcher l'abondance des OTUs. Pour modéliser au mieux ces abondances, l'ensemble des caractéristiques de nos données métagénomiques doit être pris en compte.

1.1.5 Un grand nombre d'OTUs à faible prévalence

Une table d'OTUs a une dimension d'un ordre de grandeur de plusieurs milliers d'OTUs, et d'une centaine à plusieurs milliers d'échantillons. Le nombre d'OTUs dépend du type de microbiote étudié. Ce grand nombre d'OTUs reflète la diversité généralement rencontrée dans les écosystèmes microbiens.

D'un point de vue statistique, la table d'OTUs comporte plus de variables que d'individus. L'analyse de ce type de données est un problème de grande dimension. Le point de vue est inversé par rapport au cadre statistique classique où il y a plus d'individus que de variables. Les méthodes d'analyses classiques ne sont pas adaptées aux problèmes de grande dimension, au *fléau de la dimension* (Bellman, 1957). En grande dimension, on suppose que peu de variables sont pertinentes en appliquant un principe de parcimonie. Les méthodes les plus utilisées pour pallier au fléau de la dimension sont (i) les méthodes classiques de sélection de variables (choix du sous-modèle avec le plus faible taux d'erreur, e.g. selon le R^2 ou critère AIC), (ii) les méthodes de réduction de dimension comme l'analyse en composante principale (ACP) ou la régression des moindres carrés partiels (PLS) et (iii) les méthodes d'estimations contraintes de type lasso, ridge ou plus généralement elastic net (Fan et Lv, 2010).

Parmi ce grand nombre d'OTUs, la majorité sont rares et il y a seulement une poignée d'OTUs généralistes, ubiquistes, c'est-à-dire présents dans tous les échantillons (Lynch et Neufeld, 2015 ; Jousset et al., 2017). La rareté est ici dans le sens du faible taux de présence observée et non en terme d'abondance. Le taux de présence observée sur l'ensemble des échantillons vient approximer le taux de prévalence au sens

épidémiologique. Nous utiliserons ici la notion de prévalence d'un OTU comme le taux de présence observé sur le groupe d'échantillons.

Les faibles prévalences des OTUs impliquent un excès de zéros observé dans les tables d'OTUs ($\backsim$ 90% de zéros) (Paulson et al., 2013). Pour l'analyse numérique, une table d'OTUs est une matrice creuse, clairsemée (*sparse matrix*), avec beaucoup de zéros (Duff et al., 1986). Cette caractéristique limite les analyses statistiques de ces données : il sera difficile de tirer de l'information d'un OTU qui est très peu prévalent or les OTUs rares sont la majorité. Cette caractéristique sera donc essentielle à prendre en compte à l'aide de modèles adaptés.

L'effet des faibles prévalences sur les analyses statistiques n'est pas encore bien décrit dans la littérature. De plus, en épidémiologie, les agents pathogènes d'intérêt sont souvent - et heureusement - peu prévalents. Or les OTUs de faibles prévalences sont classiquement filtrés arbitrairement et empiriquement pour diminuer la proportion de zéros dans les matrices de données et ainsi améliorer la qualité des analyses (Weiss, Van Treuren et al., 2016). La détection d'associations impliquant des OTUs à faibles prévalences apparaît donc comme un challenge méthodologique pour lequel il faut étudier les limites de l'analyse et exploiter au mieux l'information présente dans les jeux de données métagénomiques.

1.1.6 Normalisation des données compositionnelles

Chaque échantillon subit un processus d'extraction de l'acide nucléique, d'amplification et de séquençage. Des biais importants apparaissent à l'issu du processus d'obtention des données. La profondeur de séquençage, aussi appelée taille de librairie, est souvent définie comme la somme des reads des différents OTUs dans un échantillon. Elle est propre à chaque échantillon, sans réalité biologique : une somme totale importante ne veut pas dire qu'il y a un plus grand nombre de bactéries dans l'échantillon. Les données de lecture d'OTUs doivent être prises en compte selon la profondeur de séquençage de chacun des échantillons : les données métagénomiques sont des données de composition (Gloor, Wu et al., 2016 ; Gloor, Macklaim et al., 2017 ; Quinn et al., 2018).

Dans un échantillon à faible profondeur de séquençage, certains OTUs en faibles proportions pourraient ne pas être détectés. Pour vérifier que tous les OTUs ont bien été récupérés, les biologistes vont faire une étude de raréfaction en observant les courbes de raréfaction (Sanders, 1968 ; Schloss et Handelsman, 2004). Pour construire ces courbes, il suffit de compter le nombre d'OTUs pour un ensemble de sous-échantillons à différents intervalles de profondeur. Lorsque ce nombre se stabilise, on observe une asymptote horizontale qui est la profondeur de séquençage suffisante pour observer tous les OTUs présents dans l'échantillon. Après une étude de raréfaction sur tous les échantillons, ceux dont les courbes ne se stabilisent pas sont considérés comme n'étant pas assez séquencés.

La profondeur de séquençage étant variable entre échantillons, ils ne sont pas directement comparables. Il faut donc transformer les données. Une des premières normalisation que les biologistes ont appliqué sur les données métagénomiques consiste à raréfier les données (Gotelli et Colwell, 2001 ; Horner-Devine et al., 2004 ; Hughes et Hellmann, 2005) - à distinguer de la raréfaction vue précédemment. Les données raréfiées sont obtenues en trois étapes : la première étape consiste à sélectionner une taille de librairie minimale, un seuil. Les échantillons ayant moins de lectures que le seuil sélectionné sont éliminés du jeu de données. Les librairies restantes sont sous-échantillonnées afin que tous les échantillons aient la même taille de librairie. Le seuil peut être choisi après étude de raréfaction, avec le maximum des profondeurs de séquencage nécessaire pour atteindre le palier des courbes de raréfaction. Une telle normalisation par sous-échantillonnage est aberrante d'un point de vue statistique étant donné la perte d'information engendrée (McMurdie et Holmes, 2014). Les données raréfiées ne prennent plus en compte la profondeur de séquençage initiale.

Afin d'intégrer l'effet de composition, les biologistes utilisent également les données relatives : ce sont les proportions, les fréquences des OTUs, relativement à l'effort de séquençage. L'opération de transformation appelée « total sum scaling » (TSS) consiste simplement à diviser les données de lecture de chaque échantillon par sa profondeur de séquençage (Paulson et al., 2013 ; Weiss, Z. Z. Xu et al., 2017 ; Kumar et al., 2018).

Les données relatives sont souvent utilisées par les biologistes pour décrire la composition des jeux de données métagénomiques puisqu'elles permettent de bien représenter la composition en pourcentage de chaque OTUs. Cependant, l'analyse de ces données peut conduire à de fausses interprétations statistiques notament dans l'étude des associations (Aitchison, 1982 ; Friedman et Alm, 2012). Il se trouve que la contrainte sur les sommes marginales par échantillons engendre un simplex où les mesures d'associations usuelles ne sont pas applicables. En 1897, Karl Pearson mettait en garde contre « les tentatives d'interprétation des corrélations entre ratios dont les numérateurs et les dénominateurs contiennent des parties communes » (Pearson, 1897).

Des transformations suggérées par Aitchinson permettent de transformer les simplex d'Aitchinson en espace réel où les mesures d'associations usuelles sont applicables. Ces transformations sont le logratio additif (alr) et le logratio centré (clr) (Aitchison, 1986). La transformation clr est largement utilisée pour les données microbiennes.

La profondeur de séquençage, calculée comme la somme des lectures d'un échantillon, n'est pas un bon estimateur de la taille de l'effet produit par la variabilité de la profondeur de séquençage. En effet, la somme comme la moyenne ne sont pas des estimateurs statistiques robustes. La transformation clr utilise la moyenne géométrique pour prendre en compte la taille de cet effet. La moyenne géométrique est moins sensible aux valeurs extrêmes que la moyenne arithmétique. Des méthodes initialement développées pour la transcriptomique proposent d'estimer la taille de l'effet à l'aide de statistiques robustes comme la médiane et la moyenne géométrique dans l'outil DESeq (Anders et Huber, 2010) ou la moyenne tronquée dans l'outil edgeR (Robinson et Oshlack, 2010). L'outil metagenomeseq (Paulson et al., 2013), prévu pour la métagénomique ciblée propose également une autre estimation de la profondeur de séquençage, la normalisation *cumulative-sum scaling* (CSS). Les données de comptage brutes sont ici divisées par la somme cumulative des comptes allant jusqu'à un centile déterminé à partir des données. Les premiers centiles étant stables et les derniers instables (de grande variabilité), le centile choisi est le premier centile pour lequel une instabilité est détectée. La transformation clr est formellement similaire aux normalisations « efficaces » fournies par DESeq et edgeR (Quinn et al., 2018).

Cette transformation utilise l'ajout d'un « pseudocount » aux données, souvent égal à 1, afin d'éviter le calcul du logarithme de zéro. Toutefois, l'utilisation de « pseudocount » peut induire des variations importantes dans l'analyse des données métagénomiques qui contiennent d'autant plus une grande majorité de zéros (Weiss, Z. Z. Xu et al., 2017). La normalisation est en effet sensible à la valeur du « pseudocount » (Costea et al., 2014). L. Chen et al., 2018 proposent une alternative robuste appelée GMPR, sans l'utilisation de « pseudocount ».

Le choix de la normalisation reste un challenge scientifique pour intégrer au mieux les données aux outils d'analyses existants (Weiss, Z. Z. Xu et al., 2017 ; Pereira et al., 2018). Une fois la variabilité de la profondeur de séquençage prise en compte, il est possible de comparer les échantillons entre eux.

1.1.7 Détection de facteurs structurant les communautés

Dans le cadre de l'analyse d'associations, nous aimerions pouvoir détecter les facteurs structurants afin de les prendre en compte dans l'analyse. Il est difficile de traiter simultanément toutes les espèces. Des tests d'hypothèses multivariés sont généralement utilisés pour évaluer les variations globales des microbiotes en fonction d'un facteur (Xia et Sun, 2017). Deux solutions existent : (i) la dimensionnalité est d'abord réduite et ensuite les hypothèses sont testées, ou (ii) des régressions sont effectuées pour chaque espèce séparément, puis la dimensionnalité des résultats est réduite en sommant les statistiques (Warton et al., 2012), en tenant compte des corrélations entre espèces (Y. Wang et al., 2012). La première méthode est la plus largement utilisée pour analyser les données de communautés. La première étape consiste à réduire la dimensionnalité, principalement à l'aide d'une mesure de diversité.

La première diversité est la diversité alpha qui consiste à déterminer la diversité locale au sein de chaque modalité de facteur. Différents indices de diversité peuvent être utilisés (T. C. Hill et al., 2003). La richesse spécifique d'un échantillon est le nombre d'OTUs présents dans l'échantillon. Au-delà de la richesse spécifique, les indices de diversité spécifique, comme ceux de Shannon et Simpson, prennent également en compte

la répartition des abondances des OTUs. Ce sont des indices d'équitabilité (evenness), i.e. régularité des distributions des OTUs. Après calcul de cette mesure univariée, il est possible d'effectuer un test d'hypothèse classique de dépendance à un ou plusieurs facteurs.

Le type complémentaire de diversité caractérisant les différences de diversité entre échantillons est la diversité beta. Elle permet d'estimer la différence de diversité entre modalités de facteur. Les distances ou dissimilarités entre chaque paires d'échantillons sont calculées pour obtenir une matrice de distances représentant les scores de beta diversité. L'indice de Jaccard est le premier indice de beta diversité (Jaccard, 1901). L'indice de dissimilarité de Bray-Curtis est le plus utilisé par les écologues car il prend en compte les proportions des espèces (Bray et Curtis, 1957). La distance Unifrac (Unique fraction metric) permet de prendre en compte les distances phylogénétiques inter-OTUs (C. Lozupone et Knight, 2005 ; C. A. Lozupone, Hamady et al., 2007) à partir des longueurs de branche des OTUs partagés par les échantillons. La version non pondérée (unweighted Unifrac) utilise uniquement la présence/absence des OTUs. La version pondérée (weighted) prend en compte les abondances des OTUs en multipliant chaque longueur de branche par la différence d'abondance des descendants de la branche. Pour tester si la beta-diversité diffère d'un groupe à l'autre, une analyse des similarités (ANOSIM) ou une MANOVA non paramétrique (PERMANOVA ou autrement appelé NP-MANOVA) sont utilisées (Clarke, 1993 ; Anderson, 2001). Ces méthodes se basent sur les rangs et les p-valeurs sont obtenues par permutation. La PERMANOVA est la plus utilisée par les biologistes et semble plus robuste pour les données métagénomiques (Anderson et Walsh, 2013).

La question de la normalisation des données se pose pour l'utilisation de mesures de diversité construites initialement pour des données macroécologique. Bien que le choix de la métrique de diversité influence le plus la puissance des tests (Thorsen et al., 2016), le choix de la normalisation des données est aussi influant. McMurdie et Holmes, 2014 puis Weiss, Z. Z. Xu et al., 2017 ont fournit des études comparatives sur ce choix de normalisation. Pour la métrique de Bray-Curtis, il est ainsi conseillé de prendre les proportions et non les transformations en log qui « faussent

les comparaisons de communautés en supprimant les différences importantes dans les OTUs communs et en amplifiant les légères différences dans les OTUS rares » (McKnight et al., 2019) . Pour la mesure unweighted Unifrac, il est conseillé de raréfier les données (Knight et al., 2018), bien que ce ne soit pas idéal (McMurdie et Holmes, 2014). Il est également possible d'ajouter la profondeur de séquençage en paramètre du modèle (Weiss, Z. Z. Xu et al., 2017). Cette dernière solution n'a malheureusement pas encore été comparée aux modèles utilisant les données normalisées.

D'autres solutions se montrent particulièrement intéressantes. Des dévelop-pements ont été réalisés sur les mesures de diversité notamment en utilisant les corrélations entre OTUs pour construire un indice de similarité. Deux échantillons ayant peu d'OTUs en commun mais partageant des OTUs similaires en terme de corrélation aux autres OTUs de la communauté vont se retrouver proches par rapport à l'indice TINA (Schmidt et al., 2017). Cette méthode pourrait permettre de tester et valider les facteurs à prendre en compte dans l'analyse des associations. Un autre modèle prometteur propose de construire des vecteurs latents résumant le jeux de données (Sohn et H. Li, 2018). Chaque OTU est modélisé par une loi zero-inflated quasi–Poisson ayant pour moyenne une combinaison linéaire des vecteurs latents. Les vecteurs latents obtenus peuvent êtes utilisés comme les composantes principales d'une analyse en coordonnées principales et ils peuvent aussi servir à tester l'effet d'un facteur structurant.

D'une manière générale, ce chapitre illustre le travail nécessaire pour intégrer la normalisation des données métagénomiques, la prise en compte de facteurs structurants pour la modélisation des abondances d'OTUs et la modélisation des associations entre OTUs dans les jeux de données métagénomiques.

1.1.8 Modélisation statistique des OTUs

Pour modéliser l'abondance des OTUs et leur donner du sens biologique, il faut tenir compte de leurs caractéristiques. Ces données sont compositionnelles, surdispersées, potentiellement structurées et comportent un excès de zéros.

Les données de comptage de communautés Y_{ij} sont habituellement décrites par des

lois binomiales $B(n_i, p_j)$ avec n_i le nombre d'individus dans un lieu i et p_j la probabilité de présence d'une espèce j. Asymptotiquement quand n est grand, la loi binomiale tend vers la loi de Poisson qui possède de bonnes propriétés statistiques.

$$B(n_i, p_j) \hookrightarrow P(\lambda_{ij})$$

avec $\lambda_{ij} = n_i \times p_j$ la moyenne de la loi de Poisson.

Pour les données de communautés microbiennes, la profondeur de séquençage des échantillons est variable et le nombre de reads d'un OTU doit être pris en compte relativement à cette taille de librairie. Pour prendre en compte cet effet en modélisation, la taille de librairie, ou une estimation de celle-ci, va être ajoutée en « offset » dans le modèle.

$$log(\lambda_{ij}) = \beta_j + log(\sigma_i)$$

avec σ_i représentant la taille de librairie de l'échantillon i
et β_j la moyenne de l'OTU j sans l'effet de la taille de librairie.

Les données métagénomiques sont surdispersées : la variance observée est bien supérieure à la moyenne. Par conséquent, la loi de Poisson ne peut pas être utilisée. Les données de comptage surdispersées peuvent être modélisées par la loi binomiale négative (Lindén et Mäntyniemi, 2011 ; Coly et al., 2016). La loi de Poisson log-normale peut également être utilisée. La loi binomiale négative est équivalente à une loi de Poisson dont le paramètre varie suivant une loi Gamma. Biologiquement, la probabilité p_j de présence de l'espèce j est une variable aléatoire. Pour ajuster la variance et donc modéliser la surdispersion, la loi binomiale négative $BN(\lambda_{ij}, \theta_j)$ possède un paramètre supplémentaire θ_j défini par :

$$var(Y_{ij}) = \lambda_{ij} + \frac{\lambda_{ij}^2}{\theta_j}$$

La loi binomiale négative est polyvalente : elle converge en distribution vers la loi de Poisson pour θ_j grand et permet de modéliser des données surdispersées pour θ_j petit.

La loi binomiale négative peut expliquer une grande partie de la variance des données métagénomiques mais n'explique pas nécessairement l'excès de zéros observé (Cunningham et Lindenmyer, 2005 ; Gonzales-Barron et al., 2010). Biologiquement, si l'échantillon n'a pas été exposé à un OTU, cet OTU n'a aucune chance d'être présent. Un OTU peut être absent historiquement d'un échantillon sans qu'on ne maîtrise les éventuels facteurs. Les lois de type « zero-inflated » (ZI) permettent d'augmenter la probabilité d'obtenir un zéro en venant ajouter à la loi une probabilité d'obtenir un zéro dit *structurel* (Ridout et al., 1998 ; L. Xu et al., 2015). Dans la modélisation de l'abondance des OTUs, la loi $ZIBN(\lambda_{ij}, \theta_j, \pi_j)$ a un paramètre π_j supplémentaire correspondant à la probabilité de zéro structurel de l'OTU j :

$$Y_{ij} \sim \begin{cases} 0 & \text{avec probabilité } \pi_j, \\ BN(\lambda_{ij}, \theta_j) & \text{avec probabilité } 1 - \pi_j. \end{cases}$$

La modélisation de données « zero-inflated » peut fournir des informations sur les mécanismes écologiques susceptibles d'avoir généré les données. Ces lois permettent de modéliser séparément les composantes de présence/absence, et d'abondance en cas de présence. Les distributions de type « zero-inflated » sont de plus en plus utilisées pour modéliser les données métagénomiques (Paulson et al., 2013 ; E. Z. Chen et H. Li, 2016 ; Jonsson et al., 2018). La loi $ZIBN$ s'ajuste le mieux aux données dans différents cas d'études de données métagénomiques (Kurtz et al., 2015 ; R. Fang et al., 2016).

Les facteurs structurants peuvent être pris en compte dans la modélisation de l'abondance des OTUs en ajoutant d'éventuelles covariables X au modèle.

$$log(\lambda_{ij}) = \beta_j + X_i^t \gamma_j + log(\sigma_i)$$

où β_j représente la moyenne de l'OTU j, γ_j l'effet de X et σ_i la taille de librairie de l'échantillon i.
Une fois les facteurs ajoutés au modèle, l'effet « moyen » du facteur est *gommé*.

Nous avons présenté les données métagénomiques, leurs caractéristiques et une

modélisation de celles-ci. L'intégration des contraintes exposées est un challenge méthodologique majeur dans le cadre de ma thèse.

1.2 Analyse des réseaux d'associations microbiennes

L'analyse des associations entre OTUs est étudiée pour dégager de la connaissance sur la structure des microbiotes, étudier les interactions entre microbes, et identifier les éléments clés pour assurer leur gestion. L'analyse des réseaux d'associations nécessite la maîtrise de la notion d'association statistique et des outils d'inférence de réseaux. Revenons tout d'abord sur la question biologique : la détection de potentielles interactions biologiques.

1.2.1 Des interactions aux associations

Nous souhaitons étudier les potentielles interactions entre OTUs à partir des données de microbiotes. Les interactions biologiques sont difficiles à évaluer et l'étude des associations statistiques apparaît comme le premier outil permettant de visualiser les patterns de co-occurrences d'OTUs.

Pour notre recherche de potentielles interactions à partir des associations statistiques, nous aurions besoin du postulat irréaliste : « Deux OTUs sont en interaction biologique, si et seulement si une association statistique est observable ». L'étude des deux implications suivantes permettraient d'en savoir plus sur le lien entre association et interaction :

— « Si deux OTUs sont en interaction biologique, alors une association statistique est observable » et sa contraposée « Si aucune association n'est observée, alors les deux OTUs ne sont pas en interaction. »

— « Si une association est observée alors les deux OTUs sont en interaction biologique » et sa contraposée « Si deux OTUs ne sont pas en interaction biologique, alors aucune association statistique n'est observable. »

La compréhension de ces assertions semble essentielle afin de se rapprocher au plus près de la biologie et déceler les limites des analyses d'associations. La distance entre

association et interaction est difficile à évaluer. Certains ont proposé des modèles dynamiques pour simuler des interactions et tester si les mesures d'associations étaient capables de retrouver ces interactions (Weiss, Z. Z. Xu et al., 2017). Toutes les interactions ne semblent pas être observables à partir des corrélations. Dans notre propos, nous nous baserons sur l'étude des corrélations tout en ayant conscience qu'il n'est pas possible de conclure à une interaction biologique.

Nous présenterons dans un premier temps un panel de mesures d'associations statistiques pour définir cette notion d'association. Nous décrirons ensuite les différentes méthodes pour inférer des réseaux d'associations.

1.2.2 Mesure d'associations

Pour identifier une association statistique entre deux OTUs à partir de données transversales d'enquêtes de composition, il est possible d'utiliser un indice de corrélation, une distance, une mesure de similarité/dissimilarité. Il est également possible de détecter des schémas de co-présence/co-exclusion à l'aide des données de présence/absence des OTUs ou à partir des détections spécifiques des agents pathogènes qui peuvent être réalisées en analyses supplémentaires aux analyses 16S. Faust et Raes, 2016 fournissent un bon aperçu des mesures d'associations utilisées avec les données 16S.

Le test de Fischer exact et le test du χ^2 d'indépendance sont utilisés pour déceler une relation entre deux variables qualitative/binaire. Au niveau des coefficients de corrélation, le coefficient Phi donne une mesure de la corrélation analogue à la corrélation de Pearson (Yule, 1912). Le coefficient Q de Yule et les odd-ratio sont aussi des mesures d'associations entre données binaires (Tan et al., 2004).

Pour les données d'abondance, la première mesure de corrélation est le coefficient de Pearson. Il découle directement de la mesure de la covariance. La corrélation de Spearman est une mesure non paramétrique qui est obtenue en mesurant la corrélation de Pearson sur les données de rangs. La corrélation de Kendall est une autre mesure non paramétrique. Une étape de rééchantillonage permet d'améliorer les tests de significativité des corrélations (Bishara et Hittner, 2012)

La mesure d'une distance permet d'apprécier la proximité entre deux variables et peut être utilisée pour mesurer une association statistique entre deux variables même si ce n'est pas son objectif premier. Un test statistique par simulation permet ensuite de tester si la distance mesurée est typique de la distribution nulle où les deux variables sont indépendantes. Une distance utilisée en présence de données écologiques de comptage de communautés est la distance de Bray-Curtis. Il existe un grand nombre de distances utilisées par les écologistes, pour ne citer que les principales : euclidienne, Manhattan, Mahalanobis, Jaccard (Goslee et Urban, 2007).

Des mesures issues de la théorie de l'information permettent de détecter des relations non linéaires et non monotones comme l'information mutuelle (MI) et le coefficient d'information maximale (MIC) qui permettent de détecter les formes de corrélations monotones classiques mais aussi des formes non monotones (Reshef et al., 2011).

Une mesure d'association proposée par Aitchinson permet de pallier au problème de compositionnalité sans besoin de normaliser les données : la variance du log ratio des variables (Aitchison, 1986).

Certaines de ces mesures sont plus facilement implémentables dans le cadre d'une étude systématique des associations entre OTUs : les approches réseaux.

1.2.3 Méthodes réseaux

Pour l'étude des associations à partir de données métagénomiques, le grand nombre d'OTUs à analyser nécessite l'utilisation de méthodes adaptées pour répondre aux problèmes méthodologiques qui en découlent, comme la multiplicité des tests ou l'inférence parcimonieuse. Pour cela, nous avons recours aux méthodes de « réseaux ».

Il existe deux grandes familles d'inférence de réseaux : les réseaux basés sur les corrélations et les modèles graphiques. De nombreux articles de revues exposent ces différentes méthodes appliquées aux données métagénomiques (Faust et Raes, 2012 ;

Layeghifard et al., 2017 ; Jiang et Hu, 2016 ; C. Li et al., 2016 ; Röttjers et Faust, 2018 ; Dohlman et Shen, 2019).

Pour obtenir un réseau à partir des corrélations (relevance network, correlation-based network), il suffit de calculer la matrice de corrélations, c'est-à-dire toutes les corrélations par paires et ensuite seuiller les valeurs de corrélations significatives (Faust, Sathirapongsasuti et al., 2012). On obtient ainsi un réseau graphique en reliant les OTUs représentés par les nœuds à l'aide de liens si la corrélation est supérieure au seuil de significativité. Les différentes mesures d'associations exposées dans la section précédente peuvent être utilisées. La normalisation des données et le choix du seuil de significativité sont très importants ici. La méthode CONET (Faust et Raes, 2016) propose une étape de permutation et bootstrap afin d'évaluer une p-valeur empirique et de mieux valider les liens. Cette méthode permet de palier aux nombreux biais provenant des caractéristiques des données métagénomiques. L'inférence du réseau à partir des corrélations nécessite d'effectuer un grand nombre de tests, une correction de la p-valeur pour des tests multiples doit être effectuée (Storey, 2002). Le principal défaut de ces méthodes est que les associations obtenues prennent en compte les effets indirects dus aux associations entre les autres paires d'OTUs. Une méthode de déconvolution de réseau (Feizi et al., 2013) a été développée pour distinguer les effets directs. Cette méthode est très peu utilisée. Les méthodes SparCC (Friedman et Alm, 2012) et CCLasso (H. Fang et al., 2015) se basent sur la variance du log ratio pour mesurer les associations et sur un principe de parcimonie pour ne pas inférer trop de liens.

Les modèles graphiques gaussiens se basent sur les propriétés des champs de Markov : l'indépendance est conditionnelle, i.e. la dépendance entre deux OTUs est considérée en enlevant l'effet des autres OTUs. Lorsque les données sont normales, cela revient à calculer des corrélations partielles. Un zéro dans la matrice de précision (inverse de la matrice de corrélation) est nécessaire et suffisant à l'indépendance conditionnelle, i.e. à l'absence de liens (Whittaker, 1990). Meinshausen et Bühlmann, 2006 et Friedman, Hastie et al., 2008 proposent deux algorithmes d'inférence du modèle graphique gaussien régularisé par pénalisation L1 : les premiers proposent une méthode de « sélection du voisinage » et les seconds la méthode de « lasso graphique » , aussi

appelée glasso. La pénalisation L1 appelée lasso est utilisée pour réduire la dimension des données et sélectionner les corrélations les plus pertinentes. De nombreuses méthodes ont été développées sur cette base comme SPIEC-EASI (Kurtz et al., 2015), gCODA (H. Fang et al., 2017) ou CD-trace (Yuan et al., 2019).

L'utilisation de copule gaussienne permet d'élargir le cadre gaussien à d'autres distributions (Anderson, Valpine et al., 2019 ; Popovic et al., 2019). Dans ce cadre, il est d'ailleurs possible d'utiliser des données mixtes : binaires, ordinales ou continues (Dobra et Lenkoski, 2011 ; Abegaz et Wit, 2015).

Il est enfin notable qu'il est possible de construire des modèles hiérarchiques pour aborder la reconstruction de réseaux. Ce sont des modèles flexibles permettant de décomposer la complexité des phénomènes biologiques en une série de sous modèles plus simples. Ces modèles se sont démocratisés notamment grâce à la diffusion de logiciels facilitant la modélisation et l'inférence comme Winbugs (Lunn et al., 2000) et JAGS (Plummer, 2003). Il existe différents modèles hiérarchiques bayésiens appliqués à la modélisation des données métagénomique (Yang et al., 2017 ; Ovaskainen et al., 2017 ; Björk et al., 2018). Les procédures d'estimation des paramètres peuvent être gourmandes en ressources informatiques et la complexité augmente rapidement avec le nombre de variables. Le modèle hiérarchique Poisson log-normal (PLN) est optimisé pour répondre à cette complexité (Biswas et al., 2016 ; Chiquet, Mariadassou et al., 2018).

1.3 Présentation du travail de thèse

Dans ce contexte, le sujet de ma thèse était d'étudier le potentiel des approches réseaux dans la détection d'associations entre OTUs au sein de données métagénomiques. Une attention particulière devait être portée sur les agents pathogènes à travers l'étude des interactions microbiennes qui influencent leur dynamique. Dans les trois chapitres suivants, dont deux sont etayés par des manuscrits acceptés ou déjà soumis pour publication, je présente les travaux effectués sur ce sujet.

Dans le cadre de premiers travaux, j'ai étudié les pratiques et les outils d'analyse utilisables pour inférer des réseaux d'association au sein de métagénomes. J'ai déterminé, compte-tenu des propriétés des données métagénomiques, leur efficacité et leurs limites, la nature des informations permettant l'identification d'associations. Ces travaux m'ont permis de déterminer des pistes pour améliorer l'étude des associations microbiennes.

Sur la base des connaissances accumulées, j'ai développé un package d'analyse des associations entre OTUs visant à inférer les associations pertinentes au sein de métagénomes. L'outil développé prend en compte les spécificités des données métagénomiques et offre la possibilité de considérer l'effet de facteurs structurants sur la distribution de l'abondance des OTUs. Ce package s'avère particulièrement efficace par rapport aux outils courrament utilisés dans le domaine.

À travers la participations dans différents projets de métagénomique, j'ai pu confirmer la pertinence de l'outil développé et identifier des pistes d'améliorations permettant de faire face aux problématiques biologiques actuelles. Les bases techniques qui permettaient les évolutions envisagées sont présentées. Enfin, je consacre le dernier chapitre de cette thèse à une discussion des résultats obtenus qui se concentre sur les principaux problèmes rencontrés dans le cadre des études que j'ai conduites.

Chapitre 2

À la recherche d'associations fiables

Les communautés microbiennes contiennent des milliers d'unités taxonomiques opérationnelles (OTUs) dont la plupart sont rares, entraînant un excès de zéros dans les données. Cette caractéristique de la communauté peut entraîner des difficultés méthodologiques : des simulations ont montré que les méthodes de détection d'associations par paires d'OTUs donnent des résultats problématiques. Lorsqu'il existe une forte proportion de zéros dans une table d'OTUs, les performances des outils de détection d'associations sont altérées.

Notre objectif était de comprendre l'impact de la rareté des OTUs sur la détection des associations. En fonction de la proportion de zéros dans les données, nous avons exploré la capacité des statistiques communes à identifier les associations, la sensibilité des mesures d'associations alternatives et la performance des outils d'inférence de réseau.

En étudiant à l'aide de développements mathématiques et de simulations, j'ai pu constater qu'une grande partie des associations, en particulier des associations négatives, ne peuvent pas être testées de manière fiable. Cette contrainte entrave l'identification d'agents biologiques candidats pouvant être utilisés pour lutter contre les agents pathogènes rares. L'identification d'associations testables pourrait servir de méthode objective pour filtrer les jeux de données au lieu des approches empiriques actuelles. Cette stratégie de filtrage pourrait réduire considérablement les temps de calcul et la qualité de l'inférence de réseaux. Différentes possibilités pour améliorer l'analyse

des associations au sein du microbiote sont discutées.

En plus de la publication exposée dans les pages suivantes, j'ai présenté ces travaux à l'oral à deux occasions. Au tout début de ma thèse, j'ai présenté les premières étapes de ce travail dans le cadre du Colloque Apprentissage de réseaux : de la théorie aux applications en biologie et écologie (2016, Toulouse). A l'issue des travaux, j'ai présenté les résultats obtenus dans le cadre de la conférence internationale Pathobiome 2018 qui s'est déroulée à Ajaccio.

PLOS ONE

RESEARCH ARTICLE

Rarity of microbial species: In search of reliable associations

Arnaud Cougoul*, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui

UMR Epidemiology of Animal and Zoonotic Diseases, Université Clermont Auvergne, INRA, VetAgro Sup, Saint-Genès-Champanelle, France

* arnaud.cougoul@inra.fr

OPEN ACCESS

Citation: Cougoul A, Bailly X, Vourc'h G, Gasqui P (2019) Rarity of microbial species: In search of reliable associations. PLoS ONE 14(3): e0200458. https://doi.org/10.1371/journal.pone.0200458

Editor: Hauke Smidt, Wageningen University, NETHERLANDS

Received: June 7, 2018

Accepted: February 28, 2019

Published: March 15, 2019

Data Availability Statement: All microbiota data used in the article are available in the Qiita database (https://qiita.ucsd.edu) and on the TARA Ocean project web page (http://ocean-microbiome.embl.de). All data are described in Table 2, Part 5 of S1 Appendix. * Arctic freshwater systems (EBI Study Accession: PRJEB15630, ID Qiita 1883) * Gut bacteria of Peruvian rainforest ants (EBI Study Accession: PRJEB15630, ID Qiita 10343) * Honeybees from Puerto Rico (EBI Study Accession: PRJEB14927, ID Qiita 1064) * Soil from California vineyards (EBI Study Accession: PRJEB15630, ID Qiita 10082) * The Global Sponge Microbiome (DOI: 10.1038/ncomms11870, ID

Abstract

The role of microbial interactions in defining the properties of microbiota is a topic of key interest in microbial ecology. Microbiota contain hundreds to thousands of operational taxonomic units (OTUs), most of them rare. This feature of community structure can lead to methodological difficulties: simulations have shown that methods for detecting pairwise associations between OTUs, which presumably reflect interactions, yield problematic results. The performance of association detection tools is impaired when there is a high proportion of zeros in OTU tables. Our goal was to understand the impact of OTU rarity on the detection of associations. We explored the utility of common statistics for testing associations; the sensitivity of alternative association measures; and the performance of network inference tools. We found that a large proportion of pairwise associations, especially negative associations, cannot be reliably tested. This constraint could hamper the identification of candidate biological agents that could be used to control rare pathogens. Identifying testable associations could serve as an objective method for filtering datasets in lieu of current empirical approaches. This trimming strategy could significantly reduce the computational time needed to infer networks and network inference quality. Different possibilities for improving the analysis of associations within microbiota are discussed.

Introduction

Microbiota play key roles in ecosystem processes, from eukaryote physiology [1] to global biogeochemical cycles [2]. Research often focuses on comparing microbiota found in similar environments to identify the major forces shaping their structure [3] and function [4]. Microbial interactions are probably one such force [5, 6].

The most common technique for describing microbiota is 16S rRNA sequencing [7]. Association network analysis is then often employed to characterize potential microbial interactions [8]. Such analyses require identifying pairwise associations between the occurrence or abundance of bacterial operational taxonomic units (OTUs) [9]. However, microbiota frequently contain hundreds to thousands of OTUs, most of them rare [10–12]. Consequently, a typical matrix describing the abundance of OTUs among similar microbiota will include a high proportion of zeros. Simulations have illustrated that an excess of zeros impairs the

PLOS ONE | Rarity of microbial species: In search of reliable associations

Qiita 1740) * Tree leaves (DOI:10.1111/j.1462-2920.2010.02258.x, ID Qiita 396) * HMP healthy human (DOI:10.1038/nature11234, ID Qiita 1928) * TARA Ocean Project (DOI: 10.1126/science.1261359) http://ocean-microbiome.embl.de/data/miTAG.taxonomic.profiles.release.tsv.gz.

Funding: The work is funded by the French National Institute for Agricultural Research (http://institut.inra.fr/en). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

efficiency of association network analysis [13, 14]. To avoid this problem, rare OTUs are filtered out beforehand. Current trimming procedures are empirical in nature and restrictive. They may rely on OTU prevalence [13, 15], mean abundance [16], or diversity [17]. Moreover, simulations have found that association network analyses more efficiently detect negative relationships (i.e., amensal, competitive) than positive relationships (i.e., mutualistic, commensal) [13]. It is not yet clear whether this result is due to the distribution of OTU prevalence.

Precisely defining the conditions under which positive and negative associations can be reliably tested should improve current research on microbial interactions. This approach could help design studies that have adequate statistical power; identify potential paths for improving data analysis; and, accounting for its constraints, clarify the interpretation of association network analyses.

Below, we analyzed the effect of low OTU prevalence, a common pattern in real microbiota, on association measures calculated from occurrence data and read abundance data. More specifically, we theoretically and empirically calculated the extrema of common prevalence-based association measures. These extrema were used to define which OTU associations could be reliably tested. We investigated whether alternative association measures and cutting-edge association network analysis tools were also affected by low OTU prevalence. This methodological strategy allowed us to (i) define the extent to which prevalence and sample size affect the results of microbiota association analyses; (ii) demonstrate that negative associations cannot be captured in most cases; and (iii) show that there is little added value obtained from analyzing abundance data as compared to occurrence data. We discuss our findings in the context of current analytical procedures and tools with a view to proposing potential solutions to the issues we identified.

Materials and methods

Methods for detecting associations among microbes have progressed rapidly as the to obtain microbiota data has become more widespread. Here, we determined how an excess of zeros affected classical correlation measures by examining the latter's testability. We also considered alternative association measures and explored the relationship between method type and association network inference quality.

Prevalence affects the distribution of association statistics, which can lead to problems with the testability of correlation coefficients. For instance, a statistic's minimum and/or maximum can fall within the expected confidence interval obtained from the classical distributions used to approximate expected values. This issue can arise with both occurrence data and abundance data.

Model for occurrence data: Fisher test and Phi coefficient

First, we explored how to define testability when occurrence data are used. Co-occurrence networks are commonly reconstructed using the hypergeometric law that underlies Fisher's exact test [9, 18, 19]. For fixed prevalence values, the probability of observing the minimum or maximum number of co-occurrences may be higher than the alpha level (traditionally set to 5%) [20, 21]. In such a case, neither negative nor positive associations, respectively, can be significantly detected. Limits on testability can be studied by enumerating all the possible combinations of associations based on prevalence (detailed in Part B.7 in S1 Appendix). The combinatorics that ensue from the hypergeometric law provide numerical solutions for determining association testability.

The Phi coefficient [22] can be used to establish equations for exploring association testability, which provide an analytical solution. The Phi coefficient ϕ is a measure of association

between two binary variables X_A and X_B.

$$\phi = \sqrt{\frac{P_{11} - P_A P_B}{P_A (1 - P_A) P_B (1 - P_B)}}, \tag{1}$$

where P_A, P_B are the prevalence values for two OTUs, X_A and X_B, and P_{11} is the prevalence of their co-occurrence. The prevalence of an OTU is

$$\text{prevalence} = \frac{\text{number of non} - \text{zero samples}}{\text{total number of samples}}. \tag{2}$$

The extrema of Phi [23] depend exclusively on P_A and P_B (S1 Fig and Part B in S1 Appendix).

$$\begin{aligned} \min(\phi) &= \max\left(-\sqrt{\frac{P_A P_B}{(1 - P_A)(1 - P_B)}}, -\sqrt{\frac{(1 - P_A)(1 - P_B)}{P_A P_B}}\right) \\ \max(\phi) &= \min\left(-\sqrt{\frac{P_A(1 - P_B)}{P_B(1 - P_A)}}, -\sqrt{\frac{P_B(1 - P_A)}{P_A(1 - P_B)}}\right) \end{aligned} \tag{3}$$

Under the null hypothesis (H_0) that the occurrences of X_A and X_B are independent, Phi can be approached thanks to Pearson's chi-squared test:

$$\phi^2 = \frac{\chi^2}{N}, \tag{4}$$

where N is the total number of samples and χ^2 is a chi-squared distribution with one degree of freedom [24]. This latter distribution is thus used to build a confidence interval with which to test departure from the null hypothesis. Furthermore, we can describe cases where it would be impossible to reliably test associations based on this confidence interval because the genuine minimum and/or maximum of ϕ fall within the confidence interval.

Model for read abundance data: Pearson and Spearman correlations

Second, we explored how to define testability when read abundance data are used. We first employed the Pearson correlation coefficient [25], which is a measure of association between two continuous variables, X_A and X_B.

$$r = \frac{\mathrm{E}(X_A X_B) - \mathrm{E}(X_A)\,\mathrm{E}(X_B)}{\sigma_{X_A}\,\sigma_{X_B}} \tag{5}$$

We demonstrated that the minimum of the Pearson correlation coefficient depends only on OTU prevalence (see the proof in Part C in S1 Appendix and the illustration in S2 Fig).

$$\min(r) = -\sqrt{\frac{P_A P_B}{(1 - P_A)(1 - P_B)}} > -1, \quad \text{if } P_A + P_B < 1 \tag{6}$$

We can then define a confidence interval based on the following assumption: if X_A and X_B follow two uncorrelated normal distributions,

$$r = \frac{t}{\sqrt{N - 2 + t^2}} \tag{7}$$

where t has a Student's t-distribution with degrees of freedom $N - 2$.

We demonstrated that the result for the correlation minimum (Eq (6)) is identical for the Spearman correlation approach (Part C.7 in S1 Appendix). The Spearman correlation is the

Pearson correlation applied to the ranks of X_A and X_B. The Spearman correlation coefficient follows the same expected distribution described by Eq (7) when X_A and X_B are independent. This fact makes it possible to relax the assumption of normality of the Pearson correlation test, a hypothesis not respected in the analysis of the microbiota data.

To estimate the proportion of unreliable tests, we considered two distributions for OTU prevalence: (i) a uniform law, to study the influence of sample size N and prevalence P_A, P_B and (ii) a truncated power law, to take into account the real structure of microbiota data. We also compared the results for the testability limits for the two types of data and highlighted a correlation between the two associated measures.

Simulated responses of association measures

We found that, theoretically, OTU prevalence has an impact on the observable minimum Pearson and Spearman correlation coefficients. We therefore explored the behavior of alternative association measures. We analyzed the relationship between OTU prevalence and the values of five measures used to infer association networks: Pearson and Spearman correlation coefficients, Bray Curtis dissimilarity, mutual information, and the maximal information coefficient (MIC) [9, 26]. Bray Curtis dissimilarity is an ecological statistic that we employed here to quantify compositional dissimilarity between OTUs. Mutual information and the MIC are two measures that were developed from information theory. Both are used to capture nonlinear or non-monotonic relationships. We generated two correlated variables to analyze the responses of the association measures. The zero-inflated negative binomial (ZINB) distribution appears to best fit microbiota data [27, 28]. We generated a bivariate normal sample of size $N = 50$ and simulated three correlation levels: a negative correlation ($r = 1$), a positive correlation ($r = 1$), and a null correlation ($r = 0$), which served as a reference. The copula theory allows normally distributed data to be marginally transformed into ZINB-distributed data [29]. OTU prevalence was modeled using the probability of structural zeros. For the ZINB distribution, dispersion was 0.5, and the mean was 1000. This situation corresponded to two OTUs of high abundance. Prevalence values ranged from 0.05 to 0.95 in 0.05 steps. We calculated the value of each association measure for all possible pairs of prevalence. We conducted 100 simulations and retained the median value for each prevalence pair.

Association network analysis tools

We studied the relationship between OTU prevalence and the quality of inference provided by association network analysis tools. Three inference tools were studied: CoNet [30], SPIEC-EASI [15], and SparCC [16]. We simulated datasets containing 50 samples and 100 OTUs. The data followed a multivariate normal distribution and contained with 100 known associations, of which half were positive and half were negative. From the adjacency matrix, we calculated a correlation matrix where the target matrix condition was 100, as described in [15]. Using the copula theory, we then transformed the normally distributed data into ZINB-distributed data [29]. Prevalence was modeled using the probability of structural zeros. All the OTUs had the same prevalence, which was the variable study parameter. For the ZINB distribution, dispersion was 0.5, and the mean was 1000. Finally, we used the different tools to infer the association network and measured tool ability to pick up on positive or negative associations. We independently examined the proper classification of negative associations and positive associations. Inference quality was assessed based on the area under the ROC curve (AUC) and the area under the precision-recall curve (AUPRC) [31].

Data filtering before association network inference

We analyzed the effect of data filtering methods on network inference quality. We simulated datasets containing 300 samples and 300 OTUs following a ZINB distribution, as described in the previous paragraph. The datasets contained 1000 associations, half positive and half negative. As above, the target matrix condition was 100. OTU prevalence followed a power law distribution where $k = -1.5$. Minimum prevalence was 5/300 to avoid simulating a situation in which OTUs were missing from all 300 samples, which would not be taken into account in network inference. For the ZINB distribution, dispersion was 0.5, and the mean was 1000. We implemented data filtering in CoNet and SPIEC-EASI (S2 File). For CoNet, we did not compute the p-values of the problematic pairs we identified. For SPIEC-EASI, after normalizing the data with the centered log-ratio (clr) transformation [32], we assigned a zero weight to the problematic pairs during the graphical lasso estimation [33], which corresponded to a strong regularization parameter for these pairs. SparCC's algorithm did not allow problematic pairs to be filtered. To generate benchmarks for data filtering, we inferred association networks under three different conditions: (i) for all OTU pairs; (ii) for fully testable pairs only (i.e., after removing problematic pairs; alpha level of 5%); and (iii) for OTUs that had been filtered based on a prevalence threshold. In this latter case, the goal was to be able to compare the results of filter based on testability with those obtained using a conventional filter based on prevalence. To do this, we removed enough low prevalence OTUs to have at least the same number of filtered pairs as in (ii). We performed 20 simulations of each. We then measured the AUC values associated with network inference. Inference quality was based only on the associations that remained after filtering.

Results

Testability given a uniform prevalence distribution

When occurrence data were used, four inequations (Eqs (7-10) in S1 Appendix) defined reliable tests based on the chi-squared distribution and OTU prevalence (Fig 1C). The proportion of non-testable associations (i.e., neither positive nor negative correlations could ever be significant) rapidly fell as N increased (Fig 1D). The proportion of associations with partial testability (i.e., either only positive or negative correlations could ever be significant) never exceeded 0.25 (Fig 1D). When $N = 300$, the proportion of fully testable associations (both positive and negative correlations could be significant) exceeded 0.80 (Fig 1D). We showed numerically that there was consistency between the proportion of Fisher's exact tests affected by prevalence and the analytical results (Fig 1A and 1B). There were slightly more non-testable associations when Fisher's exact test was used, as compared to the Phi coefficient, and slightly fewer associations with partial testability.

When read abundance data were used, some negative correlations were not testable based on the Student's distribution (Eq (33) in S1 Appendix and Fig 1E). This problem became less pronounced as N increased, and the proportion of testable associations reached 0.95 at $N = 300$ (Fig 1F).

Testability given realistic community structure

Prevalence distributions are highly unbalanced in microbiota because of the large number of rare OTUs (Fig 2A). Accordingly, we modeled OTU prevalence using a truncated power law distribution; the latter reflects observed community structure (Part E in S1 Appendix and S3 Fig). OTU prevalence was fitted according to a truncated power law, with k ranging from −2

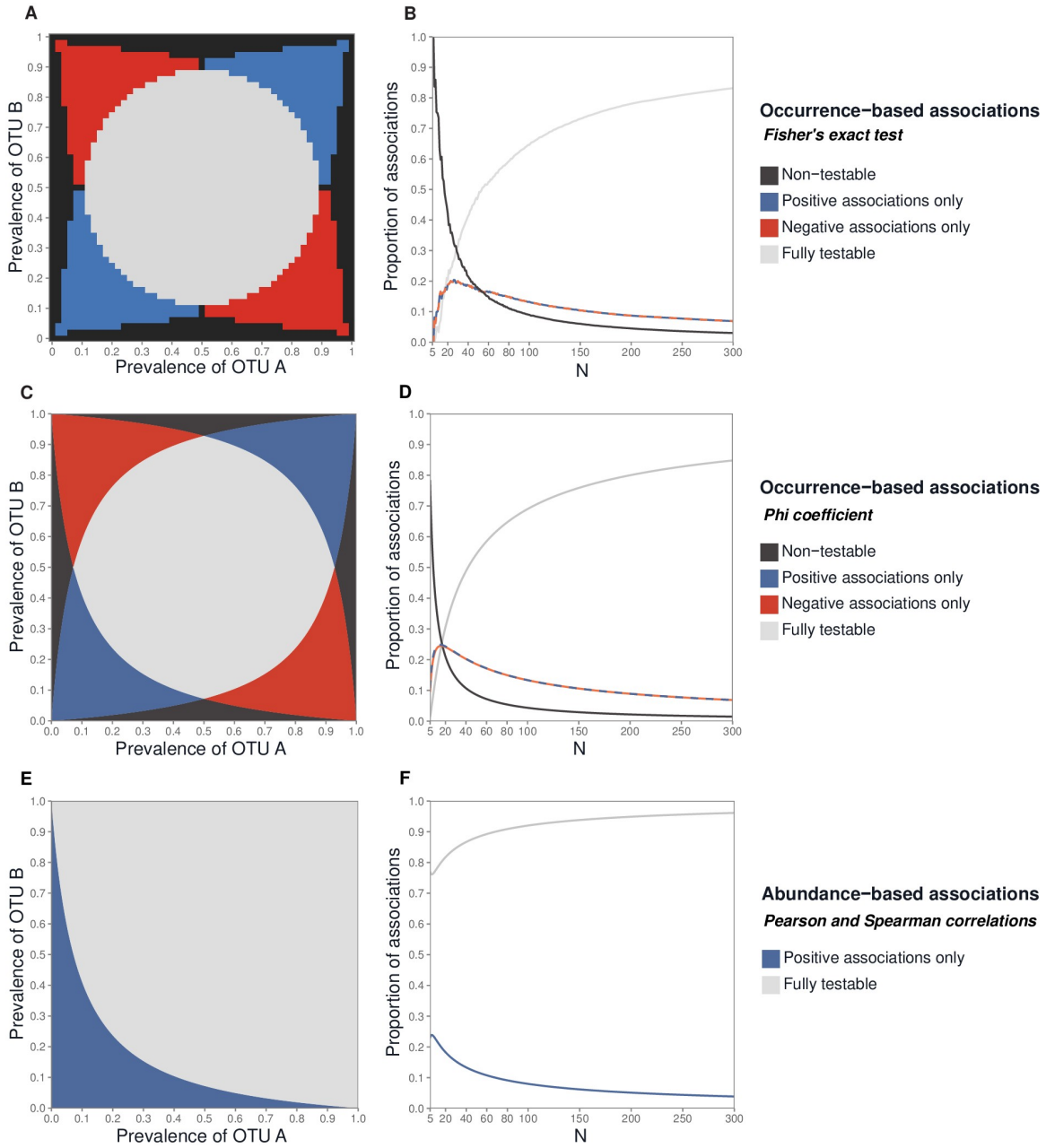


Fig 1. Testability of pairwise associations for the occurrence data and for the read abundance data. For the occurrence data: the testability zones defined by OTU prevalence for the Fisher's exact test (**A**), and the proportion of testable associations as a function of *N* assuming prevalence follows a uniform distribution (**B**). Testability zones defined by OTU prevalence for the Phi coefficient (**C**), and the proportion of testable associations as a function of *N* assuming prevalence follows a uniform distribution (**D**). For the read abundance data: testability zones defined by OTU prevalence for the Pearson and Spearman correlation coefficients (**E**), and the proportion of testable associations as a function of *N* assuming prevalence follows a uniform distribution (**F**). The alpha level for the tests was 5%. For (A), (C) and (E), $N = 50$.

https://doi.org/10.1371/journal.pone.0200458.g001

to −1: the smaller the *k*, the higher the proportion of rare species. The use of such a distribution means that, for most OTU pairs, both OTUs had a low prevalence (Fig 2B).

For the occurrence data, there was thus a large proportion of associations for which negative correlations could never be significant (> 0.50 for $k = -1$, > 0.90 for $k = -2$); this

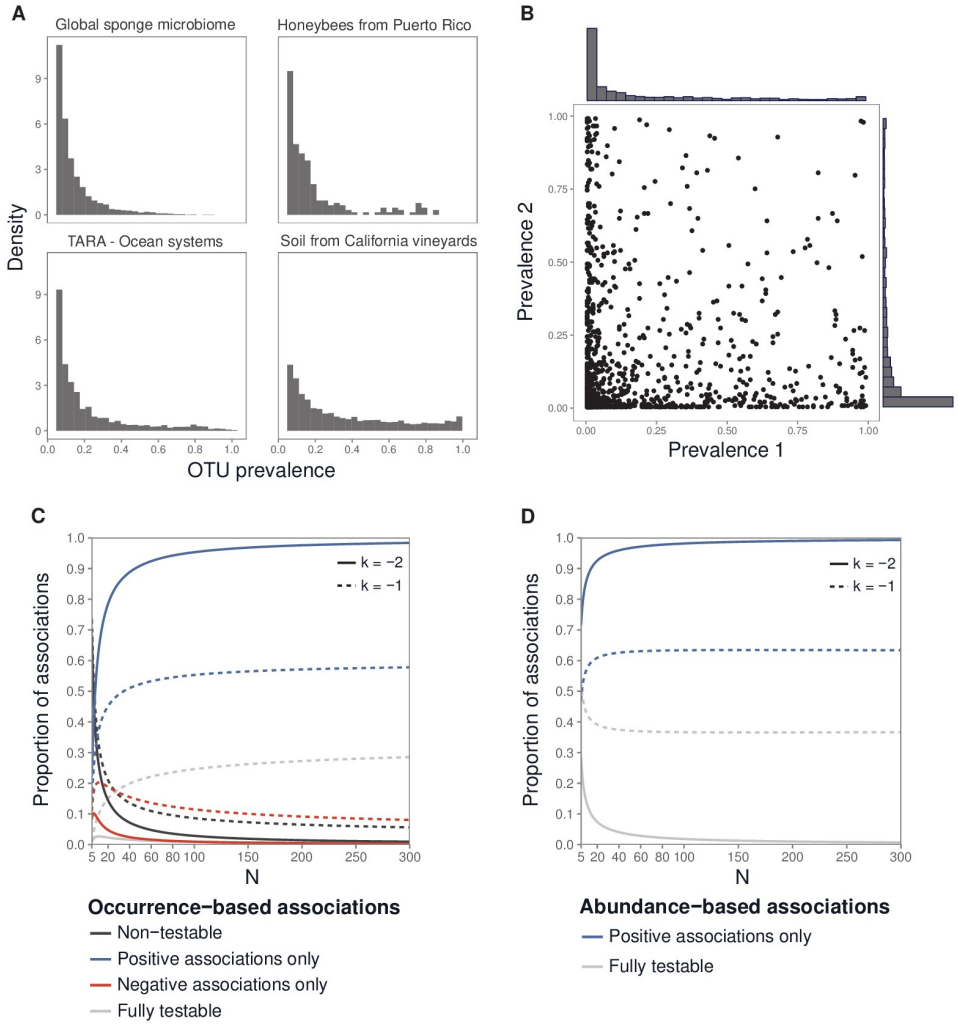


Fig 2. Association testability under realistic conditions of microbial community structure. (**A**) Histograms of OTU prevalence in several microbiota characterized by 16S rRNA sequencing. Data were taken from the Qiita database [34] and the TARA Ocean Project [35]. The microbiota are described elsewhere (Part E in S1 Appendix). To better illustrate the skewed distributions, only prevalence values of greater than 5% were included. (**B**) Distribution of all pairs of OTU prevalences from microbiota data for soil from California vineyards. Each point represent a pair of OTU prevalences. Proportion of testable associations as a function of N when $k = -2$ or -1 for the occurrence data (**C**) and the read abundance data (**D**).

https://doi.org/10.1371/journal.pone.0200458.g002

proportion increased as N increased (Fig 2C). This counter-intuitive result is due to the accumulation of rare OTUs as N increases under the power law assumption. Fewer than 10% of associations were non-testable when N was greater than 50 (Fig 2D).

For the read abundance data, when $N = 100$, a large proportion of negative correlations were non-testable when $k = -1$ (proportion: 0.60) and $k = -2$ (proportion: 0.95) (Fig 2D).

Comparison between the two data types

We compared the association statistics for both data types under conditions of low OTU prevalence such as those observed in actual microbiota data (Part D in S1 Appendix). A formal decomposition of variance and covariance illustrates the structural relationship of the

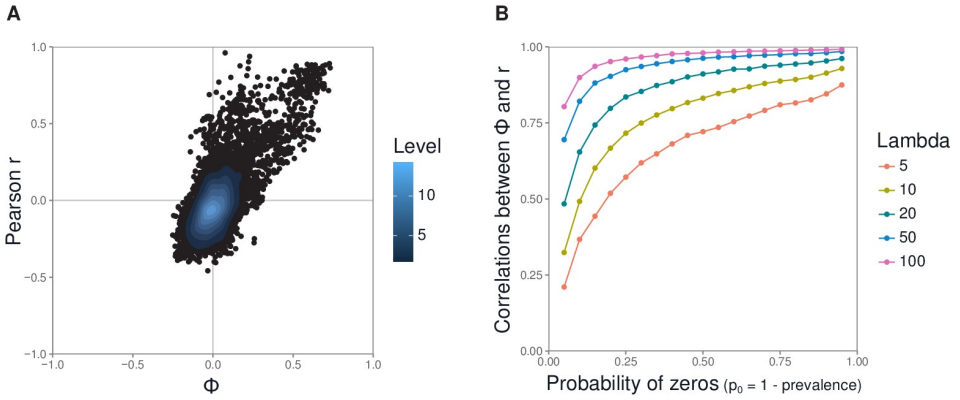


Fig 3. Correlation between the Phi coefficient and the Pearson coefficient. (**A**) Correlation in honeybees microbiota data (Part E in S1 Appendix). Each point corresponds to the association coefficients for an OTU pair. Read abundance data were normalized using clr. (**B**) Correlation computed from simulations of OTU abundances modeled using a zero-inflated Poisson (ZIP) distribution (Part D.2 in S1 Appendix). The parameters are the probability of structural zeros, p_0, and the value of the Poisson parameter, λ. In biological terms, the probability of structural zeros corresponds to the prevalence (prevalence = $1 - p_0$), and the Poisson parameter corresponds to read abundance. For each pair of p_0 and λ values, we generated 100 samples of two hypothetical OTUs, X_A and X_B, whose abundances followed a ZIP distribution with those parameter values. We then calculated ϕ and r for the samples. The correlation between the two coefficients was assessed by repeating this process 10^5 times.

https://doi.org/10.1371/journal.pone.0200458.g003

correlation coefficients calculated from the occurrence data and the read abundance data (Eq (2), Part A in S1 Appendix). The observed values of the Phi coefficient ϕ and the Pearson correlation coefficient r for OTU pairs in microbial datasets (Fig 3A) illustrated that the minimum of the statistics is particularly affected as explained above. Furthermore a correlation was observed between the two measures for real microbiota datasets (cor = 0.78 and $R^2 = 0.62$ for honeybees microbiota data, Fig 3A). Simulations allowed us to delve deeper into the expected correlation between the two measures. The association tests that can be performed using occurrence data versus read abundance data tend to be similar, and prevalence influences association testability in the same way. More specifically, association measures for the two data types become correlated as prevalence decreases (Fig 3B).

Impact of OTU prevalence on other association measures

We studied the relationship between OTU prevalence and the responses of five common association measures (Fig 4) using simulated data. There were differences in the abilities of the measures to capture negative associations. The Pearson correlation coefficient did a poor job of picking up on negative associations. The Spearman correlation coefficient did better: it was able to pick up on negative associations. OTU prevalence had a strong effect on the Spearman correlation coefficient, as noted above. Bray Curtis dissimilarity and mutual information did a poor job of capturing negative associations: their responses for the dataset containing the associations were the same as their responses for the null dataset. The MIC responded well, especially when prevalence was high. The Spearman correlation coefficient and the MIC were the only measures that could properly capture negative correlations, but they were nonetheless affected by low prevalence.

In the case of the positive associations, all five measures showed a greater degree of sensitivity. However, OTU prevalence still exerted an influence, even if it was less dramatic than for negative associations. For the negative associations, measures were altered when the sum of the two prevalences decreased, along the first bisector. For the positive associations, measures

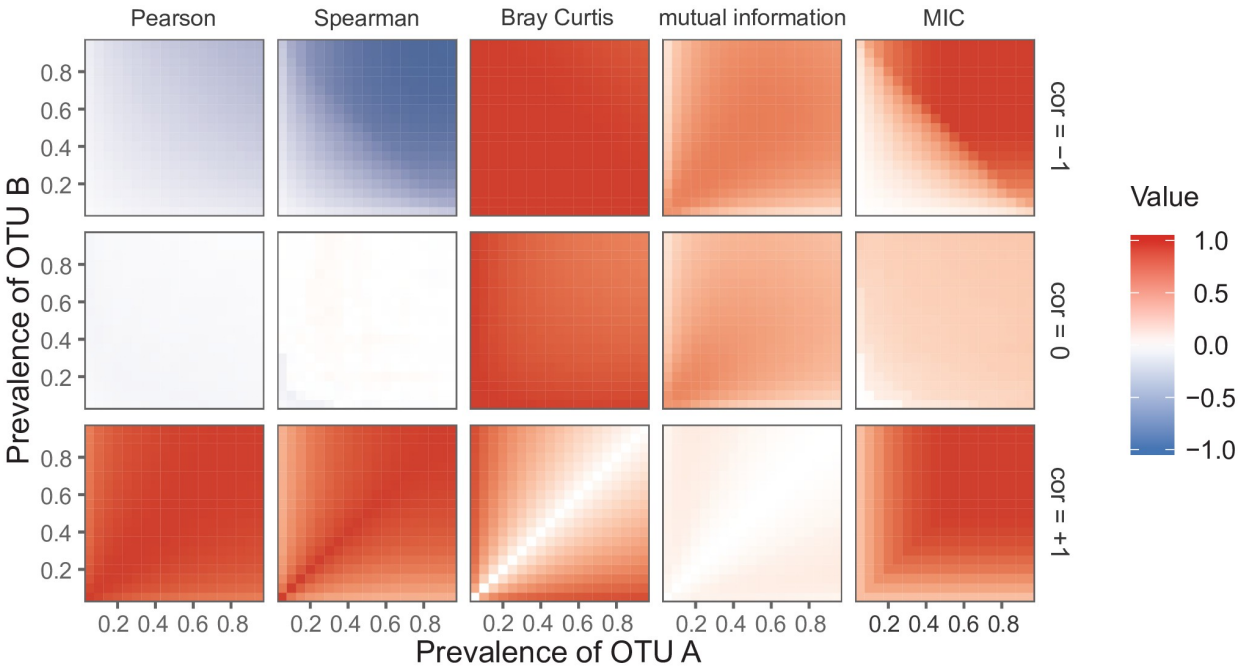


Fig 4. Relationship between OTU prevalence and the responses of five association measures for a simulated dataset. Two zero-inflated negative binomial (ZINB) distributions ($N = 50$, $\mu = 1000$, $\theta = 0.5$, $p_0 = 1 - \text{prevalence}$) were created using all pair of prevalences from 0.05 to 0.95 (steps of 0.05) and for three correlation levels. For the graphs, the correlation level is −1 in the first row, 0 in the second row, and +1 in the third row. The five association measures are represented in different columns. A total of 100 simulations were performed, and the median values were plotted.

https://doi.org/10.1371/journal.pone.0200458.g004

were affected when one of prevalences decreased, along the prevalence axes. Consequently, the mechanisms that limit the ability to measure positive associations are different from those tied to negative associations.

Impact of prevalence on network inference quality

We compared the ability of three recently developed tools to infer association networks within simulated microbiota data: all three had difficulties detecting associations when faced with a high proportion of zeros (i.e., low OTU prevalence; Fig 5). Positive associations were easier to detect, but low prevalence still had an effect. Examining the characteristics of the ROC curves, limitations occurred at a prevalence level of 0.2. When paired OTUs had prevalences below this level, they fell completely within a zone of partial testability, where only positive associations could be tested (compare with Fig 1E).

Impact of data filtering on association network inference quality

We analyzed the effect of filtering data on the quality of association network inference (Fig 6) using simulated data. In our dataset, problematic pairs (at an alpha level of 5%) represented, on average, 70% of the total number of associations. During prevalence-based filtering, we removed the less prevalent OTUs, with a view to eliminating the same proportion of associations as in testability-based filtering.

The results obtained with CoNet and SPIEC-EASI were quite similar. When the data were unfiltered, negative associations were less well recovered than were positive associations, as mentioned previously. Under these conditions, the AUC values were below 0.8. When the

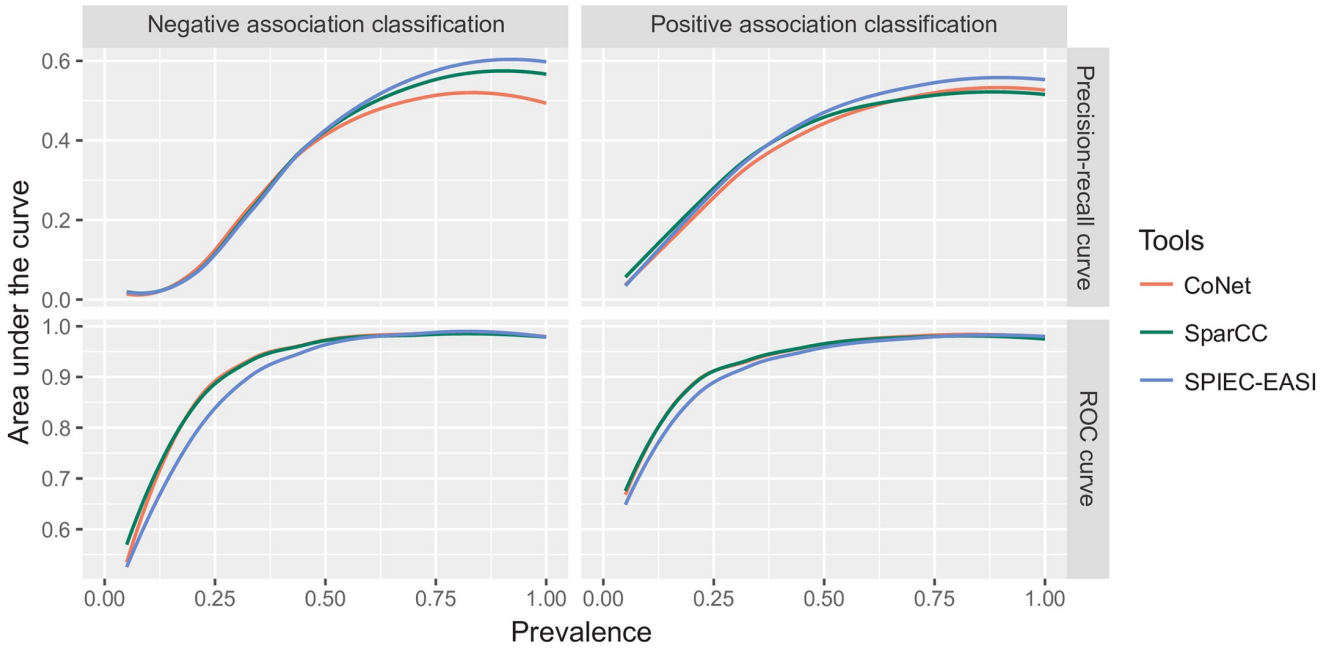


Fig 5. Performance of three association network analysis tools as a function of OTU prevalence. Datasets of 100 OTUs were generated using a ZINB distribution ($N = 50$, $\mu = 1000$, $\theta = 0.5$, $p_0 = 1 - \text{prevalence}$). A covariance structure was imposed on the datasets-there were 100 associations, of which half were positive and half were negative. All OTUs had the same prevalence, which varied from 0.05 to 1 in 20 log steps. For each prevalence value, 20 simulations were performed. The plots show the means of a LOESS regression. The left-hand graphs represent the classification of the negative associations, and the right-hand graphs represent the classification of the positive associations. The top and bottom graphs show the AUPRC and AUC values, respectively.

https://doi.org/10.1371/journal.pone.0200458.g005

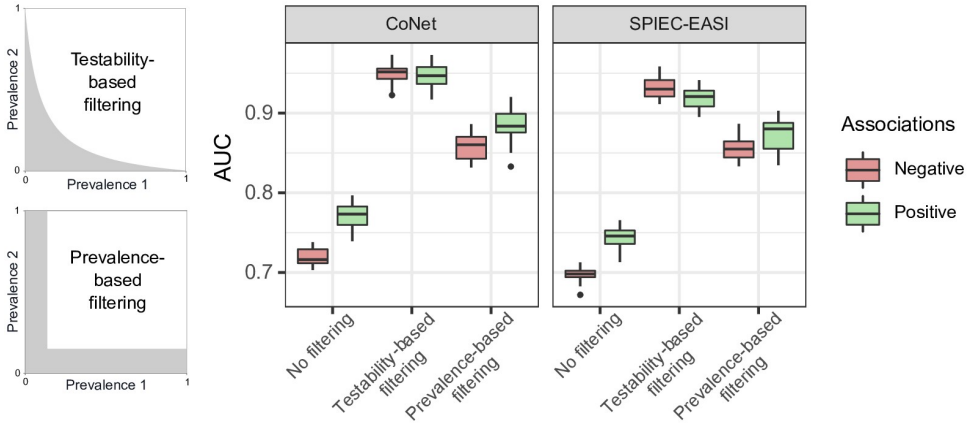


Fig 6. Impact of data filtering on association network inference quality. Performance of CoNet and SPIEC-EASI depending on data filtering methods: no filtering, testability-based filtering, and prevalence-based filtering. In testability-based filtering, problematic associations were removed (alpha level of 5%). In prevalence-based filtering, the lowest-prevalence OTUs were removed to obtain the same number of filtered associations as for testability-based filtering. Datasets of 300 OTUs were generated using ZINB distributions ($N = 300$, $\mu = 1000$, $\theta = 0.5$, $p_0 = 1 - \text{prevalence}$). Prevalences were simulated using a power law distribution where $k = -1.5$. A covariance structure was imposed on the datasets: there were 1000 associations, half positive, half negative. The target matrix condition was 100. A total of 20 simulations were performed to obtain the boxplots of the areas under the ROC curve.

https://doi.org/10.1371/journal.pone.0200458.g006

data were filtered, the quality of inference improved. When the data were filtered by testability, the AUC values for both negative and positive associations were greater than 0.9. Furthermore, the AUC values for negative associations were the same as the AUC values for positive associations (and were even higher when SPIEC-EASI was used). When prevalence-based filtering was used, the AUC values were lower. For our simulated dataset, testability-based filtering thus yielded better results than the more common, prevalence-based filtering procedure.

Network inference could be carried out significantly faster when the data were filtered. The mean calculation times for unfiltered, testability-filtered, and prevalence-filtered data were as follows: 122, 72, and 19 seconds, respectively, for SPIEC-EASI and 2041, 667, and 661 seconds, respectively, for CoNet.

Discussion

We showed that it is impossible to reliably test a large proportion of the pairwise associations between OTUs in microbiota using classical association measures and common association network analysis tools. Indeed, in our simulations employing realistic community structure (i.e., most OTUs are rare), we discovered the following: (i) correlations, especially negative correlations, could not be tested for most associations using classical statistics; (ii) alternative association measures was also affected by low OTU prevalence; and (iii) cutting-edge network analysis tools also struggle when OTU prevalence is low. These findings clarify previous modeling results [13] and underscore a major analytical challenge in this domain. This issue cannot be solved via the use of statistics adapted to non-linear relationships, the permutation and bootstrap tests proposed by CoNet, or the clr transformation procedure employed by SPIEC-EASI. It also has important practical implications. For example, this constraint could hamper the identification of candidate biological agents that could be used to control rare pathogens.

We defined equations that can be used to quickly identify *a priori* whether OTU associations can be tested. Applying stringent standards (i.e., analyzing only fully testable associations) drastically reduced the number of tests required to infer an association network. We propose a way to implement this filtering strategy in CoNet and SPIEC-EASI: by assuming there is no association for problematic pairs in the correlation matrix of OTU abundances when an association network is being inferred. By limiting test number, the time needed for network inference was drastically reduced. We showed that identifying testable associations could serve as an alternative to current, empirical strategies for filtering microbiota datasets. Indeed, we found that inference quality may be better if data are filtered to remove problematic pairs of OTUs rather to remove low-prevalence OTUs.

We found that association testability tended to be similar for occurrence data and read abundance data. More specifically, association measures calculated using the two data types became correlated as prevalence decreased. This fact raises questions about the information that is actually being captured by current methods for quantifying OTU associations. These questions have both computational implications—it is unclear that current models are able to make the most of abundance data—and biological implications—the two data types could reveal the operation of different biological processes involved in interactions. Zero-inflated distributions can be used to explicitly model occurrence and abundance. They aim to differentiate structural zeros, due to OTU absence, from sampling zeros, due to limited sequencing depth. Since zeros can be ambiguous, presence-absence patterns likely change with sequencing depth. As a result, the minima and maxima of the Pearson correlation coefficient and the Phi coefficient will depend on this depth. Fitting OTU abundances using such distributions appears to be a promising solution for improving the inference of microbial associations [27, 36].

The low prevalence of OTUs in metagenomics datasets greatly limits our ability to carry out broad-scale analyses. Based on the results obtained in this study, we believe that advances in the discovery of microbial associations should be made by systematically integrating available information into the models being used. Initial attempts to develop statistical models that incorporate previous findings into metagenomics analyses have yielded promising results [37]. From a biological point of view, this approach would benefit from the development of a database dedicated to microbial interactions. Open and shared microbiota datasets, like those present on the Qiita collaborative platform (https://qiita.ucsd.edu), could be used to benchmark statistical models, and contributing to such databases could improve our knowledge of microbiota.

Supporting information

S1 Fig. Extrema of the Phi coefficient as a function of OTU prevalence. Minimum (**A**) and maximum (**B**) of the Phi coefficient as a function of prevalence. Computed from Eq (3).
(PDF)

S2 Fig. Minimum of the Pearson correlation coefficient as a function of OTU prevalence. Minimum of the Pearson correlation coefficient r as a function of prevalence. Computed from Eq (6).
(PDF)

S3 Fig. Prevalence structure of real microbial communities. (**A**) Histograms of OTU prevalence in several microbiota characterized by 16S rRNA sequencing. The microbiota are described in Part E in S1 Appendix. (**B**) Probability density function of the same prevalence data (log-log scale), which were fitted to a truncated power law distribution; the power law coefficient k was estimated by maximizing log-likelihood.
(PDF)

S4 Fig. Proportion of testable associations as a function of the power law coefficient *k*. Proportion of testable associations as a function of k when $N = 50$, 100, or 300 for the occurrence data (**A**) and for the read abundance data (**B**).
(PDF)

S1 Appendix. Supplementary material. 1. Notation and decomposition of variance and covariance. 2. Threshold method for binary data. 3. Threshold method for quantitative data. 4. Similarity of the Phi and Pearson correlation coefficients. 5. Distribution of OTU prevalence in real microbiota.
(PDF)

S1 File. R code for carrying out testability-based data filtering. There are two functions, one for each data type: occurrence data and abundance data.
(R)

S2 File. Material used in the simulations. Code files, data files, and result files associated with each of the figures. Implementation of testability-based filtering in CoNet and via the graphical lasso method with clr transformation (SPIEC-EASI-like).
(ZIP)

Acknowledgments

This work was funded by two INRA metaprogrammes: Meta-omics of microbial ecosystems (MEM) and Integrated management of animal health (GISA). We thank Ioana Molnar for the mathematical advice and Jessica Pearce-Duvet for proofreading the manuscript. We also thank

Leo Lahti and the other anonymous reviewer for their constructive comments that helped us improve this work.

Author Contributions

Conceptualization: Arnaud Cougoul, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Data curation: Arnaud Cougoul.

Formal analysis: Arnaud Cougoul.

Funding acquisition: Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Investigation: Arnaud Cougoul, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Methodology: Arnaud Cougoul, Xavier Bailly, Patrick Gasqui.

Project administration: Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Supervision: Xavier Bailly, Patrick Gasqui.

Validation: Arnaud Cougoul, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Visualization: Arnaud Cougoul, Xavier Bailly, Patrick Gasqui.

Writing – original draft: Arnaud Cougoul, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui.

Writing – review & editing: Arnaud Cougoul, Xavier Bailly, Patrick Gasqui.

References

1. Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI. Worlds within worlds: evolution of the vertebrate gut microbiota. Nature Reviews Microbiology. 2008; 6(10):776–788. https://doi.org/10.1038/nrmicro1978 PMID: 18794915
2. Newman DK. Geomicrobiology: How Molecular-Scale Interactions Underpin Biogeochemical Systems. Science. 2002; 296(5570):1071–1077. https://doi.org/10.1126/science.1010716 PMID: 12004119
3. Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, et al. Microbiota and Host Nutrition across Plant and Animal Kingdoms. Cell Host & Microbe. 2015; 17(5):603–616. https://doi.org/10.1016/j.chom.2015.04.009
4. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial ecology: building predictive understanding of community function and dynamics. The ISME Journal. 2016; 10(11):2557–2568. https://doi.org/10.1038/ismej.2016.45 PMID: 27022995
5. Gibson TE, Bashan A, Cao HT, Weiss ST, Liu YY. On the Origins and Control of Community Types in the Human Microbiome. PLOS Computational Biology. 2016; 12(2):e1004688. https://doi.org/10.1371/journal.pcbi.1004688 PMID: 26866806
6. Gonze D, Lahti L, Raes J, Faust K. Multi-stability and the origin of microbial community types. The ISME Journal. 2017; 11(10):2159–2166. https://doi.org/10.1038/ismej.2017.60 PMID: 28475180
7. Konopka A. What is microbial community ecology? The ISME Journal. 2009; 3(11):1223–1230. https://doi.org/10.1038/ismej.2009.88 PMID: 19657372
8. Layeghifard M, Hwang DM, Guttman DS. Disentangling Interactions in the Microbiome: A Network Perspective. Trends in Microbiology. 2017; 25(3):217–228. https://doi.org/10.1016/j.tim.2016.11.008 PMID: 27916383
9. Faust K, Raes J. Microbial interactions: from networks to models. Nature Reviews Microbiology. 2012; 10(8):538–550. https://doi.org/10.1038/nrmicro2832 PMID: 22796884
10. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences. 2016; 113(21):5970–5975. https://doi.org/10.1073/pnas.1521291113
11. Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiology Reviews. 2016; 40(5):686–700. https://doi.org/10.1093/femsre/fuw017 PMID: 27358393
12. Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. The ISME Journal. 2017; 11(4):853–862. https://doi.org/10.1038/ismej.2016.174 PMID: 28072420

13. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal. 2016; 10(7):1669–1681. https://doi.org/10.1038/ismej.2015.235 PMID: 26905627

14. Mainali KP, Bewick S, Thielen P, Mehoke T, Breitwieser FP, Paudel S, et al. Statistical analysis of co-occurrence patterns in microbial presence-absence datasets. PLOS ONE. 2017; 12(11):e0187132. https://doi.org/10.1371/journal.pone.0187132 PMID: 29145425

15. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology. 2015; 11(5): e1004226. https://doi.org/10.1371/journal.pcbi.1004226 PMID: 25950956

16. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology. 2012; 8(9):e1002687. https://doi.org/10.1371/journal.pcbi.1002687 PMID: 23028285

17. Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Frontiers in Microbiology. 2014; 5(MAY):1–14.

18. Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Research. 2010; 20(7):947–959. https://doi.org/10.1101/gr.104521.109 PMID: 20458099

19. Li C, Lim KMK, Chng KR, Nagarajan N. Predicting microbial interactions through computational approaches. Methods. 2016; 102:12–19. https://doi.org/10.1016/j.ymeth.2016.02.019 PMID: 27025964

20. Tarone RE. A Modified Bonferroni Method for Discrete Data. Biometrics. 1990; 46(2):515. https://doi.org/10.2307/2531456 PMID: 2364136

21. Carlson J, Heckerman D, Shani G. Estimating false discovery rates for contingency tables; 2009.

22. Yule GU. On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society. 1912; 75(6):579. https://doi.org/10.2307/2340126

23. Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. Biometrika. 2006; 93(1):197–206. https://doi.org/10.1093/biomet/93.1.197

24. Guilford JP. The phi coefficient and chi square as indices of item validity. Psychometrika. 1941; 6 (1):11–19. https://doi.org/10.1007/BF02288569

25. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 1896; 187:253–318.

26. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, et al. Detecting Novel Associations in Large Data Sets. Science. 2011; 334(6062):1518–1524. https://doi.org/10.1126/science.1205438 PMID: 22174245

27. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Computational Biology. 2014; 10(4):e1003531. https://doi.org/10.1371/journal.pcbi.1003531 PMID: 24699258

28. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017; 5(1):27. https://doi.org/10.1186/s40168-017-0237-y PMID: 28253908

29. Trivedi PK, Zimmer DM. Copula Modeling: An Introduction for Practitioners. Foundations and Trends® in Econometrics. 2006; 1(1):1–111. https://doi.org/10.1561/0800000005

30. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape. F1000Research. 2016; 5:1519. https://doi.org/10.12688/f1000research.9050.1 PMID: 27853510

31. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015; 10(3):e0118432. https://doi.org/10.1371/journal.pone.0118432 PMID: 25738806

32. Aitchison J. The statistical analysis of compositional data: monographs in statistics and applied probability. Chapman & Hall, London. 1986.

33. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9(3):432–441. https://doi.org/10.1093/biostatistics/kxm045 PMID: 18079126

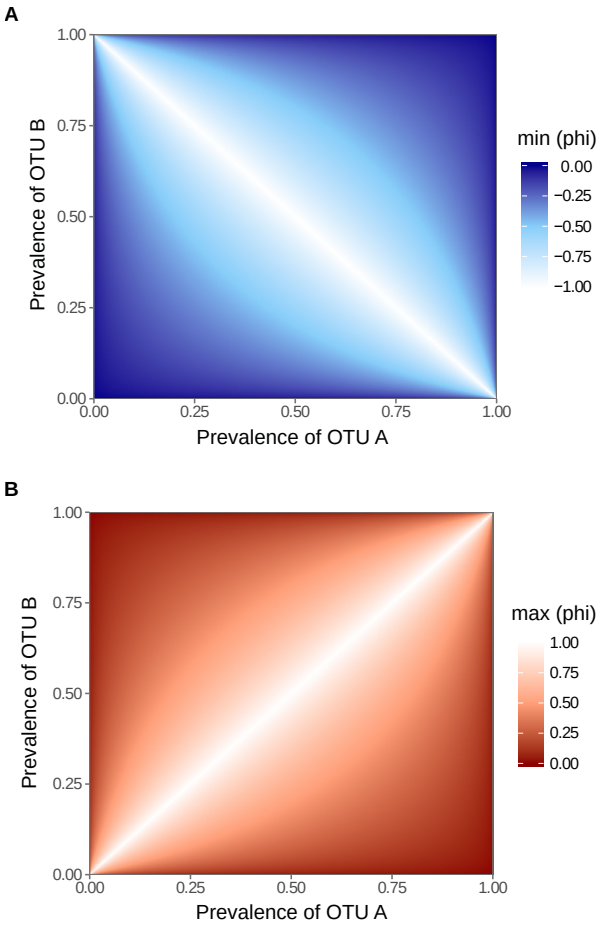
34. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nature Methods. 2018; 15(10):796–798. https://doi.org/10.1038/s41592-018-0141-9 PMID: 30275573

35. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015; 348(6237):1261359–1261359. https://doi.org/10.1126/science.1261359 PMID: 25999513

36. Jonsson V, Österlund T, Nerman O, Kristiansson E. Modelling of zero-inflation improves inference of metagenomic gene count data. Statistical Methods in Medical Research. 2018; p. 096228021881135. https://doi.org/10.1177/0962280218811354 PMID: 30474490

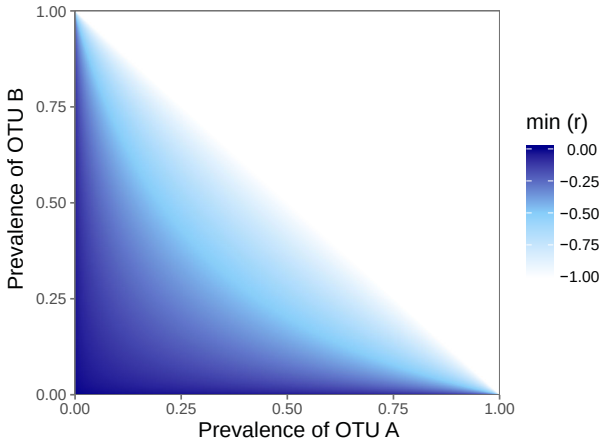
37. Lo C, Marculescu R. MPLasso: Inferring microbial association networks using prior microbial knowledge. PLOS Computational Biology. 2017; 13(12):e1005915. https://doi.org/10.1371/journal.pcbi.1005915 PMID: 29281638

Matériels supplémentaires



S1 Fig. Extrema of the Phi coefficient as a function of OTU prevalence. Minimum **(A)** and maximum **(B)** of the Phi coefficient as a function of prevalence. Computed from Eq (3).

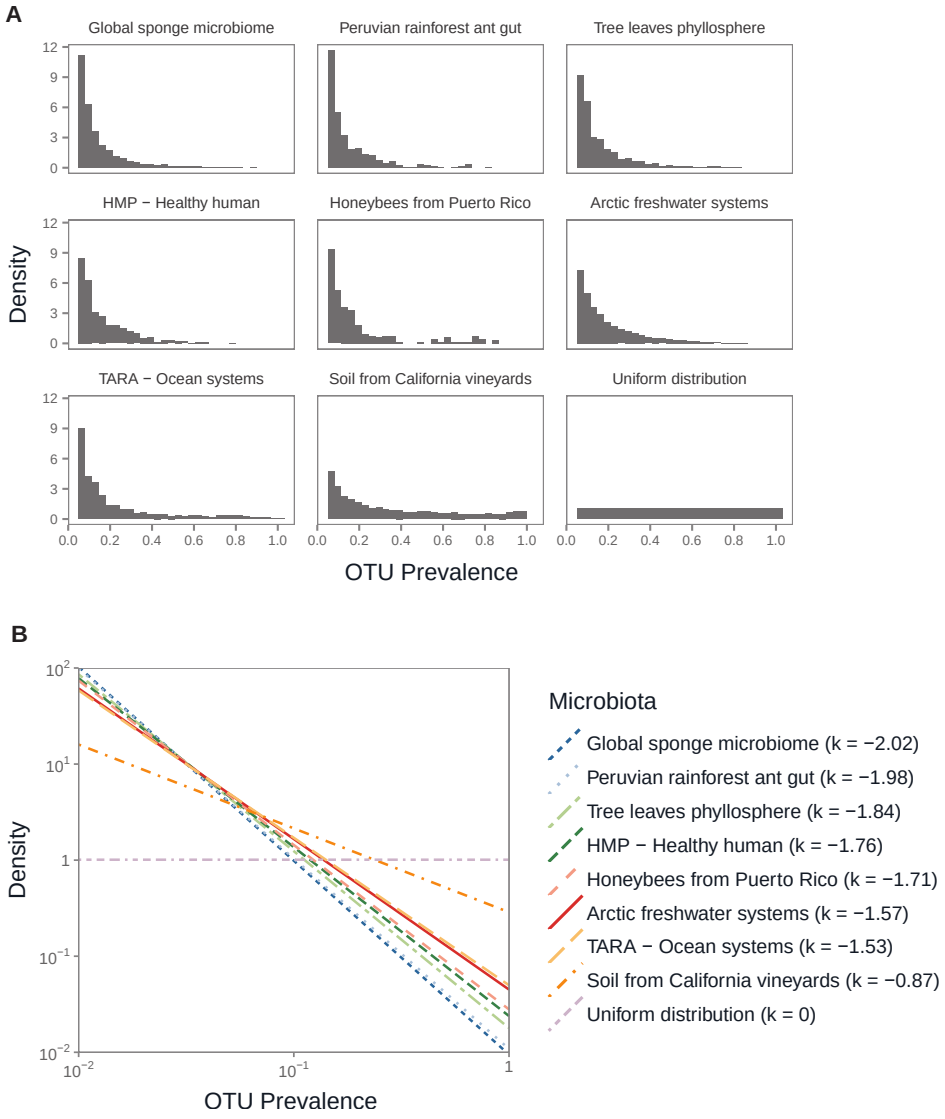
https://doi.org/10.1371/journal.pone.0200458.s001



S2 Fig. Minimum of the Pearson correlation coefficient as a function of OTU prevalence.

Minimum of the Pearson correlation coefficient r as a function of prevalence. Computed from Eq (6).

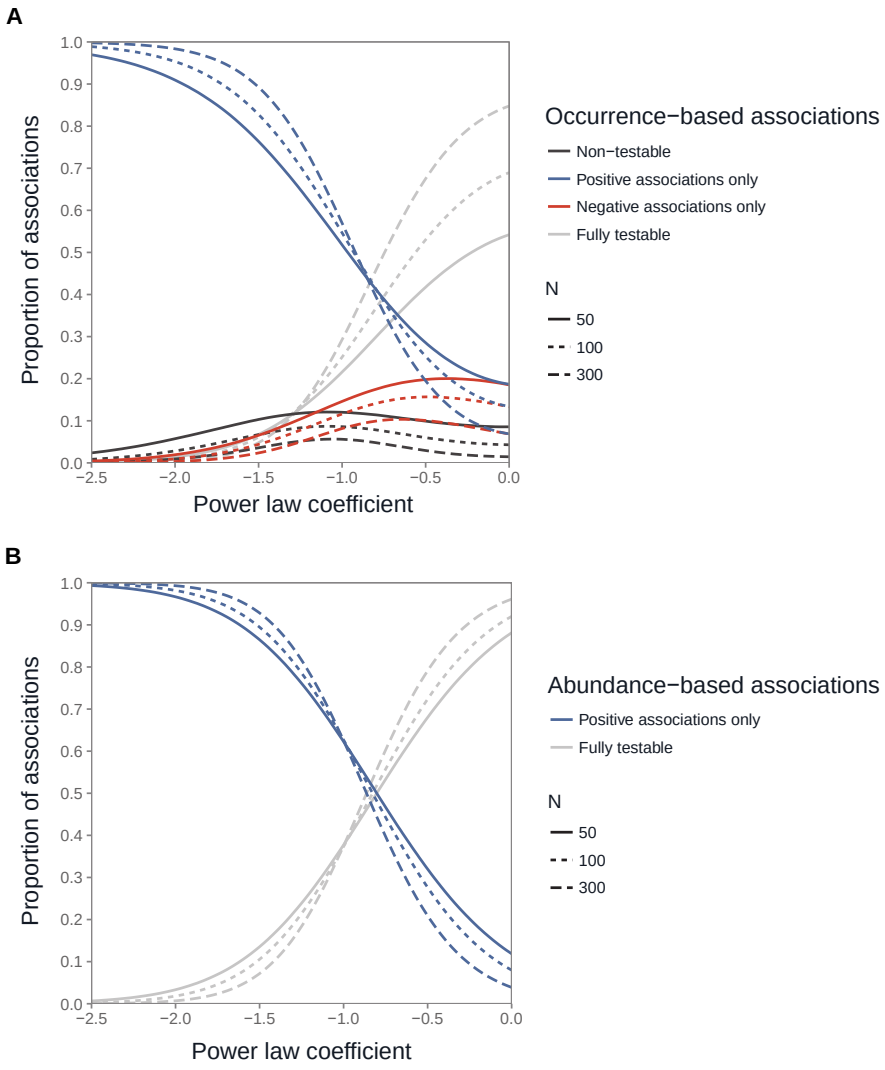
https://doi.org/10.1371/journal.pone.0200458.s002



S3 Fig. Prevalence structure of real microbial communities.

(A) Histograms of OTU prevalence in several microbiota characterized by 16S rRNA sequencing. The microbiota are described in Part E in S1 Appendix. **(B)** Probability density function of the same prevalence data (log-log scale), which were fitted to a truncated power law distribution ; the power law coefficient k was estimated by maximizing log-likelihood.

https://doi.org/10.1371/journal.pone.0200458.s003



S4 Fig. Proportion of testable associations as a function of the power law coefficient k.

Proportion of testable associations as a function of k when $N = 50$, 100, or 300 for the occurrence data **(A)** and for the read abundance data **(B)**.

https://doi.org/10.1371/journal.pone.0200458.s004

Rarity of microbial species: In search of reliable associations

Arnaud Cougoul, Xavier Bailly, Gwenaël Vourc'h, Patrick Gasqui

S1 Appendix. Supplementary Material

Contents

Supplementary Material

In the supplementary material below, we describe how we established our thresholds for occurrence data (i.e., represented by binary variables) and read abundance data (i.e., represented by positive continuous variables). We then discuss the link between the two threshold types. Finally, we describe the 16S data from several microbial communities that we used to characterise prevalence patterns.
First, we present the notation and decomposition of variance and covariance as a function of OTU co-occurrence.

A. Notation and decomposition of variance and covariance

We consider two OTUs whose abundances are modelled by two random variables, X_A and X_B (Tables 1 a and b). Our threshold is based on presence or absence of OTU, so we created a contingency table whose categories are defined by variable presence or absence.

	$X_B = 0$	$X_B \neq 0$	Total
$X_A = 0$	N_{00}	N_{01}	$\overline{N_A} = N - N_A$
$X_A \neq 0$	N_{10}	N_{11}	N_A
Total	$\overline{N_B} = N - N_B$	N_B	N

Table 1a. Contingency table of the presence/absence of two OTU read-abundance variables X_A and X_B where the entries are sample counts.

	$X_B = 0$	$X_B \neq 0$	Total
$X_A = 0$	P_{00}	P_{01}	$\overline{P_A} = 1 - P_A$
$X_A \neq 0$	P_{10}	P_{11}	P_A
Total	$\overline{P_B} = 1 - P_B$	P_B	1

Table 1b. Contingency table of the presence/absence of two OTU read-abundance variables X_A and X_B where the entries are proportions.

N is the number of microbiota samples; N_{00} is the number of co-absences of X_A and X_B; N_{11} is the number of co-occurrences of X_A and X_B; and $P_{11} = N_{11}/N$ is the proportion of co-occurrences of the two OTUs. $P_A = N_A/N$ and $P_B = N_B/N$ are the marginal probabilities of X_A and X_B, respectively (i.e., individual OTU prevalence). Since the OTUs are observed at least once, $P_A, P_B \in [1/N\,, 1]$.

We can calculate the mean and estimated variance of X_A and X_B using the non-zero values of X_A or X_B. Consequently, $\mu_{X_A} = P_A\ \mu_{X_A|\,X_A\neq 0}$, and $\mu_{X_B} = P_B\ \mu_{X_B|\,X_B\neq 0}$.

The estimated variances of X_A and X_B can be calculated as follows:

$$\begin{aligned}{\sigma_{X_A}}^2 &= \widehat{Var}(X_A) = \tfrac{1}{N}\sum_N\left(X_A - \mu_{X_A}\right)^2 = P_A\ \left(\sigma_{X_A|\,X_A\neq 0}\right)^2 + \ P_A\ \overline{P_A}\ \left(\mu_{X_A|\,X_A\neq 0}\right)^2 \\ {\sigma_{X_B}}^2 &= \widehat{Var}(X_B) = \tfrac{1}{N}\sum_N\left(X_B - \mu_{X_B}\right)^2 = P_B\ \left(\sigma_{X_B|\,X_B\neq 0}\right)^2 + P_B\ \overline{P_B}\ \left(\mu_{X_B|\,X_B\neq 0}\right)^2\end{aligned} \quad (1)$$

The estimated covariance of X_A and X_B can be decomposed based on whether or not X_A and X_B co-occur (i.e., X_A and X_B are non-null or not). If $\widehat{Cov}(X_A, X_B) = \mu_{X_A \times X_B} - \mu_{X_A} \times \mu_{X_B}$, then

$$\widehat{Cov}(X_A, X_B) = \overbrace{\left[P_{11}\, \widehat{Cov}\,(X_A, X_B)_{|\, X_A, X_B \neq 0}\right]}^{\text{"exclusively quantitative" covariance}} + \underbrace{\left[\, P_{11}\left(\mu_{X_A|\, X_A, X_B \neq 0} \times \mu_{X_B|\, X_A, X_B \neq 0}\right) - \left(\mu_{X_A} \times \mu_{X_B}\right)\right]}_{\text{"qualitative" covariance}} \quad (2)$$

"Exclusively quantitative" covariance

When the data are reduced into binary variables, $\widehat{Cov}\,(X_A, X_B)_{|\, X_A, X_B \neq 0} = 0$ because $\{X_A |\, X_A, X_B \neq 0\}$ and $\{X_B |\, X_A, X_B \neq 0\}$ are constants. Then $\left[P_{11}\, \widehat{Cov}\,(X_A, X_B)_{|\, X_A, X_B \neq 0}\right]$ is part of the covariance of X_A and X_B only because of the quantitative aspect of data.

"Qualitative" covariance

The second part of the covariance $\left[P_{11}\left(\mu_{X_A|\, X_A, X_B \neq 0} \times \mu_{X_B|\, X_A, X_B \neq 0}\right) - \left(\mu_{X_A} \times \mu_{X_B}\right)\right]$ is the difference between the mean product for the whole population and the mean product for the co-occurring elements only. Consequently, it can be explained by OTU co-occurrences (qualitative in nature).

When the data are reduced into binary variables (based on equations (1) and (2)):

$\widehat{Cov}(X_A, X_B) = P_{11}\left(\mu_{X_A|\, X_A, X_B \neq 0} \times \mu_{X_B|\, X_A, X_B \neq 0}\right) - \left(\mu_{X_A} \times \mu_{X_B}\right) = P_{11} - P_A\, P_B$

${\sigma_{X_A}}^2 = P_A\, {\sigma_{X_A \neq 0}}^2 + P_A\, (1 - P_A)\, {\mu_{X_A \neq 0}}^2 = P_A\, \overline{P_A}$ and ${\sigma_{X_B}}^2 = P_B\, \overline{P_B}$.

Therefore, the correlation of X_A and X_B, $cor(X_A, X_B) = \frac{\widehat{Cov}(X_A, X_B)}{\sigma_{X_A}\, \sigma_{X_B}}$, will depend only on P_{11}, P_A, and P_B.

B. Threshold method for binary data

Our method is based on the properties of discrete statistics. As binary data are discrete data, statistical tests have discrete distributions, as do *p*-values. Moreover, the minimum observable *p*-value for fixed marginal values can be higher than the alpha level (usually set to 5%), which means the test yields useless results [1,2]. In other words, for two OTUs with fixed prevalence, if all the possible values of an association index fall within the expected confidence interval, the association is simply not testable. Below, we will illustrate how OTU prevalence can thus shape potential correlations.

In this section, we detail how we developed our threshold method for binary data (i.e., OTU occurrence). First, we describe the association index used and show that it is bounded. Second, we present how we defined its testability. Third, we examine the consequences of our threshold method for network inference. Fourth, we present the testability limits on Fisher's exact test as a function of prevalence.

1. Measure of associations for binary data

The combinatorics that ensue from the hypergeometric law provide only simulated solutions for determining the testability of associations. In contrast, the Phi coefficient [3] can be used to establish equations for exploring association testability and give an analytical solution. The Phi coefficient is mathematically related to the common chi-square test. Since Fisher's exact test and Pearson's chi-square

test are asymptotically equivalent, we used the Phi coefficient as the basis for our threshold method. Moreover, we showed that the testability results were equivalent for both tests (see section A.7 and S1 Fig 3). Phi is also equivalent to the Pearson correlation coefficient in situations with binary data (coded by 0 and 1), a property that was helpful when extending our threshold method to quantitative situations (see sections 3 and 4).

Consider two random binary variables, $\widetilde{X_A}$ and $\widetilde{X_B}$, which represent the presence or absence of two OTUs. Working from Table 1, the Phi coefficient for the association between $\widetilde{X_A}$ and $\widetilde{X_B}$ is calculated as follows:

$$\phi = \frac{P_{11} - P_A P_B}{\sqrt{P_A \overline{P_A} P_B \overline{P_B}}} \text{ if } P_A,\ P_B \in]0,1[\ , \text{ and } \phi = 0 \text{ if not} \tag{3}$$

2. Bounds of the Phi coefficient as a function of prevalence

Based on the Boole–Fréchet inequality for logical conjunction, for the marginal probabilities $P_A,\ P_B \in]0,1[$, it follows that

$$max(0,\ P_A + P_B - 1) \ \leq\ P_{11} \ \leq\ min(P_A,\ P_B) \tag{4}$$

Given equations (3) and (4) and because ϕ is a continuous and monotonic function of P_{11}:

$$-1 \ \leq\ \phi_{min} \ \leq\ \phi \ \leq\ \phi_{max} \ \leq\ +1 \tag{5}$$

where

$$\phi_{min} = max\left(-\left(\frac{P_A P_B}{\overline{P_A}\,\overline{P_B}}\right)^{1/2},\ -\left(\frac{\overline{P_A}\,\overline{P_B}}{P_A P_B}\right)^{1/2}\right) \tag{5a}$$

$$\phi_{max} = min\left(\left(\frac{P_A \overline{P_B}}{P_B \overline{P_A}}\right)^{1/2},\ \left(\frac{P_B \overline{P_A}}{P_A \overline{P_B}}\right)^{1/2}\right) \tag{5b}$$

[4]

Therefore, ϕ is bounded and ϕ_{min} and ϕ_{max} depend exclusively on P_A and P_B.

3. Distribution of the Phi coefficient under the null hypothesis of independence

Under the null hypothesis (H_0) that the occurrences of two OTUs, $\widetilde{X_A}$ and $\widetilde{X_B}$, are independent, ϕ can be determined thanks to the Pearson's chi-squared test: $\phi^2 = \chi^2/N$, where N is the total number of observations and χ^2 is the chi-squared statistic for a 2x2 contingency table whose data follow a chi-squared distribution and for which there is 1 degree of freedom [5].

Since we know the distribution of ϕ, we can obtain the confidence interval at an alpha level of α. The confidence interval of a χ_1^2 distribution is $CI_{1-\alpha}(\chi_1^2) = [0, b]$, where b is defined by $P(\chi_1^2 > b) = \alpha$ (e.g., for $\alpha = 5\%$, $b \approx 1.96^2 \approx 3.84$).

The confidence interval of ϕ at an alpha level of α can be calculated as follows:

$$CI_{1-\alpha}(\phi) = \left[-\sqrt{K}, \sqrt{K}\ \right], \text{ where } K = {}^{b}/_{N} \tag{6}$$

4. Determining the testability of occurrence-based associations

We now examine the testability of the Phi coefficients calculated from pairs of OTU prevalence values. We do so by determining if the extrema of Phi occur within the confidence interval. There are two ways in which we may have trouble detecting significant associations:

A) If $\phi_{min} > -\sqrt{K}$, then we will not be able to detect a significant negative association.
B) If $\phi_{max} < \sqrt{K}$, then we will not be able to detect a significant positive association.

As ϕ_{min} and ϕ_{max} depend exclusively on P_A and P_B, we now consider the conditions under which P_A and P_B adopt problematic values.

We can split the first case (**A**) in two subcases because ϕ_{min} can have two different values depending on the specific values of P_A and P_B:

A1) If $P_A + P_B < 1$, then $max(0,\ P_A + P_B - 1) = 0$. Based on equations (3), (4), and (5a),

$$\phi_{min} = -\left(\frac{P_A\, P_B}{\overline{P_A}\ \overline{P_B}}\right)^{1/2}$$

A2) If $P_A + P_B \geq 1$, $max(0,\ P_A + P_B - 1) = P_A + P_B - 1$. Based on equations (3), (4), (5a),

$$\phi_{min} = -\left(\frac{\overline{P_A}\ \overline{P_B}}{P_A\, P_B}\right)^{1/2}$$

We can then resolve the inequation $\phi_{min} > -\sqrt{K}$.

A1) For $P_A + P_B < 1$, $\phi_{min} > -\sqrt{K} \quad \Leftrightarrow \quad \left(\frac{P_A\, P_B}{\overline{P_A}\ \overline{P_B}}\right)^{\frac{1}{2}} < \sqrt{K}$

$\Leftrightarrow \quad \frac{P_A\, P_B}{(1-P_A)(1-P_B)} < K$ (all variables are positive)

$$\Leftrightarrow \quad P_B < \frac{1-P_A}{1+\frac{1-K}{K}P_A} \tag{7}$$

A2) For $P_A + P_B \geq 1$, $\phi_{min} > -\sqrt{K} \quad \Leftrightarrow \quad \frac{(1-P_A)(1-P_B)}{P_A\, P_B} < K$

$$\Leftrightarrow \quad P_B > \frac{-1+P_A}{-1+(1-K)P_A} \tag{8}$$

If inequations (7) or (8) are true, a negative association cannot be detected.

The second case (**B**) can be similarly split up because ϕ_{max} can also have two values:

B1) If $P_A \leq P_B$, then $min(P_A,\ P_B) = P_A$. Based on equations (3), (4), and (5b),

$$\phi_{max} = \left(\frac{P_A\ \overline{P_B}}{P_B\ \overline{P_A}}\right)^{1/2}$$

B2) If $P_A \geq P_B$, then $min(P_A,\ P_B) = P_B$. Based on equations (3), (4), and (5b),

$$\phi_{max} = \left(\frac{P_B\ \overline{P_A}}{P_A\ \overline{P_B}}\right)^{1/2}$$

We can now solve the inequation $\phi_{max} < \sqrt{K}$.

B1) If $P_A \leq P_B$, $\phi_{max} < \sqrt{K} \Leftrightarrow \frac{P_A(1-P_B)}{P_B(1-P_A)} < K$

$$\Leftrightarrow P_B > \frac{P_A}{K+(1-K)P_A} \quad (9)$$

B2) If $P_A \geq P_B$, $\phi_{max} < \sqrt{K} \Leftrightarrow \frac{P_B(1-P_A)}{P_A(1-P_B)} < K$

$$\Leftrightarrow P_B < \frac{P_A}{\frac{1}{K}+\frac{K-1}{K}P_A} \quad (10)$$

If inequations (9) or (10) are true, a positive association cannot be detected.

Using the four inequations (7), (8), (9), and (10), we can delimit zones within which there is full, partial, or no testability. The characteristics of the tests in these zones will be detailed in the introduction to the next section.

For the two OTUs, P_A and P_B form a $[1/N, 1]^2$ square (Figure 1 below); $1/N$ is the smallest observable value. The testability zones in this square can be defined using four border functions that result from the inequations:

$$F_1(x) = \frac{1-x}{1+\frac{1-K}{K}x};\ F_2(x) = \frac{-1+x}{-1+(1-K)x};\ F_3(x) = \frac{x}{K+(1-K)x};\ F_4(x) = \frac{x}{\frac{1}{K}+\frac{K-1}{K}x} \quad (11)$$

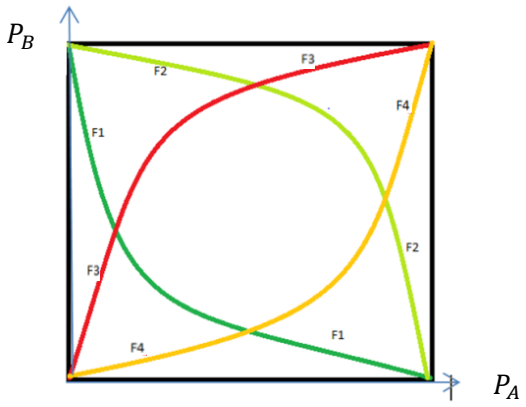


Figure 1. The four border functions delimiting testability

Emerging from these border functions are four graph intersections that are defined by:

$$\begin{aligned} F_1(x) = F_3(x) &= \frac{1}{2} \text{ at } x = \frac{K}{K+1} \\ F_2(x) = F_4(x) &= \frac{1}{2} \text{ at } x = \frac{1}{K+1} \\ F_2(x) = F_3(x) &= \frac{K}{K+1} \text{ at } x = \frac{1}{2} \\ F_1(x) = F_4(x) &= \frac{1}{K+1} \text{ at } x = \frac{1}{2} \end{aligned} \quad (12)$$

5. Proportion of associations in each testability zone

The zones defined by the border functions (11) contain different proportions of associations that can be categorised as fully testable, partially testable, or non-testable using our threshold method. The first zone, $A_{bilateral}$, contains associations for which both positive and negative correlations can be reliably tested. The second zone, $A_{unilateral}$, contains associations for which only positive correlations can be reliably tested (subzone $A_{positive}$) and for which only negative correlations can be reliably tested (subzone $A_{negative}$). Finally, the third zone, $A_{irrelevant}$, contains associations that cannot be reliably tested at all.

The distribution of prevalence values is treated as identical for all OTUs. Therefore, P_A and P_B have the same distribution and play symmetrical roles. However, these distribution patterns are not necessarily uniform. We examined two types of distributions—the uniform distribution and the truncated power law distribution; the latter fit the prevalence patterns of OTUs in real microbiota (see section 5).

For the uniform distribution of prevalence, the probability density function is

$$f(x) = \begin{cases} \dfrac{1}{1-\dfrac{1}{N}} = \dfrac{N}{N-1} & if\ \dfrac{1}{N} \le x \le 1 \\ 0 & if\ not \end{cases} \tag{13}$$

For the truncated power law distribution of prevalence, the probability density function is

$$f(x) = \begin{cases} Cx^k & if\ \dfrac{1}{N} \le x \le 1 \\ 0 & if\ not \end{cases} \tag{14}$$

and, following normalization, we arrive at $\int_{\frac{1}{N}}^{1} f(x)\,dx = C\,\frac{\left[x^{k+1}\right]_{1/N}^{1}}{k+1} = 1$, so $C = \frac{k+1}{1-\left(\frac{1}{N}\right)^{k+1}}$.

When $k = 0$, we have a uniform distribution with the interval $\left[\frac{1}{N}, 1\right]$.

To computationally define the different zones, analytical formulas can be used in the case of the uniform distribution but not in the case of the power law distribution. Consequently, in the latter situation, we chose to proceed by numerical integration. Since the current form of the R function *integrate* (in the stats package) does not deal well with the power law, we used a Monte Carlo approach. This consisted of generating random prevalence values in accordance with the observed prevalence distribution (see section 2.6) and counting how many fell within each of the zones.

To simplify the zone-defining equations below, we have used the following notation:

(F_1+): "$y > F_1(x)$" and (F_1-): "$y < F_1(x)$"

and the same notation applies in the cases of F_2, F_3 and F_4.

$\wedge$ denotes the logical conjunctions.

From the four inequations (7, 8, 9, 10) and the border function (11), the proportions of associations that fall within each zone are determined as follows:

$$A_{bilateral} = \iint_{\{(F_1+)\,\wedge\,(F_4+)\,\wedge\,(F_2-)\,\wedge\,(F_3-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y$$

$$A_{positive} = \iint_{\{(F_3-) \wedge (F_1-) \wedge (F_4+)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y + \iint_{\{(F_3-) \wedge (F_2+) \wedge (F_4+)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y$$

$$A_{negative} = \iint_{\{(F_1+) \wedge (F_3+) \wedge (F_2-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y + \iint_{\{(F_1+) \wedge (F_4-) \wedge (F_2-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y$$

$$A_{irrelevant} = \iint_{\{(F_3+) \wedge (F_1-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y + \iint_{\{(F_2+) \wedge (F_4-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y + \iint_{\{(F_2+) \wedge (F_3+)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y + \iint_{\{(F_1-) \wedge (F_4-)\}} f(x)f(y)\,\mathrm{d}x\mathrm{d}y$$

$$A_{unilateral} = A_{positive} + A_{negative}$$

$$A_{bilateral} + A_{unilateral} + A_{irrelevant} = \iint f(x)f(y)\,\mathrm{d}x\mathrm{d}y = 1$$

6. Defining testability zones using a Monte Carlo method

To compute Monte Carlo integrations, it is necessary to generate random prevalence values using the observed distribution of prevalence. For the uniform distribution, many pseudorandom number generators exist. However, for the truncated power law distribution, we had to employ an inverse transformation method that is rooted in the following property:

If V follows a power law, then $F(V) = U$ is uniformly distributed (interval of $[0,1]$) and $F^{-1}(U) = V$.

We therefore needed to define the inverse cumulative distribution function. Let F be the cumulative distribution function of the truncated power law distribution as defined in (14).

If $F(x) = \int_{\frac{1}{N}}^{x} f(t)\,dt = C\int_{\frac{1}{N}}^{x} t^k\,dt = \frac{C}{k+1}\left(x^{k+1} - \left(\frac{1}{N}\right)^{k+1}\right) = \frac{x^{k+1} - \left(\frac{1}{N}\right)^{k+1}}{1 - \left(\frac{1}{N}\right)^{k+1}}$,

then $F^{-1}(x) = \left(\left(1 - \left(\frac{1}{N}\right)^{k+1}\right)x + \left(\frac{1}{N}\right)^{k+1}\right)^{\frac{1}{k+1}}$.

We can then generate a power law distribution from a uniform distribution using the following equation:

$$V = \left(\left(1 - \left(\frac{1}{N}\right)^{k+1}\right)U + \left(\frac{1}{N}\right)^{k+1}\right)^{\frac{1}{k+1}} \quad (15)$$

7. Testability limits on Fisher's exact test

Co-occurrence networks are commonly reconstructed using the hypergeometric law that underlies Fisher's exact test [6–8].

From an observed 2x2 contingency table (Table 1), Fisher showed that the probability P of obtaining such a set was given by the hypergeometric distribution:

$$P = \frac{\binom{N_{00}+N_{01}}{N_{00}} + \binom{N_{10}+N_{11}}{N_{10}}}{\binom{N}{N_{00}+N_{10}}} = \frac{(N_{00}+N_{01})!\ (N_{10}+N_{11})!\ (N_{00}+N_{10})!\ (N_{01}+N_{11})!}{N_{00}!\ N_{01}!\ N_{10}!\ N_{11}!\ N!} \quad (16)$$

where $\binom{n}{k}$ is the binomial coefficient and ! indicates the factorial.

This equation can be written according to N_A, N_B, N and N_{11}:

$$P = \frac{\binom{N-N_A}{N_{11}+(N-N_A)+(N-N_B)-N} + \binom{N_A}{N-N_{11}-(N-N_A)}}{\binom{N}{N-N_B}} = \frac{\binom{N-N_A}{N_B-N_{11}} + \binom{N_A}{N_{11}}}{\binom{N}{N_B}} \quad (17)$$

Based on the Boole–Fréchet inequality for logical conjunction, for the marginal counts $N_A,\ N_B \in]0, N[$, it follows that

$$max(0,\ N_A + N_B - \mathrm{N}) \leq N_{11} \leq min(N_A,\ N_B) \quad (18)$$

We have two extreme situations:

a) Observe the minimum number of co-occurrences, $N_{11} = min(N_{11}) = max(0,\ N_A + N_B - \mathrm{N})$

b) Observe the maximum number of co-occurrences, $N_{11} = max(N_{11}) = min(N_A,\ N_B)$

We can calculate the probability P associated with these two situations **a)** and **b)**. A bilateral test can also be performed. As in the fisher.test function of R, the p-value is computed by summing the probability for all table with probabilities less than or equal to that of the observed table.

For two given OTUs with prevalence $P_A = \frac{N_A}{N}$ and $P_B = \frac{N_B}{N}$, we have 4 possibilities in the testability limits on Fisher's exact test:

- If the p-values associated with the two configurations **a)** and **b)** are lower than the alpha level (5%), the two extremes situations **a)** and **b)** correspond to significant associations. We have no limit on the test.
- If the p-value associated with the configuration **a)** is greater than the alpha level, then we will not be able to detect a significant negative association.
- If the p-value associated with the configuration **b)** is greater than the alpha level, then we will not be able to detect a significant positive association.
- If the p-values associated with the configurations **a)** and **b)** are greater than the alpha level, then we will not be able to detect a significant positive or negative association.

C. Threshold method for quantitative data

In this section, we detail how we developed our threshold method for quantitative data (i.e., OTU read abundance). First, we introduce our system of notation and the primary elements of our proof. Second, we present the situation, in which correlations are bounded by an excess of zeroes, and describe the minimum correlation value. Third, we show how we defined association testability. Finally, we examine the consequences of our threshold method for network inference.

1. Introduction

In this section, X_A and X_B are two random variables that represent quantitative data. The Pearson correlation coefficient [9] is used to characterise the pairwise associations in OTU read abundance. We were specifically interested in understanding how the number of zeroes in the data could influence the correlation coefficient.

We use same notations as in section 1. $\overline{N_A}$ and $\overline{N_B}$ represent the number of zeros associated with X_A and X_B, respectively. N_{00} is the number of co-absences of X_A and X_B, and N_{11} is the number of co-occurrences.

Based on Table 1 and the Boole–Fréchet inequalities, we can deduce the following:

$$max(0;\ \overline{N_A} + \overline{N_B} - N) \leq N_{00} \leq min(\overline{N_A}; \overline{N_B}) \tag{19}$$

$$N_{11} = N - \overline{N_A} - \overline{N_B} + N_{00} \tag{20}$$

$$N_{11} \geq max(N - \overline{N_A} - \overline{N_B};\ 0) \tag{21}$$

For pairs of $\overline{N_A}$ and $\overline{N_B}$, we distinguish two cases:

i. $\boldsymbol{\overline{N_A} + \overline{N_B} \leq N}$

The number of zeros is sufficiently low such that there are no raw restrictions on possible correlations. Indeed, it is simple to build two non-restricted correlations that approach infimum -1 and supremum $+1$:

$$\begin{pmatrix} X_A \\ X_B \end{pmatrix} = \begin{pmatrix} \overbrace{0,\ 0,\ \dots, 0,}^{\overline{N_A}} & \overbrace{a, a, \dots, a,}^{N_{11}} & 2a, 2a, \dots, 2a \\ 2a, 2a, \dots, 2a, & a, a, \dots, a, & \underbrace{0,\ 0,\ \dots, 0}_{\overline{N_B}} \end{pmatrix} \text{ where } a > 0$$

In this case, the correlation coefficient is $r = -1$.

$$\begin{pmatrix} X_A \\ X_B \end{pmatrix} = \begin{pmatrix} \overbrace{0,0, \dots, 0,}^{\overline{N_A}} & 0, 0, \dots, 0, & \overbrace{a, a, \dots, a}^{N_{11}} \\ \underbrace{0,0, \dots, 0,}_{\overline{N_B}} & h, h, \dots, h, & a, a, \dots, a \end{pmatrix} \text{where } a, h > 0$$

The correlation tends toward the supremum, $r \xrightarrow[h \to 0]{} +1$ or $r \xrightarrow[a \to +\infty]{} +1$.

ii. $\boldsymbol{\overline{N_A} + \overline{N_B} > N}$

Based on equations (19) and (21), $N_{00} \geq \overline{N_A} + \overline{N_B} - N > 0$ and $N_{11} \geq 0$. Consequently, N_{11} can equal zero, meaning that there are enough zeros associated with X_A and X_B that X_A and X_B may not co-occur. In this situation, information on quantitative correlations is degraded. We can prove that r, the Pearson correlation coefficient, has a minimum, r_{min}, that is different from -1:

$$r_{min} \leq r \leq 1,$$

where $r_{min} = -\left(\frac{N_A\, N_B}{\overline{N_A}\ \overline{N_B}}\right)^{1/2} > -1$

2. Determining the lower bound of the Pearson correlation coefficient

Given $\overline{N_A}$ and $\overline{N_B}$, we wished to determine the minimum possible correlation between X_A and X_B. We highlight that a lower bound of the Pearson correlation exists between two positive variables and prove that it can be reached under certain conditions.

For the association between X_A and X_B, the Pearson correlation coefficient is calculated as follows:

$$r = \frac{\widehat{Cov}(X_A, X_B)}{\sigma(X_A)\,\sigma(X_B)} = \frac{\mu(X_A X_B) - \mu(X_A)\,\mu(X_B)}{\sigma(X_A)\,\sigma(X_B)}$$

If $X_A, X_B \geq 0$, then $\mu(X_A X_B) \geq 0$ and

$$r \geq \frac{-\mu(X_A)\,\mu(X_B)}{\sigma(X_A)\,\sigma(X_B)} \quad (22)$$

where equality holds if and only if $\mu(X_1 X_2) = 0$

Consequently, the mean of $X_A X_B$ is null if and only if there are no co-occurrences. In other words,

$$\mu(X_A X_B) = 0 \text{ if and only if } N_{11} = 0 \quad (23)$$

- If $\mu(X_1 X_2) = 0$, then $\sum X_1 X_2 = 0$. Each element of the sum are positive then $\sum X_1 X_2 = 0$ imply that all elements are null and there are no co-occurrences (i.e., $N_{11} = 0$).
- If there are no co-occurrences, then $X_1 X_2 = 0$ and $\mu(X_1 X_2) = 0$.

From equations (22) and (23), we can conclude that

$$-\frac{\mu(X_A)\,\mu(X_B)}{\sigma(X_A)\,\sigma(X_B)} \leq r \quad (24)$$

where equality holds if and only if $N_{11} = 0$

Moreover, if $\overline{N_A} + \overline{N_B} \leq N$, then, from equation (21), we know that $N_{11} \neq 0$. Therefore,

$$N_{11} = 0 \Longrightarrow \overline{N_A} + \overline{N_B} > N \quad (25)$$

We now want to control $-\frac{\mu(X_A)\,\mu(X_B)}{\sigma(X_A)\,\sigma(X_B)}$ and find its minimum. We therefore maximise $\frac{\mu(X_A)}{\sigma(X_A)}$ and $\frac{\mu(X_B)}{\sigma(X_B)}$ separately. μ/σ corresponds to the inverse coefficient of variation.

3. Maximising the inverse coefficient of variation

Below, we illustrate how to maximise the inverse coefficient of variation for X_A. We will show that $\frac{\mu(X_A)}{\sigma(X_A)} \leq \sqrt{\frac{N_A}{\overline{N_A}}}$.

We can express variance using the König–Huygens formula:

$$\widehat{Var}(X_A) = \mu\left({X_A}^2\right) - \mu(X_A)^2$$

If $\mu(X_A) \neq 0$, then

$\frac{\widehat{Var}(X_A)}{\mu(X_A)^2} = \frac{\mu({X_A}^2)}{\mu(X_A)^2} - 1$ and $\frac{\widehat{Var}(X_A)}{\mu(X_A)^2} = \frac{\frac{1}{N}\sum_{i=1}^{N} {X_A}_i^2}{\left(\frac{1}{N}\sum_{i=1}^{N} X_{A_i}\right)^2} - 1$, which means

$$\frac{\widehat{Var}(X_A)}{\mu(X_A)^2} = N\frac{\sum_{i=1}^{N} {X_A}_i^2}{\left(\sum_{i=1}^{N} X_{A_i}\right)^2} - 1 \quad (26)$$

We are now interested in $\frac{\sum_{i=1}^{N} {X_A}_i^2}{\left(\sum_{i=1}^{N} X_{A_i}\right)^2}$, and we will show that $\frac{\sum_{i=1}^{N} {X_A}_i^2}{\left(\sum_{i=1}^{N} X_{A_i}\right)^2} \geq \frac{1}{N_A}$.

Let V, W be two vectors of $\mathbb{R}^N$. As per the Cauchy–Schwarz inequality,

$$\left(\sum_{i=1}^{N} V_i \times W_i\right)^2 \leq \sum_{i=1}^{N} V_i^2 \sum_{i=1}^{N} W_i^2$$

where equality holds if and only if V and W are collinear.

Let $V = Y$ be the vector of non-null elements of X_A (for Y, vector size is equal to N_A); $W = 1_{N_A}$, a constant vector of size N_A. In this case, the Cauchy–Schwarz inequality becomes the following:

$$\left(\sum_{i=1}^{N_A} Y_i \times 1\right)^2 \leq \sum_{i=1}^{N_A} Y_i^2 \sum_{i=1}^{N_A} 1^2$$

where equality holds if and only if $Y = \lambda \times 1_{N_A}$, where $\lambda > 0$ (i.e., Y is a constant vector).

As $\sum_{i=1}^{N_A} Y_i = \sum_{i=1}^{N} X_{A_i}$ and $\sum_{i=1}^{N_A} Y_i^2 = \sum_{i=1}^{N} {X_A}_i^2$,

$\left(\sum_{i=1}^{N} X_{A_i}\right)^2 \leq N_A \sum_{i=1}^{N} {X_A}_i^2$, then

$$\frac{\sum_{i=1}^{N} {X_A}_i^2}{\left(\sum_{i=1}^{N} X_{A_i}\right)^2} \geq \frac{1}{N_A} \quad (27)$$

where equality holds if and only if the non-null elements of X_A are constant.

Based on equations (26) and (27), we now observe that

$$\frac{\widehat{Var}(X_A)}{\mu(X_A)^2} \geq \frac{N}{N_A} - 1 \Leftrightarrow \frac{\widehat{Var}(X_A)}{\mu(X_A)^2} \geq \frac{N - N_A}{N_A} \Leftrightarrow \frac{\sqrt{\mu(X_A)^2}}{\sqrt{\widehat{Var}(X_A)}} \leq \frac{\sqrt{N_A}}{\sqrt{N - N_A}}$$

Finally,

$$\frac{\mu(X_A)}{\sigma(X_A)} \leq \frac{\sqrt{N_A}}{\sqrt{\overline{N_A}}} \quad (28)$$

where equality holds if and only if the non-null elements of X_A are constant.

The maximum occurs where $\frac{\mu(X_A)}{\sigma(X_A)}$ is $\sqrt{\frac{N_A}{\overline{N_A}}}$.

The approach is equivalent for X_B, so we can conclude that

$$\frac{\mu(X_B)}{\sigma(X_B)} \leq \frac{\sqrt{N_B}}{\sqrt{\overline{N_B}}} \tag{29}$$

where equality holds if and only if the non-null elements of X_B are constant.

4. Determining the minimum Pearson correlation coefficient when there are many zeros

Based on equations (24), (28), and (29),

$$-\sqrt{\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}}} \leq r$$

where equality holds if and only if $N_{11} = 0$ and the non-null elements of X_A and X_B are constant. (30)

It therefore stands to reason that

$$\text{if } \overline{N_A} + \overline{N_B} > N, \text{ then } -\sqrt{\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}}} > -1 \tag{31}$$

$\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}} = \frac{(N-\overline{N_A})(N-\overline{N_B})}{\overline{N_A}\,\overline{N_B}} = \frac{N\left(N-(\overline{N_A}+\overline{N_B})\right)+\overline{N_A}\,\overline{N_B}}{\overline{N_A}\,\overline{N_B}} = \frac{N\left(N-(\overline{N_A}+\overline{N_B})\right)}{\overline{N_A}\,\overline{N_B}} + 1$

If $\overline{N_A} + \overline{N_B} > N,\ N - (\overline{N_A} + \overline{N_B}) < 0$ and $\frac{N\left(N-(\overline{N_A}+\overline{N_B})\right)}{\overline{N_A}\,\overline{N_B}} + 1 < 1$

Therefore, $\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}} < 1$ and $-\sqrt{\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}}} > -1$.

Finally, based on equations (30) and (31), when $\overline{N_A} + \overline{N_B} > N$,

$$-1 < r_{min} \leq r \leq 1$$

where r can attain $r_{min} = -\sqrt{\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}}}$ if and only if $N_{11} = 0$ and the non-null elements of X_A and X_B are constant. (32)

5. Constraints on the testability of the Pearson correlation coefficient

When X_A and X_B follow two uncorrelated normal distributions, $r \sim \frac{t}{\sqrt{N-2+t^2}}$, where t is a Student's t statistic with degrees of freedom $N-2$. We can then determine a confidence interval: $CI_{1-\alpha}(r) = \left[-\sqrt{K};\ \sqrt{K}\right]$, where K depends on α and N.

Returning to our measures of OTU prevalence, if $P_A = \frac{N_A}{N}$ and $P_B = \frac{N_B}{N}$, then $r_{min} = -\sqrt{\frac{N_A N_B}{\overline{N_A}\,\overline{N_B}}} = -\sqrt{\frac{P_A P_B}{\overline{P_A}\,\overline{P_B}}}$. The constraint is the same as in the case of binary data.

If r_{min} falls within the confidence interval, we can conclude that negative associations cannot be detected.

$$-\sqrt{\frac{P_A P_B}{\overline{P_A}\,\overline{P_B}}} > -\sqrt{K}$$
$$\Leftrightarrow \quad P_B < \frac{1 - P_A}{1 + \frac{1-K}{K} P_A} \tag{33}$$

Accordingly, if inequation (33) is true, then negative associations are not testable.

The border function that defines the testability zones in the square formed by $P_A \times P_B$ is as follows:

$$F_1(x) = \frac{1 - x}{1 + \frac{1-K}{K} x} \tag{34}$$

6. Proportion of associations in each testability zone

Using the border function (34), we observed that two zones existed. The first zone, $A_{bilateral}$, contains associations for which both positive and negative correlations can be reliably tested. The second zone, $A_{positive}$, contains associations for which only positive correlations can be reliably tested. As for the binary data (sections 2.5 and 2.6), we explored the testability of abundance-based associations using the uniform distribution and the truncated power law distribution. In the latter case, we again employed a Monte Carlo approach.

Based on the border function (34), the proportions of associations that fall within each zone can be determined as follows:

$$A_{bilateral} = \iint_{\{(F_1+)\}} f(x) f(y)\, \mathrm{d}x\mathrm{d}y$$

$$A_{positive} = \iint_{\{(F_1-)\}} f(x) f(y)\, \mathrm{d}x\mathrm{d}y$$

$$A_{bilateral} + A_{positive} = \iint f(x) f(y)\, \mathrm{d}x\mathrm{d}y = 1$$

(Same notation as in section 2.5)

7. Spearman correlation invariance

The Spearman correlation between two continuous variables X_A and X_B is calculated as follows:

$$\rho_{Spearman}(X_A, X_B) = r_{Pearson}(rg(X_A), rg(X_B))$$

where $rg(X)$ is the function that associates the ranks of X.

The identical values will be assigned to the average of their positions in the ascending order of the values, which is equivalent to averaging over all possible permutations.

If we call $\overline{N_A}$ the number of zeros in X_A, the $\overline{N_A}$ zero values will be identical values and will be assigned to the rank $mean(\{1, \dots, \overline{N_A}\})$, $\{1, \dots, \overline{N_A}\}$ being all possible rank values for these $\overline{N_A}$ null values.

As $mean(\{1, \dots, \overline{N_A}\} = \frac{\overline{N_A}\,(\overline{N_A}-1)}{2}$,

We are now interested by $Y_A = rg(X_A) - \frac{\overline{N_A}\,(\overline{N_A}-1)}{2}$ and $Y_B = rg(X_B) - \frac{\overline{N_B}\,(\overline{N_B}-1)}{2}$.

If $X_A = 0$, $rg(X_A) = \frac{\overline{N_A}\,(\overline{N_A}-1)}{2}$ and $Y_A = rg(X_A) - \frac{\overline{N_A}\,(\overline{N_A}-1)}{2} = 0$

Zeros of X_A are zeros of Y_A, and the same for X_B and Y_B.

Moreover,

$$r_{Pearson}(Y_A, Y_B) = r_{Pearson}\left(rg(X_A) - \frac{\overline{N_A}\,(\overline{N_A}-1)}{2},\ rg(X_B) - \frac{\overline{N_B}\,(\overline{N_B}-1)}{2}\right)$$

As correlation is invariant by translation:

$$r_{Pearson}(Y_A, Y_B) = r_{Pearson}(rg(X_A),\ rg(X_B)) = \rho_{Spearman}(X_A, X_B)$$

We thus constructed two variables Y_A and Y_B which:

- have the same null values than X_A and X_B.
- have a Pearson correlation equal to the Spearman correlation of X_A and X_B
- are two positive continuous variables with the same limitations on their Pearson correlation depending on prevalence as described in the part 3.5.

Thus, when we study Spearman correlation, we implicitly make a Pearson correlation with the same number of zeros and then the same limitations as we have previously mentioned.

8. Data transformation

Since the correlation is invariant by translation (see the paragraph above), if a positive transformation $t()$ transforms all the null values in a single value z_0, it suffices to study the correlation $cor(t(X_A) - z_0, t(X_B) - z_0)$ to return to the general problem. The limit on the testability of the correlation will be the same for this type of transformation.

For microbial data, this works for Total Sum Scaling (TSS) and rarefying.

The centered log ratio (clr), the cumulative sum scaling (CSS) and DESeq transformation use a pseudo-count that did not produce the theoretical results obtained, although the simulations show that the problem is still present for the clr transformations and this is also probably the case for the others.

Use of a pseudo count to avoid log(0) is not ideal because clustering results have been shown to be very sensitive to the choice of pseudo-count, due to the nonlinear nature of the log transform[10,11].

D. Similarity of the Phi and Pearson correlation coefficients

In this section, we show that testability constraints tend to be similar with both occurrence and abundance data. We also examine the degree of correlation between the correlation coefficients calculated using the two data types.

1. Testability constraints on occurrence and abundance data

The distribution of the correlation coefficient for two normally distributed independent variables is $r \sim \frac{t_{N-2}}{\sqrt{N-2+t_{N-2}^2}}$.

As $t_{N-2} \xrightarrow[N\to+\infty]{} \mathcal{N}(0,1)$ (i.e., there is distribution convergence) and $\frac{\sqrt{N-2+t_{N-2}^2}}{\sqrt{N}} \xrightarrow[N\to+\infty]{} 1$, then $r \xrightarrow[N\to+\infty]{} \frac{\mathcal{N}(0,1)}{\sqrt{N}}$. Since the distribution of the square of the Phi coefficient is $\phi^2 \sim \frac{\chi_1^2}{N} \sim \frac{\mathcal{N}(0,1)^2}{N}$ under the null hypothesis of independence, the Pearson correlation coefficient will asymptotically attain the same confidence interval as the Phi coefficient: their lower bounds converge upon $\sqrt{b/N}$ (sections 2.3 and 3.5).

We now underscore that the Phi and Pearson correlation coefficients have the same lower bound when the two OTUs have low levels of prevalence: $r_{min} = \phi_{min} = -\sqrt{\frac{P_A P_B}{\overline{P_A}\,\overline{P_B}}}$.

When N is large enough, the testability of positive associations will be the same for binary data and quantitative data. This pattern will be all the more pronounced given that, in real microbiota, OTU prevalence is greatly skewed to the right: positive associations represent the majority of associations to be tested.

2. Correlation between Phi and Pearson coefficients

In section 1, we showed that variance can be decomposed in a quantitative part and a qualitative part (equation (2)). Here, we use the results of a simulation to explore how the strength of the correlation between the values of the Phi coefficient and the Pearson coefficient is related to OTU prevalence. We are most interested in what happens when prevalence is low.

OTU abundances X_A and X_B are modelled by a zero-inflated Poisson (ZIP) distribution using the following probability mass function:

$$f(x) = \begin{cases} p_0 + (1-p_0).e^{-\lambda} & if\ x = 0 \\ (1-p_0).\dfrac{\lambda^x e^{-\lambda}}{x!} & if\ x = 1, 2 \ldots \end{cases}$$

where the probability of structural zeros, p_0, is the result of a Bernoulli process and λ is the mean of the Poisson portion of the distribution (i.e., the Poisson parameter). In the simulation, X_A and X_B had the same values for p_0 and λ.

The probability of structural zeros p_0 represents the complementary probability of prevalence P, i.e. $p_0 = 1 - P$. As p_0 increases (i.e., prevalence decreases), the correlation between the Phi coefficient and the Pearson coefficient increases (Figure 3A in the article). The correlation also strengthens as λ increases. When prevalence is below 0.25, the correlation is greater than 0.75 for all values of λ.

If OTU prevalence follows a ZIP distribution, we can conclude that the values of the Phi coefficient and the Pearson coefficient will be correlated, especially when OTU prevalence is low.

E. Distribution of OTU prevalence in real microbiota

To characterise actual OTU distribution patterns, we employed data from the QIITA database (qiita.ucsd.edu) and the TARA Ocean Project (ocean-microbiome.embl.de) [12]. The biom files were processed using the R package *biomformat*. We deliberately chose different kinds of microbiota so as to represent as wide a diversity of microbial communities as possible (Table 2). We used OTU rather than species tables.

The prevalence values were fitted to a truncated power law distribution as described by equation (14), and the power law coefficient k was estimated by maximizing the log-likelihood [13].

Source	Samples	OTUs	Median of Prevalence	Mean sequencing depth	Estimated k
Arctic freshwater systems (ID Qiita 1883)	3153	32347	0.004440216	47903.11	-1.567
Gut bacteria of Peruvian rainforest ants (ID Qiita 10343)	471	9819	0.004246285	34773.16	-1.981
HMP healthy human [14] (ID Qiita 1928)	6000	10730	0.0006666667	4538.797	-1.758
Honeybees from Puerto Rico (ID Qiita 1064)	387	3789	0.002583979	14974.18	-1.711
Soil from California vineyards (ID Qiita 10082)	237	13149	0.05907173	23479.96	-0.873
Sponge (ID Qiita 1740)	1403	24447	0.00427655	42056.75	-2.018
Tree leaves [15] (ID Qiita 396)	107	4218	0.01869159	936.7477	-1.841
TARA Ocean Project [12]	139	24798	0.02158273	34168.53	-1.534

Table 2. Sources of the microbiota we analysed and the associated number of samples, number of OTUs, and estimates of the power law coefficient k.

References

1. Tarone RE. A Modified Bonferroni Method for Discrete Data. Biometrics. 1990;46: 515. doi:10.2307/2531456

2. Carlson J, Heckerman D, Shani G. Estimating false discovery rates for contingency tables. Microsoft, Redmond, WA. 2009.

3. Yule GU. On the Methods of Measuring Association Between Two Attributes. J R Stat Soc. 1912;75: 579. doi:10.2307/2340126

4. Chaganty NR, Joe H. Range of correlation matrices for dependent Bernoulli random variables. Biometrika. 2006;93: 197–206. doi:10.1093/biomet/93.1.197

5. Guilford JP. The phi coefficient and chi square as indices of item validity. Psychometrika. 1941;6: 11–19. doi:10.1007/BF02288569

6. Chaffron S, Rehrauer H, Pernthaler J, von Mering C. A global network of coexisting microbes from environmental and whole-genome sequence data. Genome Res. 2010;20: 947–959. doi:10.1101/gr.104521.109

7. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. Nature Publishing Group; 2012;10: 538–550. doi:10.1038/nrmicro2832

8. Li C, Lim KMK, Chng KR, Nagarajan N. Predicting microbial interactions through computational approaches. Methods. Elsevier Inc.; 2016;102: 12–19. doi:10.1016/j.ymeth.2016.02.019

9. Pearson K. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. Philos Trans R Soc A Math Phys Eng Sci. 1896;187: 253–318. doi:10.1098/rsta.1896.0007

10. Costea PI, Zeller G, Sunagawa S, Bork P. A fair comparison. Nat Methods. Nature Publishing Group; 2014;11: 359–359. doi:10.1038/nmeth.2897

11. Paulson JN, Bravo HC, Pop M. Reply to: A fair comparison. Nat Methods. 2014;11: 359–360. doi:10.1038/nmeth.2898

12. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science (80-). 2015;348: 1261359–1261359. doi:10.1126/science.1261359

13. Deluca A, Corral Á. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. Acta Geophys. 2013;61: 1351–1394. doi:10.2478/s11600-013-0154-9

14. Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, et al. Structure, function and diversity of the healthy human microbiome. Nature. Nature Publishing Group; 2012;486: 207–214. doi:10.1038/nature11234

15. Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N. The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. Environ Microbiol. 2010;12: 2885–2893. doi:10.1111/j.1462-2920.2010.02258.x

Chapitre 3

Inférence de réseaux d'associations microbiennes

Les microorganismes vivent souvent en relation symbiotique avec leur environnement et jouent un rôle central dans de nombreux processus biologiques. Ils forment des systèmes complexes d'espèces en interactions. Comprendre les mécanismes qui régissent ces écosystèmes est donc un défi scientifique majeur. L'acquisition de données sur le microbiote par le biais de la métagénomique ciblée devient de plus en plus facile, avec des échantillons de plus grande taille. Les réseaux basés sur les corrélations par paires et les modèles graphiques sont couramment utilisés pour identifier les réseaux d'interaction putatifs formés par les espèces de micro-organismes, mais ces méthodes ne prennent pas toujours en compte tous les aspects des données de composition microbienne. En effet, les réseaux basés sur les corrélations ne permettent pas de distinguer les corrélations directes des corrélations indirectes et les modèles graphiques simples ne peuvent inclure de covariables en tant que facteurs environnementaux qui déterminent l'abondance du microbiote. De plus, les normalisations existantes sont souvent basées sur une transformation logarithmique, ce qui est quelque peu arbitraire et affecte donc les résultats de manière inconnue.

L'étude de la bibliographie sur les méthodes d'inférence de réseaux d'associations et l'étude de sensibilité des outils existants dans le contexte d'analyses metagénomiques nous a conduit à identifier plusieurs pistes de développement intéressantes pour aborder ces problèmes. Malgré le foisonnement de publications et de logiciels émergents sur

ce sujet, nous avons considéré que les pistes identifiées n'étaient pas encore testées et implémentées de manière satisfaisante. Je me suis donc engagé dans le développement d'une nouvelle méthode, appelée MAGMA, pour détecter les interactions entre OTUs, qui prend en compte la structure bruitée des données issues de la métagénomique, impliquant (i) un excès de zéros, (ii) une surdispersion, (ii) une compositionnalité et (iv) une éventuelle inclusion de covariables. La méthode est basée sur le modèle graphique gaussien de copules dans lesquels nous modélisons les distributions marginales avec des modèles linéaires généralisés « zero-inflated » binomiale négative. L'inférence est basée sur une procédure d'imputation par la médiane efficace, associée au lasso graphique.

Dans une étude complète de simulation, nous montrons que notre méthode surpasse les méthodes existantes d'inférence de réseaux d'associations microbiennes. De plus, l'analyse d'un jeu de données microbiennes 16S avec notre méthode révèle une nouvelle biologie intéressante.

Afin de développer ce projet, j'ai établi une interaction avec le professeur Ernst Wit de l'université de Groningen aux Pays-Bas. J'ai obtenu trois financements, via le metaprogramme MEM, le programme de formation doctorale EIRA et le réseau européen COSTNET, qui m'ont permis de travailler en collaboration avec ce spécialiste de l'analyse de réseaux durant quatre mois sur deux séjours.

La méthode développée a été implémentée sous la forme d'un package R mis à disposition de la communauté scientifique. Notre projet de publication, qui a fait l'objet d'une pré-publication, doit être prochainement re-soumis à un journal scientifique à la lumière des commentaires obtenus lors d'une première soumission infructueuse auprès de la revue bioinformatics. Ces travaux ont été présentés dans le cadre des rencontres annuelles du pôle des microbiologistes clermontois en 2019.

MAGMA: inference of sparse microbial association networks

Arnaud Cougoul[1*], Xavier Bailly[1] and Ernst C. Wit[2]

February 1, 2019

1 UMR Epidemiology of Animal and Zoonotic Diseases, Universit Clermont Auvergne, INRA, VetAgro Sup, Saint-Genès-Champanelle, France

2 Institute of Computational Science, Università della Svizzera italiana, Lugano 6900, Switzerland

* To whom correspondence should be addressed.

Contents

Abstract

Motivation: Microorganisms often live in symbiotic relationships with their environment and play a central role in many biological processes. They form a complex interacting system with emergent functionality. Understanding the mechanisms that govern this ecosystem is an important scientific challenge, made easier through recently technologies facilitating the acquisition of microbiota metagenomic data. Until now correlation-based network analysis and graphical modelling have been used to identify the putative interaction networks formed by the species of microorganisms, but existing methods do not take into account all features of microbiota data. Indeed, correlation-based network cannot distinguish between direct and indirect correlations and simple graphical models cannot include covariates as environmental factors that shape the microbiota abundance. Furthermore, the compositional nature of the microbiota data is often ignored or existing normalizations are often based on ad hoc transformations that affect the results in unknown ways.

Results: We have developed a novel method, called MAGMA, for detecting interactions between microbiota that takes into account the noisy structure of the microbiota data involving an excess of zero counts, overdispersion, compositionality and possible covariate inclusion. The method is based on Copula Gaussian graphical models whereby we model the marginals with zero-inflated negative binomial generalized linear models. The inference is based on an efficient median imputation procedure combined with the graphical lasso. We show that our method beats all existing methods in recovering microbial association networks in an extensive simulation study. The analysis of two 16S microbial data studies with our method reveals interesting new biology.

Availability and implementation: MAGMA is implemented as an R-package and is freely available at `https://gitlab.com/arcgl/rmagma`. Upon acceptance the package will also be released on CRAN.

Contact: arnaud.cougoul@inra.fr

Supplementary information: The Git repository also includes the scripts used to prepare the material of this paper.

1 Introduction

Microbiota are ubiquitous and play a central role in biological processes [1]. High-throughput sequencing allows to study the composition, structure and diversity of complex microbial communities. In the wake of technological development, there has been during the last years a multiplication of projects querying the structure and properties of specific microbiota. Among others, some large projects targeted the human microbiome, e.g., the MetaHIT project [2, 3] and the HMP project [4, 5, 6], planktonic and coral ecosystems of the different oceans (TARA Oceans) project [7, 8], or the earth's multiscale microbial diversity (EMP) project [9, 10].

Microbiota are by nature complex systems of interconnected taxa. Interactions among microbes are an important factor that shape the structure and properties of microbiota. From an ecological point of view, interactions appear to structure [11], stabilize [12] and regulate the diversity [13] of microbial communities. In the biomedical field the dysbiosis of the human gut microbiota is associated with multiple pathologies such as obesity [14], diabetes [15] and mental illness [16]. Metagenomics opens a field of exploration of potential associations between the microbiome and several complex diseases [17]. Global modifications of a microbiota can also have implications for the dynamics of a bacteria of particular interest. In epidemiology the infection of a host by a pathogen can be facilitated by some microbial species through various interaction processes [18]. Conversely, some microbial species may have antagonistic interactions with pathogens that could be used in biological control [19, 20].

Identifying potential microbial interactions from metagenomic data is therefore a topical scientific challenge. Methodological developments are needed to improve this identification, taking into account the noisy and stochastic structure of the genomic measurement process of the microbiota.

1.1 Metagenomic data characteristics

Metagenomic data from 16S rRNA sequencing consists of sequencing reads originating from thousands of different bacterial groups obtained from hundreds to thousands of samples [21]. In order to reflect the microbial composition and the relative frequency of each bacterial group among samples, sequencing reads are clustered in Operating Taxonomic Units (OTU) [22], e.g., bacterial species. The number of OTUs considered depends both on the studied microbial community and the criteria used to cluster sequences in OTUs.

Metagenomic read counts are sparse and overdispersed. Most of OTUs are rare and occur in only a few samples. The sequencing read data therefore have a large amount of zeros [23], which often in naive analyses causes spurious associations [24]. Zero-inflated (ZI) distributions thus appear to be the most appropriate to model OTU abundances. Furthermore, the abundance of an OTU, defined as the number of reads assigned to this OTU, does not follow a usual count distribution such as a zero-inflated Poisson distribution as typically overdispersion is observed when the OTU is present. It has been shown that ZI negative binomial or ZI lognormal provide a good fit [25, 26].

Another aspect to take into consideration is the sequencing depth of a sample, which is defined as the sum of all OTU read counts in a sample. Sequencing depths are unequal among samples due to experimental effects [27]. From this perspective, metagenomic sequencing read data should be considered compositional in nature [28, 29]. For a given sample, each OTU read abundance "depends" on the other OTU reads through the sequencing depth. various ways have been suggested for taking the sequencing

depth into account when analyzing the observed read counts for an OTU. A common way is to circumvent the compositional nature of the data and to make OTUs comparable by transforming and normalizing the OTU table before further analysis. Main methods are rarefying, scaling and log ratio transformation, but all have problematic aspects [25, 26, 30]. A typical scaling transformation method is to divide by the marginal sum of the sample, i.e., the sequencing depth. This method leads to spurious negative associations [31]. The centered log ratio (clr) transformation [32] and relative log expression [33] are most commonly used to process compositional sequencing data. The microbial data is mainly composed of zeros and the log cannot be applied without replacing zero values by a pseudo-count and is therefore not ideal [26, 34].

The diversity of microbiota among samples furthermore depends on factors that are known to structure the distribution of microbes such as environmental conditions, spatial and temporal scales. For instance, age, genetics, environment and diet are all factors that affect the human gut microbiota [35]. Seasonal changes in the microbiota of wild mice have also been observed [36]. These factors should be considered as much as possible in the analysis by integrating them as covariates in the model to separate biological interactions from the effects of structuring covariates.

1.2 Inference of microbial associations networks

In this paper, we take the perspective to consider microbial communities as a network of microbial species (or OTUs) that interact with each other. These networks are formalized by graphs consisting of vertices representing OTUs and edges representing statistical dependencies, i.e. associations, between OTUs. Network analysis is the most common approach to explore potential microbial interactions at the microbiota scale. There are two main ways to infer a microbial network: correlation-based networks and graphical models [37].

On the one hand, correlation-based networks are graphs obtained from computing and thresholding all pairwise association measures. A large number of association measures have been used in this framework, such as correlation (e.g., Pearson, Spearman), similarity (e.g., mutual information), or dissimilarity (e.g., Kullback-Leibler) measures [38]. These methodologies rely on pairwise associations between occurrences or abundances of bacterial OTUs among the microbiota. A permutation and bootstrap approach can be used to improve the robustness of the infered network [39]. The main disadvantage of pairwise association methods is that they are unable to distinguish between direct and indirect associations, thereby often ending up with dense network that give little insight in the underlying functional relations.

On the other hand, graphical models have minimal bias and better power [37, 40]. Graphical models are graphs that satisfy the Markov properties, which means that links represent conditional dependencies. In the multivariate Gaussian case, conditional dependence is equivalent to a non-zero partial correlation. In a such framework, the conditional dependencies can be read off from the inverse correlation matrix, called the precision matrix. Inference of Gaussian graphical models can be performed by neighborhood selection [41] or by lasso regularization [42]

The two main methods used for exploration of microbial interactions are SPIEC-EASI [40] and SparCC [43]. SparCC estimates linear Pearson correlations between the log-transformed components. The algorithm works by iteratively calculating a "basis correlation" under the assumption that the majority of pairs do not correlate [43]. SPIEC-EASI normalizes the data with the *clr* transformation before applying the classical framework of Gaussian graphical models described below. Both methods use a pseudo-counts to avoid zeros and can not take into account potential covariates.

Current network inference methods such as SPIEC-EASI and SparCC do not fully consider the structure of metagenomic data involving sparsity, overdispersion, compositionality or covariate inclusion. We therefore propose a novel inference framework involving copula Gaussian graphical models [44]. This model provides a general and integrative framework for network inference. We called our method MAGMA for **M**icrobial **A**ssociation **G**raphical **M**odel **A**nalysis. MAGMA allows to take into account all aspects of the data, while relying on the well-known properties of a latent Gaussian graphical model. We implemented our method in R and provide a package called `rMAGMA` available on a Git repository at `https://gitlab.com/arcgl/rmagma`.

2 Materials and Methods

Here we present an original way of integrating metagenomic data for the exploration of microbe-microbe interactions. We propose a copula Gaussian graphical model combined with GLM marginal distributions. Although full likelihood inference is possible, our MAGMA approximation is based on the estimation of

the latent data by the median of possible values. This mapping makes it possible to manage the excess of zeros, overdispersion, the compositional nature of the data and the inclusion of covariates.

2.1 Model

In the classical Gaussian graphical model (GGM), we consider a centred multivariate normal Z with a correlation matrix Θ^{-1},

$$Z \sim \mathcal{N}\left(0, \Theta^{-1}\right). \tag{1}$$

Computing the precision matrix Θ gives informations about partial correlations between elements of Z [45]. Under the multivariate Gaussian assumption, the partial correlation ρ_{ij} between i and j is given by:

$$\rho_{ij} = -\frac{\Theta_{ij}}{\Theta_{ii}\Theta_{jj}}. \tag{2}$$

Non-zero elements in the precision matrix Θ correspond to the conditional dependencies and edges in the conditional dependence graph.

The observed metagenomic count data, unfortunately, do not follow a normal distribution. Microbiota data are represented by a matrix Y of $n \times p$ dimension, where n is the number of samples and p is the number of OTUs. We assume that the joint distribution of observed variables Y can be transformed from a latent multivariate normal variable Z. The copula Gaussian graphical model defines the marginal transformations [44],

$$Y_{ij} = F_{ij}^{-1}\left(\Phi\left(Z_{ij}\right)\right), \tag{3}$$

where Φ is the cumulative distribution function (cdf) of the standard normal distribution and F_{ij}^{-1} is the inverse cdf of microbiota count Y_{ij} for the j^{th} OTU and for sample i.

The F_{ij} function is generally estimated by the empirical cdf $\hat{F}_j$ [46, 47], but this is not appropriate here as F_{ij} will certainly depend on the sequencing depth of sample i and therefore cannot be constant across samples. Instead, we assume that OTU read abundances are distributed according to a zero-inflated negative binomial (ZINB). We introduce the original mapping function:

$$F_{ij} \sim \text{ZINB}\left(\lambda_{ij}, \theta_j, \pi_j\right), \tag{4}$$

where λ_{ij} is the mean of the negative binomial part for sample i and species j, θ_j is the dispersion parameter and π_j is the probability of the structural zeros. The mean λ_{ij} is defined by the equation:

$$\log\left(\lambda_{ij}\right) = \beta_j + X_i\,{}^t\gamma_j + \log(\sigma_i). \tag{5}$$

β_j is modelling the mean of species j, γ_j is the effect of covariates X on species j and σ_i is the library size or sequencing depth for sample i.

With this parametric mapping function, we can model the high proportion of zeros in data by the use of a zero-inflated distribution. We model overdispersion by the negative binomial distribution. We model sequencing depth to take into account compositionality by an offset. And we also model the effect of covariates, either qualitative or quantitative, on the mean of microbial abundance.

2.2 MAGMA inference

Full likelihood inference of the above model is involved. We propose here a computational approximation of the maximum likelihood. If Y were continuous data, then observed variables could be projected into the latent space by the inverse mapping,

$$Z_{ij} = \Phi^{-1}\left(F_{ij}\left(Y_{ij}\right)\right). \tag{6}$$

But since Y_{ij} are discrete count data, F_{ij}^{-1} is not injective and the projection in the latent space is not unique. F_{ij} is a step function and Z_{ij} can take all the values in the interval $\left[\Phi^{-1}\left(F_{ij}\left(Y_{ij}-1\right)\right), \Phi^{-1}\left(F_{ij}\left(Y_{ij}\right)\right)\right]$.

To approximate the copula Gaussian graphical model, the nonparanormal normal score approach [48] takes the right bound value $\Phi^{-1}\left(\hat{F}_{ij}\left(Y_{ij}\right)\right)$ and winsorizes the data for the highest observed values to avoid infinite values. The nonparanormal SKEPTIC transformations [49] use the asymptotic relationships

between the Pearson correlation and the Spearman or Kendall rank correlations. Instead, we propose to transform the count data using the median point of the Z distribution of reachable values,

$$\tilde{Z}_{ij} = \Phi^{-1}\left(\frac{\hat{F}_{ij}\left(Y_{ij}-1\right)+\hat{F}_{ij}\left(Y_{ij}\right)}{2}\right). \tag{7}$$

$\tilde{Z}_{ij}$ thus defined is the median of the normal distribution between $\Phi^{-1}\left(\hat{F}_{ij}\left(Y_{ij}-1\right)\right)$ and $\Phi^{-1}\left(\hat{F}_{ij}\left(Y_{ij}\right)\right)$. With this estimation, we do not need to winsorized the data nor rely on dubious asymptotic relationships that certainly do not hold.

For estimating the library size σ_i of sample i in (5), the sample sequencing depth ignores the fact that different biological samples may express different 16S RNA repertoires [50]. We estimate the library size using the geometric mean of pairwise ratios (GMPR) [51]. GMPR is specifically intended for compositional zero-inflated data as the microbiome sequencing data. For each pair of samples i and i', the median of count ratios of nonzero counts is computed,

$$r_{ii'} = \underset{\left\{j \,|\, Y_{ij}, Y_{i'j} \neq 0\right\}}{\text{median}} \left(\frac{Y_{ij}}{Y_{i'j}}\right). \tag{8}$$

The ratio $r_{ii'}$ represents how much, on average, the OTU read counts of sample i are above or below those of a sample i'. If $r_{ii'} = 2$, the OTU of the sample i will have on average 2 times more read counts than those of sample i'. To estimate the library size factor of a sample i, we then compute the geometric mean of all the ratios $r_{ii'}$ involving the sample i. This is the average difference between the abundance of an OTU found in sample i and its abundance in the other samples,

$$\hat{\sigma}_i = \left(\prod_{i'=1}^{n} r_{ii'}\right)^{1/n}. \tag{9}$$

The GLM (5) is then estimated with off-sets $\{\log(\hat{\sigma}_i)\}_i$, which then allows us to calculate the quasi-normal data $\tilde{Z} = \{\tilde{Z}_{ij}\}_{ij}$ according to (7).

Finally, we propose to infer the association network from the transformed data $\tilde{Z}$ of the observed variable Y. In this way, we approximately infer the copula Gaussian graphical model, taking into account the characteristics of the microbial data to infer relevant associations between OTUs. We use graphical lasso (glasso) inference [42] from the R `huge` package to estimate a sparse precision matrix. In the sparse estimation of the precision matrix Θ, the problem is to maximize the penalized log likelihood

$$l_{pen}\left(\tilde{Z}, \Theta\right) = \log|\Theta| - \text{trace}\; S\Theta - \rho\,\|\Theta\|_1. \tag{10}$$

S denotes the empirical covariance of the $\tilde{Z}$ transformed data matrix, $\|\Theta\|_1$ is the L_1 norm and $\rho \in \mathbb{R}_0^+$ is a sequence of non-negative penalty parameters.

Penalized inference of graphical models results in a collection of OTU networks associated with the estimated precision matrix $\widehat{\Theta}_\rho$ for different values of ρ. In order to infer the most parsimonious network given the available data, one need to weigh the fit of the data relative to the complexity of the data [52, 53]. To select the penalty parameter ρ, we consider three approaches: rotation information criterion (ric) [54], stability approach for regulation selection (stars) [55] and extended Bayesian information criterion (ebic) [56]. All these approaches are encoded in the R package `huge` used by MAGMA.

In summary, MAGMA inference comprises of the following steps:

1. Adjust the marginal OTU abundances to ZINB distributions according to equations (4) and (5).
2. Approximate the latent data Z according to equation (7).
3. Estimate a sparse precision matrix $\widehat{\Theta}_\rho$ according to equation (10).
4. Select the penalty ρ^* that best balances fit and complexity via ric/stars/ebic.
5. Identify the OTU network from non-zero elements of $\widehat{\Theta}_{\rho*}$.

3 Results and Discussion

We studied the efficiency of the MAGMA tool to infer a network of microbial associations. With this aim, we first analyzed the behavior of MAGMA on simulated data in section 3.1. We measured the quality of network inference under different conditions and compared MAGMA with other network approaches. In section 3.2 we applied MAGMA to data from the Human Microbiome Project.

3.1 Simulation study

In this section, we first describe how we generate simulation data. We then studied six different aspects of the MAGMA model with respect to this simulated data: (i) its consistency, i.e., whether it converges to the true network with increasing number of samples n, (ii) its robustness, i.e., whether it is able to deal with deviations from ZINB read counts, (iii) its ability to infer the network with varying interaction strengths, (iv) how its ability to reconstruct the network depends on different network topologies, (v) its ability to account for confouding by integrating a covariates, and finally, (vi) we compared MAGMA with existing tools for the inference of microbial association networks. The ability of the procedure to recover the simulated microbial network was measured via the area under the ROC curve (AUC) along the ρ-path of the inferred networks.

3.1.1 Generation of realistic data sets

To measure the performance of network inference tools, we should simulate datasets of known structure and tried to recover the associations that we simulated. SPIEC-EASI [40] proposes a simulation procedure, however it is unable to reproduce variations in sequencing depth, which is considered an essential feature [57]. Our procedure first generates an association network G with d vertices and e edges (Figure 1). The topology of the generated network can be selected to be either band, block, cluster, hub,

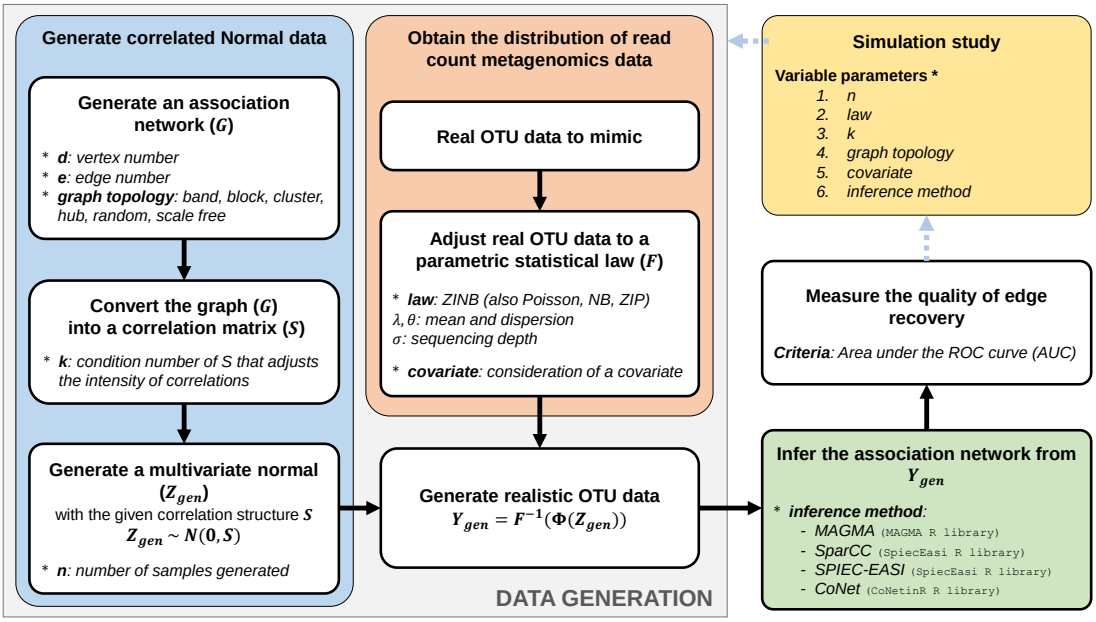


Figure 1: Workflow of the generation of realistic data for the inference benchmarking.

random or scale free as defined in [40]. We associate the simulated graph with an inverse correlation matrix fixing the condition number of the matrix k regulating the strength of correlations. We then generate multivariate normal data with the obtained correlation structure.

Then we need to transform the latent normal data into the observed read count data. To mimic the structure of real data, we relied on the 16S data of the microbiome of Puerto Rico honey bees obtained by MG Dominguez-Bello [58, study ID 1064]. We filtered the data, keeping the 80 OTUs with a prevalence greater than 15% and 286 samples with a sequencing depth greater than 100 reads. The average sequencing depth was 19,000. The data has been fitted according to some parametric distribution, e.g., the ZINB used in our network inference, but also other distributions: Poisson, zero-inflated Poisson and

negative binomial. Using the copula transformation, we project the multivariate normal data into read counts using the selected marginal distributions combined with the logarithmic link function involving covariates and an offset.

3.1.2 Effect of the sample size

Data were simulated with different number of samples n. We then inferred the association network and measured the quality of edge recovery as shown in Figure 2A. As the number of samples increases, the AUC increases and tends to one. Asymptotically, the method correctly recovers all the simulated links. The approximation of the copula Gaussian graphical model made by MAGMA allowed to recover the network with hundreds of samples. With 200 simulated OTUs, 200 to 300 of samples are sufficient recover almost the entire network correctly.

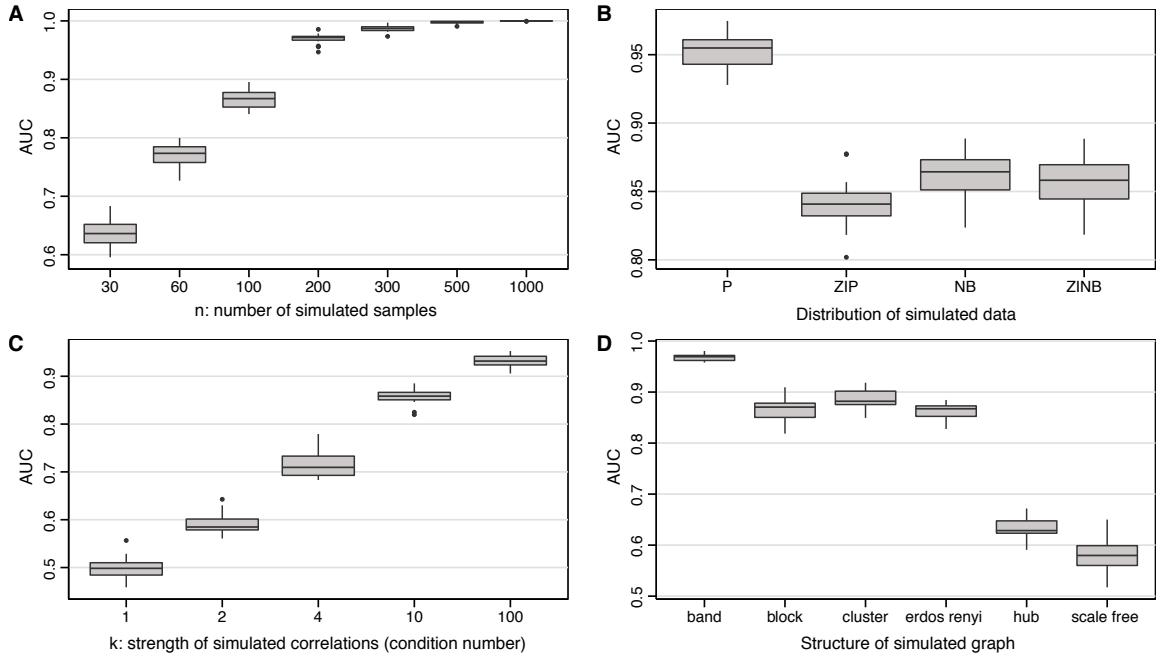


Figure 2: Effect of varying parameters on the quality of network inference. Boxplots of the AUC criterion according to: (**A**) the number of simulated samples n varying from 30 to 1000; (**B**) the distributions of the simulated data (Poisson, Zero-inflated Poisson, Negative Binomial and Zero-inflated Negative Binomial); (**C**) the condition number of the simulated correlation matrix k varying from 1 to 100; (**D**) the structure of simulated graph (band, block, cluster, hub, random and scale free). If they did not vary, the parameters were fixed at: $n = 100$, $k = 10$, random graph structure (Erdos-Renyi), marginal count data simulated according to a ZINB. We considered 20 simulation iterations for networks of size 200 with an average degree of 2.

3.1.3 Effect of the distribution of read counts

To measure the flexibility of our method to model misspecification, we varied the distribution of the simulated data. The results are shown in Figure 2B. Zero-inflated Poisson, negative binomial and zero-inflated negative binomial all performed roughly similar, suggesting that MAGMA is quite robust to model misspecification, as it assumes underlying ZINB data. It is striking that dependence networks with underlying Poisson distributed read count data were able to be reconstructed significantly better (average AUC > 0.95), suggesting that the zero-inflation and, particularly, over-dispersion makes network reconstruction more difficult (average AUC ≈ 0.85).

3.1.4 Effect of the strength of partial correlations

The strength of the simulated correlations was modelled by the condition number of correlation matrix of the simulated data. A low condition number corresponds to small values of the coefficients of the

correlation matrix and this will produce weak links. The results are shown in Figure 2C. With a condition number of 1, its lowest possible value, the correlations have no strength and the results obtained were the same as a random draw with an average AUC of 0.5. The AUC increases rapidly with an increasing condition number. For $k = 4$, we found an AUC at 0.72 and for $k = 10$ the AUC was already at 0.85.

3.1.5 Effect of network topologies

As Figure 2D shows, network topology has, perhaps surprisingly, a significant impact on network reconstruction quality. Simulations were done for different kind of graph structures. The band graph has the best reconstruction properties (average AUC > 0.95). On the other end, the recovery of a scale free or a hub network was difficult (average AUC $\in (0.55, 0.65)$). It seems that high-degree nodes pose a problem with network inference. This is a common issue also with other methods [40]. The reason why the band topology can be easily reconstructed may be because it has the lowest maximum node degree of all topologies. The results for the block, cluster and random networks were good with an AUC above 0.85.

3.1.6 Consideration of a covariate

In order to check the capacity of our method to account for confounding in the dependence network, we used MAGMA with the inclusion of a quantitative covariate. We generated read count data by adding a covariate effect with different levels of strength. The coefficients of the covariate, γ_j in (5), for all OTUs were sampled from a normal distribution with variance equal to 0, 1, 2 or 4. The mean of γ is taken to be zero, as it is just an offset, confounded with the sampling effort. The values of the unit specif covariate are sampled from a standard normal distribution.

We compare the effect of including and ignoring the covariate effect across different levels of confounding. As Figure 3 shows, when there is in fact no confounding, using MAGMA containing an irrelevant covariate does not result in more errors than using MAGMA without the covariate. The addition of an irrelevant covariate effect to the method does not have a negative impact on the AUC. As the strength of the confounding increases, MAGMA that accounts for this confounding has an increasing advantage in recovering the network structure over applying MAGMA that ignores the covariate. We conclude that careful modelling of the read count distribution F_{ij} is particularly relevant: the inference quality of the association network relative to the agnostic MAGMA increases when the covariate effect is gets stronger.

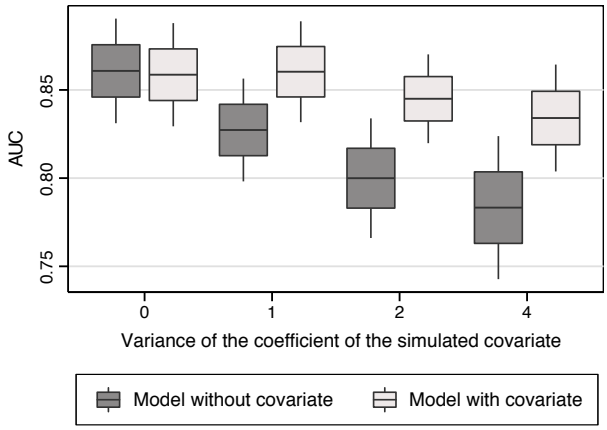


Figure 3: The network recovery ability (AUC) as a function of the confounding covariate strength γ for the agnostic MAGMA method vs. the MAGMA method with covariate effect. We considered 20 simulation iterations for networks of size 200 with an average degree of 2, number of samples $n = 100$, conditioning number $k = 10$, graph structure was random, read count data were simulated according to a ZINB.

3.1.7 Comparison with other association network approaches

We compare the existing methods regarding the presence or absence of structure among samples due to a covariate. As shown in Figure 4, MAGMA showed better performances than the three reference

methods in reconstructing microbial interactions, namely SparCC, CoNet and SPIEC-EASI. The networks recovered by CoNet are derived from the calculation of Spearman correlation p-values by permutation and bootstrap. In our simulations, this did not have an added value compared to the networks obtained from Spearman correlations thresholding. The Pearson correlation network and the graphical lasso model on raw data did not work well without data normalization: linear correlations should not be calculated from raw read count data. The graphical lasso with nonparanormal SKEPTIC transformation had a higher AUC than that obtained with SparCC and SPIEC-EASI; yet this non-parametric transformation is typically not used for the study of microbiota data. In the presence of a covariate, the performance of all competing methods degraded significantly and the AUC dropped. Under our simulations, MAGMA inference yielded the best performance. We therefore conclude that it is essential to take into account the potential covariates with structural effects on the microbiota.

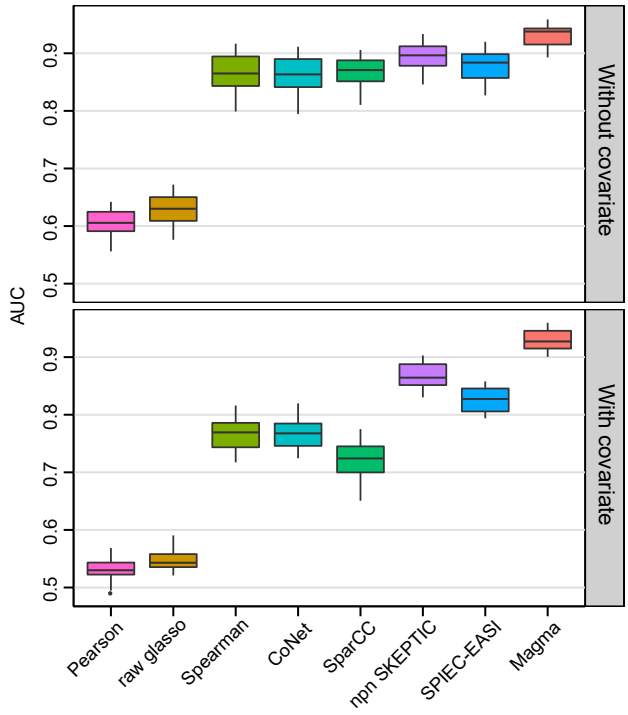


Figure 4: Comparison of inference methods considering a covariate effect. Boxplots of the AUC criterion for different network inference methods. We considered 20 simulation iterations for networks of size 200 with an average degree of 2, number of samples $n = 300$, condition number $k = 5$, graph structure was random and read count data were simulated according to a ZINB. The quantitative covariate parameter γ was drawn from a $N(0, 2)$, when covariate was effective. CoNet designates network obtained from Spearman correlation p-values by 100 iterations of permutation and bootstrap. For SparCC, the correlation threshold parameter was equal to 0.3 and 100 iterations were done in the outer loop and 20 in the inner loop. raw Glasso, npn SKEPTIC, SPIEC-EASI, and MAGMA were network obtained by graphical lasso inference from raw data, nonparanormal SKEPTIC transformation, clr transformation and MAGMA transformation respectively. For Pearson, Spearman and SparCC networks, we computed the path of inferred networks by thresholding the correlations. For CoNet, we thresholded the p-values. For the graphical lasso, we varied the regularization parameter.

3.2 Microbial data illustration: Human Microbiome Project

In this section, we present the analysis of the 16S variable region V3-5 data from the Human Microbiome Project (HMP) [4, 5]. The study collected microbiomes of healthy individual at various body sites. The data was retrieved on the qiita data platform [58, study ID 1928]. This study brings together a total of 6,000 samples from 18 different the human body sites. A total of 10,000 microbial species occupy the human ecosystem. We first studied a stool microbiota sample, comparing MAGMA with SparCC and SPIEC-EASI. Second, we analyzed the stool and saliva microbiota in a single study in order to show the usefulness of the covariate implementation in the MAGMA method.

3.2.1 Inference of gut microbiota network

Among other roles, the gut microbiota is involved in nutrient metabolism and in the prevention of colonization by pathogenic micro-organisms. Getting insight into the functioning of this microbial ecosystem is therefore a critical scientific issue. Stool HMP data contains 388 samples and 10,730 OTUs, with most OTUs being rare. We filter out OTUs present in less than 25% of the samples and remove the samples whose sequencing depth is less than 500 reads on the remaining OTUs. These samples show large stochastic variability and in a properly weighted analysis would not add much information. After this preprocessing we obtain an OTU table with 360 samples and 306 OTUs.

Figure 5A show the stool network obtained by MAGMA, SparCC and SPIEC-EASI. With the *stars* selection from the `huge` R package, MAGMA and SPIEC-EASI selected a little over 2000 edges (2356 for MAGMA, 2332 for SPIEC-EASI). For comparative purposes each network is shown with the same amount of 2000 edges. Figure 5B shows the network node degree distributions as well as the Venn diagrams of the inferred links. SparCC network has the wides distribution with high degree nodes for both positive and negative association links. Regarding positive links, the SPIEC-EASI network has more nodes characterized by low degrees than the other methods.

The three networks show a strong antagonism between the groups of the Firmicutes and Bacteroidetes phila. MAGMA network showed the most tempered opposition between this two groups and has fewer negative links (100) than the other networks (486 for SparCC and 533 for SPIEC-EASI). Less than half of the negative links recovered by SparCC and SPIEC-EASI were identical, raising questions about their veracity. The MAGMA stool network has more positive links: 25% and 30% more than SparCC and SPIEC-EASI respectively. Relative to this, only 33% of positive links recovered by MAGMA differed from those found by these two methods. Compared to other tools, MAGMA seems to identify a coherent network with sensible biological structure, and it showed a good reproducibility of results compared to the other methods.

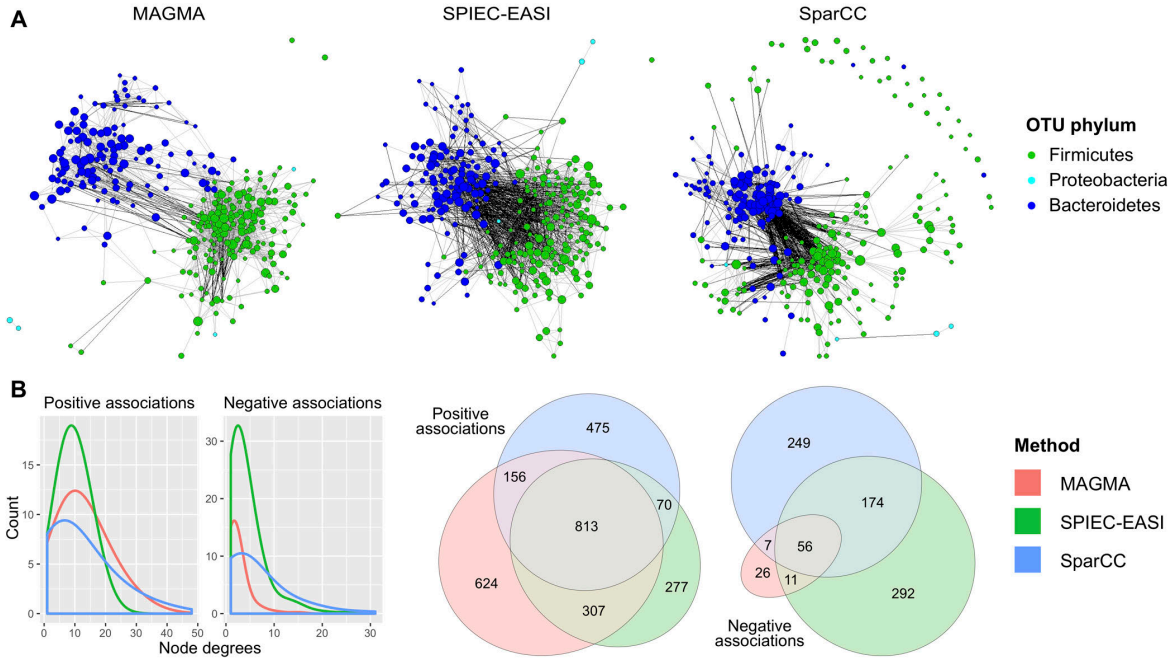


Figure 5: Stool microbiota network. (**A**) Stool microbial association network obtained from three methods MAGMA, SPIEC-EASI and SparCC. Nodes are OTUs. Black and gray links represent negative and positive associations, respectively. (**B**) positive and negative associations obtained from the three networks: smoothed histograms of node degrees and Venn diagrams representing the overlap of inferred links between the different methods.

3.2.2 Microbial network body site variation

To illustrate MAGMA's ability to account for confounding or, put differently, to analyze information across heterogeneous samples, we pool two different sets of HMP microbiota and introduce a covariate.

We group gut microbiota data and salivary microbiota data and introduce a, probably marked, "body site" effect. Again, we filter out OTUs present in less than 25% of the samples and remove samples with sequencing depth of less than 500 reads on the remaining OTUs. This results in an OTU table with 665 samples and 245 OTUs.

Figure 6 shows the two networks we obtain with and without integration of the body site covariate in MAGMA. In the network *without* covariate (Figure 6A), two sets of OTUs were stand out. A first group with Firmicutes and Bacteroidetes phyla corresponds to intestinal microbiota and a second group corresponds to salivary microbiota. The OTUs of the same group are positively associated with each other, while two OTUs of different groups are negatively related. In the network *with* covariate (Figure 6B), there were again two groups of OTUs. This time the spurious negative links between the two groups disappear, because the difference in frequency of OTUs between body sites has been taken into account by means of the body site covariate, which allows to find real functional interactions between the various OTUs. This includes various positive associations between the two groups of OTUs. In fact, there are no common negative links between network A and B. Negative correlations due to the average body site effect are shifted to 0 when normalizing by considering the covariate (Figure 6C). Positive correlations due to OTU co-presence in a specific microbiota are centered when including the covariate. Taking into account the body site in MAGMA makes it possible to obtain a *consensual* network.

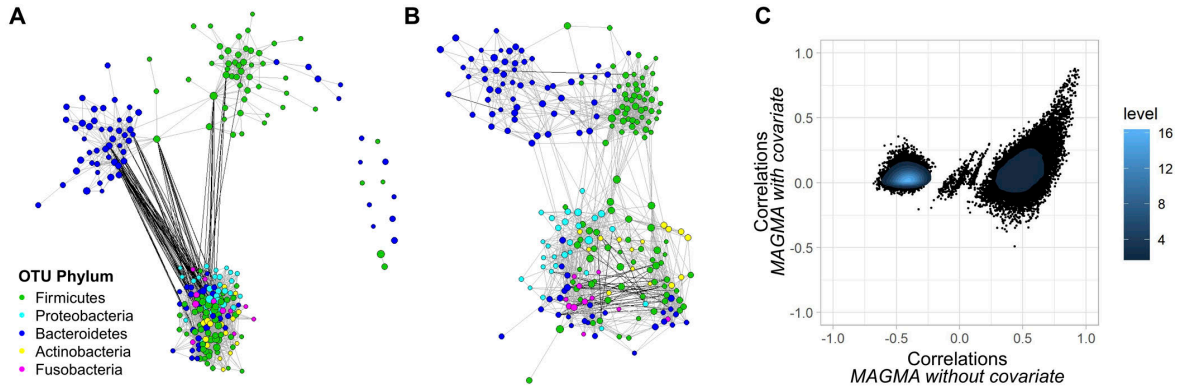


Figure 6: Association network of stool and saliva microbiota pooled data. (**A**) Stool and saliva microbial association network without body site covariate. (**B**) Stool and saliva microbial association network including body site factor. Regularization parameters for the two networks were determined with stars selection. (**C**) Correlations of MAGMA without including body site covariate versus correlations of MAGMA including body site effect. Correlations were computed from Pearson correlations of MAGMA transformed data (normalization defined by equation (7)).

4 Conclusion

We have introduced a network model that responds to the methodological challenges arising from sequencing read count data: excess of zeros, over-dispersion, compositionality and the presence of covariates. To meet these challenges, the network inference method we propose takes advantage of a GLM-inspired parametric mapping function, while being based on the well-known Gaussian graphical model. MAGMA offers a normalization approach based on the theory of copulas. Moreover, it takes into account variable sequencing depth estimating the library size effect by the geometric mean of pairwise ratios.

In the simulation studies we show that the approximations made during the transformation of the data rapidly converge towards the correct solution when the number of samples and the strength of the correlations increase. The ZINB law we propose is flexible and can deal easily with moderate amount of model misspecification. The integration of covariates improves the quality of the inference in presence of structural factors affecting the OTU read counts. Finally, MAGMA performs better than other available competitors in a wide variety of situations.

We have applied MAGMA to infer an intestinal microbial network from a HMP data, allowing for heterogeneous samples. The resulting network shows a consensual interactions that are not affected by wildly different OTU counts for various body sites. All this shows that MAGMA is a practical tool for inferring microbial functional networks from metagenomic sequence read count data.

Acknowledgments

The work was funded by two INRA metaprogrammes: Meta-omics of microbial ecosystems (MEM) and Integrated management of animal health (GISA) and by the European Cooperation for Statistics of Network Data Science (COSTNET: COST Action CA15109).

References

[1] Tobias Rees, Thomas Bosch, and Angela E. Douglas. How the microbiome challenges our concept of self. *PLOS Biology*, 16(2):e2005358, feb 2018.

[2] Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.

[3] Manimozhiyan Arumugam, Jeroen Raes, Eric Pelletier, Denis Le Paslier, Takuji Yamada, et al. Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–180, 2011.

[4] The Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, jun 2012.

[5] The Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214, jun 2012.

[6] Karoline Faust, J. Fah Sathirapongsasuti, Jacques Izard, Nicola Segata, Dirk Gevers, Jeroen Raes, and Curtis Huttenhower. Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology*, 8(7), 2012.

[7] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, et al. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359–1261359, may 2015.

[8] Gipsi Lima-Mendez, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, et al. Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1262073_1–1262073_9, may 2015.

[9] Luke R. Thompson, Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, 551(7681):457–463, nov 2017.

[10] Jack A. Gilbert, Janet K. Jansson, and Rob Knight. The Earth Microbiome project: successes and aspirations. *BMC Biology*, 12(1):69, dec 2014.

[11] Didier Gonze, Leo Lahti, Jeroen Raes, and Karoline Faust. Multi-stability and the origin of microbial community types. *The ISME Journal*, 11(10):2159–2166, oct 2017.

[12] Jacopo Grilli, György Barabás, Matthew J. Michalska-Smith, and Stefano Allesina. Higher-order interactions stabilize dynamics in competitive network models. *Nature*, 548(7666):210–213, 2017.

[13] Battle Karimi, Pierre Alain Maron, Nicolas Chemidlin-Prevost Boure, Nadine Bernard, Daniel Gilbert, and Lionel Ranjard. Microbial diversity and ecological networks as indicators of environmental quality. *Environmental Chemistry Letters*, 15(2):265–281, 2017.

[14] Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, et al. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546, aug 2013.

[15] Helle Krogh Pedersen, Valborg Gudmundsdottir, Henrik Bjørn Nielsen, Tuulia Hyotylainen, Trine Nielsen, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, 535(7612):376–381, jul 2016.

[16] Qinrui Li, Ying Han, Angel Belle C. Dy, and Randi J. Hagerman. The Gut Microbiota and Autism Spectrum Disorders. *Frontiers in Cellular Neuroscience*, 11(April), apr 2017.

[17] Jun Wang and Huijue Jia. Metagenome-wide association studies: fine-mining the microbiome. *Nature Reviews Microbiology*, 14(8):508–522, 2016.

[18] Elise Vaumourin, Gwenaël Vourc'h, Patrick Gasqui, and Muriel Vayssier-Taussat. The importance of multiparasitism: examining the consequences of co-infections for human and animal health. *Parasites & vectors*, 8:545, 2015.

[19] Boris Jakuschkin, Virgil Fievet, Loïc Schwaller, Thomas Fort, Cécile Robin, and Corinne Vacher. Deciphering the Pathobiome: Intra- and Interkingdom Interactions Involving the Pathogen Erysiphe alphitoides. *Microbial Ecology*, 2016.

[20] R. Poudel, A. Jumpponen, D. C. Schlatter, T. C. Paulitz, B. B. McSpadden Gardener, L. L. Kinkel, and K. A. Garrett. Microbiome Networks: A Systems Framework for Identifying Candidate Microbial Assemblages for Disease Management. *Phytopathology*, 106(10):1083–1096, oct 2016.

[21] Miklós Bálint, Mohammad Bahram, A. Murat Eren, Karoline Faust, Jed A. Fuhrman, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. *FEMS Microbiology Reviews*, 40(5):686–700, sep 2016.

[22] Jolinda Pollock, Laura Glendinning, Trong Wisedchanwet, and Mick Watson. The Madness of Microbiome: Attempting To Find Consensus Best Practice for 16S Microbiome Studies. *Applied and Environmental Microbiology*, 84(7):e02627–17, feb 2018.

[23] Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D. Peddada. Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, 8(NOV):1–10, nov 2017.

[24] Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal*, 10(7):1669–1681, jul 2016.

[25] Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology*, 10(4):e1003531, apr 2014.

[26] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, dec 2017.

[27] David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger, and Chris P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[28] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(November):1–6, 2017.

[29] Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(March):2870–2878, 2018.

[30] Joseph N Paulson, O Colin Stine, Héctor Corrada Bravo, and Mihai Pop. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, 10(12):1200–1202, dec 2013.

[31] J Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

[32] J. Aitchison. The statistical analysis of compositional data: monographs in statistics and applied probability. *Chapman & Hall, London*, 1986.

[33] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 2010.

[34] Paul I Costea, Georg Zeller, Shinichi Sunagawa, and Peer Bork. A fair comparison. *Nature Methods*, 11(4):359–359, mar 2014.

[35] Catherine A. Lozupone, Jesse I. Stombaugh, Jeffrey I. Gordon, Janet K. Jansson, and Rob Knight. Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230, 2012.

[36] Corinne F. Maurice, Sarah Cl Knowles, Joshua Ladau, Katherine S. Pollard, Andy Fenton, Amy B. Pedersen, and Peter J. Turnbaugh. Marked seasonal variation in the wild mouse gut microbiota. *ISME Journal*, 9(11):2423–2434, 2015.

[37] Mehdi Layeghifard, David M. Hwang, and David S. Guttman. Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, 25(3):217–228, mar 2017.

[38] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, aug 2012.

[39] Karoline Faust and Jeroen Raes. CoNet app: inference of biological association networks using Cytoscape. *F1000Research*, 5:1519, 2016.

[40] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, may 2015.

[41] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

[42] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, jul 2008.

[43] Jonathan Friedman and Eric J Alm. Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9):e1002687, sep 2012.

[44] Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics*, 5(2 A):969–993, 2011.

[45] Joe Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing, 1990.

[46] Fentaw Abegaz and Ernst C Wit. Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica*, 69(4):419–441, 2015.

[47] Pariya Behrouzi and Ernst C Wit. Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(1):141–160, 2019.

[48] Han Liu, John Lafferty, and Larry Wasserman. The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

[49] Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

[50] Marie Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[51] Li Chen, James Reeve, Lujun Zhang, Shengbing Huang, Xuefeng Wang, and Jun Chen. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ*, 6:e4600, 2018.

[52] Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. all models are wrong...: an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

[53] Ivan Vujačić, Antonino Abbruzzo, and Ernst Wit. A computationally fast alternative to cross-validation in penalized gaussian graphical models. *Journal of Statistical Computation and Simulation*, 85(18):3628–3640, 2015.

[54] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge Package for High-dimensional Undirected Graph Estimation in R. *Journal of Machine Learning Research*, 13:1059–1062, 2012.

[55] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. *Advances in neural information processing systems*, 24(2):1–14, jun 2010.

[56] Rina Foygel and Mathias Drton. Extended Bayesian Information Criteria for Gaussian Graphical Models. *Advances in Neural Information Processing Systems*, 23(1):604–612, nov 2010.

[57] Nicholas J. Gotelli. Null model analysis of species co-occurrence patterns. *Ecology*, 81(9):2606–2621, 2000.

[58] Antonio Gonzalez, Jose A Navas-Molina, Tomasz Kosciolek, Daniel McDonald, Yoshiki Vázquez-Baeza, et al. Qiita: rapid, web-enabled microbiome meta-analysis. *Nature methods*, 15(10):796–798, 2018.

Chapitre 4

Applications et adaptations de la méthode MAGMA

Sommaire

Ce chapitre présente dans un premier temps une application directe de l'outil MAGMA. Par la suite, il illustre deux autres applications envisagées suite à des collaborations de recherche, qui appellent des développements de MAGMA pour prendre en compte les propriétés spécifiques des données étudiées. Pour chaque application, le contexte biologique est présenté. Le cas échéant, les développements méthodologiques nécessaires sont exposés.

4.1 Application de la méthode MAGMA avec covariable : exemple du microbiote de la tique à différents stades de développement

Les microbiotes peuvent être structurés par différents facteurs d'ordre biotique ou abiotique. Aussi, j'ai souhaité appliquer MAGMA avec un jeu de données permettant de prendre en compte un facteur structurant d'un microbiote afin d'inférer un réseau d'associations. Nous avons dans ce cadre étudié les associations entre les bactéries portées par des tiques de l'espèce *Ixodes ricinus* échantillonnées à différents stades de développement. *Ixodes ricinus* est un modèle biologique utilisé dans mon laboratoire d'accueil, l'UMR EPIA (Épidémiologie des maladies animales et zoonotiques).

Les tiques sont vecteurs d'agents pathogènes qui provoquent des maladies transmissibles à l'homme, notamment la maladie de Lyme. La maladie de Lyme est la maladie vectorielle la plus importante en prévalence en France métropolitaine. Elle est causée par une bactérie du complexe d'espèces *Borrelia burgdorferi*.

La tique comporte trois stades de développement. La larve se nourrit sur un hôte généralement de petite taille pour muer en nymphe. Après un deuxième repas de sang, la nymphe passe au stade adulte. L'adulte a ensuite son dernier repas puis la femelle s'accouple et pond plusieurs milliers d'œufs. Le mâle meurt après l'accouplement et la femelle après la ponte.

Le microbiote sanguin de l'hôte est transmis à la tique lors des repas de sang avec potentiellement des agents pathogènes présents. C'est une importante source d'infection pour les tiques. Toutefois, certains microbes ont une probabilité de transmission verticale, de la femelle à la larve, négligeable pour ce qui est des *Borrelia* responsables de la maladie de Lyme.

Ainsi, un projet a été développé afin d'analyser les communautés bactériennes à différents stades pour comprendre i) comment les communautés bactériennes évoluent

au cours du développement de la tique et ii) détecter de potentielles interactions entre le microbiote de la tique et la présence d'agents pathogènes.

Dans cette application, mon objectif est de déterminer un réseau global où les associations présentes sont des associations valides sur l'ensemble du jeu de données. Il s'agit d'écarter les associations provenant d'un effet stade qui va structurer le jeu de données. Certaines corrélations pourraient en effet être dues à une différence d'abondance des espèces microbiennes entre stades.

Matériel et méthodes. Dans le but de décrire la diversité microbienne présente dans les tiques à différents stades de leur développement, une campagne d'échantillonnage de tiques par la méthode du drapeau a été lancée dans la forêt de Sénart pour récolter les tiques *Ixodes ricinus* responsables de la transmission d'agents pathogènes à l'homme. Les microbiotes de 155 tiques larves, 153 nymphes et 154 adultes ont été analysés par métagénomique ciblée. La composition du microbiote à chacun des stades de développement a d'abord été décrite. J'ai analysé la diversité entre échantillons à l'aide de deux distances : la distance de Bray-Curtis et les distances euclidiennes calculées à partir des données normalisées par MAGMA. J'ai représenté celles-ci à l'aide d'un positionnement multidimensionnel non-métrique (NMDS). J'ai testé l'effet stade sur la diversité bêta par NP-MANOVA (adonis R, Oksanen et al., 2013). J'ai inféré le réseau global avec MAGMA en prenant en compte l'effet du stade de développement de la tique.

Résultats. Une étude sur la diversité bactérienne dans chaque stade a permis de décrire dans un premier temps les espèces présentes, leurs nombres et les différences entre stades (Figure 4.1). Le microbiote de la larve, acquis de la mère, est différent de celui de la nymphe ou de l'adulte qui s'est nourri sur un vertébré pour se développer. Le microbiote sanguin du vertébré hôte est en partie transféré lors de ce repas. Les résultats de l'analyse de l'effet stade donne ce dernier statistiquement significatif avec une MANOVA non paramétrique sur les deux distances utilisées. Les microbiotes des nymphes et des adultes sont proches pour la distance de Brays-Curtis mais se distinguent pour la distance euclidienne calculée sur la normalisation de MAGMA. Les échantillons se répartissent sur le premier axe MDS1 selon leur stade de développement

(Figure 4.1 B.). La distance euclidienne sur les données normalisées par MAGMA permet de mieux distinguer les stades des échantillons de tiques. L'utilisation de la normalisation de MAGMA est donc envisageable en dehors de son utilisation pour les réseaux d'associations.

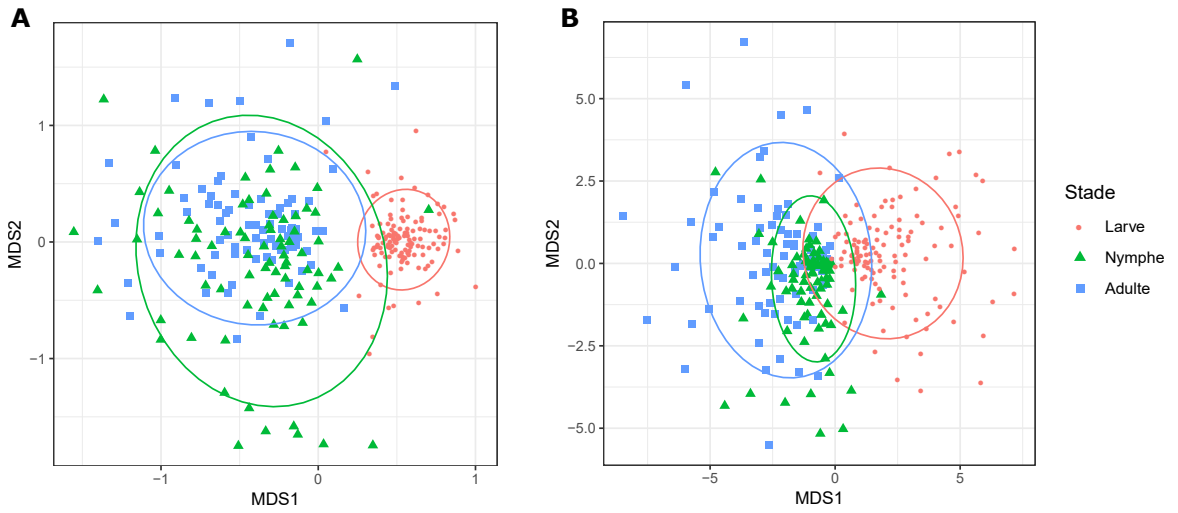


FIGURE 4.1 – Diversité de microbiotes de tiques à différents stades de développement. **A.** Projection des distances de Bray-Curtis inter-échantillons **B.** Projection des distances euclidiennes à partir des données normalisées par MAGMA. Pour A. et B. les données sont réduites à deux dimensions à l'aide d'une analyse NMDS.

Le stade de développement de la tique qui est un facteur structurant important du microbiote semble être bien pris en compte dans le réseau inféré (Figure 4.2). Dans le cas contraire, le réseau aurait pu être fragmenté en sous-groupes d'OTUs.

Le réseau obtenu présente différents résultats d'intérêt dont :

Un module principal comportant de nombreux OTUs liés par de nombreuses associations positives. Cet ensemble de taxons pourrait constituer le coeur du microbiote de la tique.

Un module périphérique liant *Wolbachia* et *Arsenophonus*, deux bactéries typiques du parasitoïde *Ixodiphagus hookeri* qui pond dans *Ixodes ricinus*, ainsi que *Spiroplasma ixodetis*, un symbionte d'arthropode capable de limiter le développement de parasitoïdes dans différents modèles biologiques. Des expériences complémentaires suggèrent que l'association tripartite identifiée illustre une interaction significative.

Ces résultats soulignent la pertinence de l'approche développée et de son implémentation.

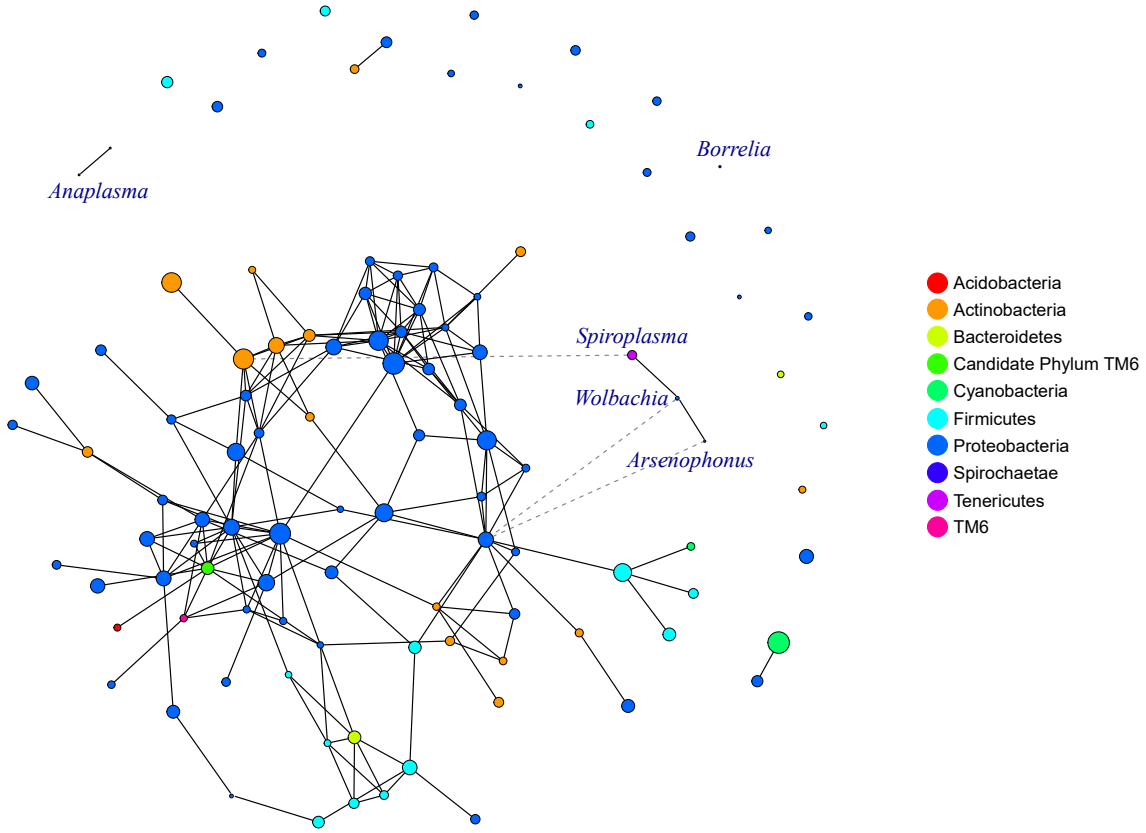


FIGURE 4.2 – Réseau d'associations au sein de microbiotes de tiques échantillonées dans la forêt de Sénart aux trois stades de la tique.
Les liens en traits plein sont des associations positives et les liens en pointillé des associations négatives. Les nœuds représentent les OTUs et la taille des nœuds est proportionnelle à leurs prévalences.

4.2 Application de MAGMA en lien avec les analyses différentielles et adaptation pour l'étude de données multi-gènes : exemple du microbiote de l'environnement de fermes laitières

Un des développements des approches de microbiote tient à l'intégration de données portant sur plusieurs règnes. Il n'est plus rare d'étudier en parallèle les bactéries, les champignons et les protozoaires présents dans un microbiote. Il reste toutefois diffi-

cile d'intégrer l'ensemble des données disponibles car elles sont produites séparément à partir d'un même échantillon. La diversité au sein de chaque phylum étant étudiée par barcoding à l'aide d'un marqueur spécifique, on analyse les données obtenues de manière indépendante pour aboutir à différents tableaux caractérisés par des distributions de tailles de librairies indépendantes.

Nous avons abordé ce problème dans le cadre du projet collaboratif Amont Saint-Nectaire. Une analyse 16S et ITS2 a permis de décrire les compositions bactériennes et fongiques des échantillons de l'environnement des fermes laitières. J'ai travaillé avec Céline Delbès et Isabelle Verdier-Metz de l'Unité mixte de Recherche sur le Fromage afin d'analyser ces deux jeux de données. Le traitement des données multi-gènes constitue un nouveau challenge et a nécessité une adaptation de MAGMA pour analyser ces deux tableaux de données. J'ai donc réalisée cette adaptation méthodologique, mais je n'ai pas pu l'appliquer aux données car les données fongiques n'étaient pas disponibles au moment de l'achèvement de ma thèse.

Dans l'attente de ces données fongiques, j'ai réalisé une application de la méthode MAGMA avec l'objectif de mettre en lien le réseau d'associations obtenu par MAGMA avec les résultats des analyses différentielles suivant deux groupes de fermes caractérisées par leurs historiques de présence d'agents pathogènes. Ce travail a été réalisé en collaboration avec Étienne Rifa, biostatisticien et Sébastien Theil, bioinformaticien de l'UMRF.

L'intérêt de la production de lait cru est un équilibre à trouver entre (i) la présence du microbiote du lait cru et du fromage qui améliorent les qualités gustatives et qui ont de nombreux bienfaits en santé humaine et (ii) la présence d'agents pathogènes et les risques sanitaires associés. La pression sanitaire et l'intensification des pratiques de production favorisent la perte de la diversité microbienne présente dans le lait. La compréhension des transferts de microbes de la ferme au lait est un enjeu majeur pour maintenir cette diversité. Des facteurs biotiques et abiotiques peuvent structurer le microbiote du lait et du fromage et entrer en jeu dans la présence des agents pathogènes.

Le projet Amont Saint-Nectaire est une collaboration avec le Pôle fromager AOP Massif central, plusieurs équipes pluridisciplinaires de l'INRA (UMRF, UMR EPIA, UMRH, UMR Territoires), les chambres d'agriculture du Cantal et du Puy de Dôme, l'Interprofession du Saint-Nectaire et 14 exploitations fermières participantes. Le projet consiste à mettre en place des approches globales en amont prenant en compte : (i) l'ensemble de l'environnement de production primaire, (ii) les agents pathogènes multiples et l'écologie microbienne à l'échelle de la ferme. La partie écologie microbienne de ce projet est un travail unique qui vise à décrire les microbiotes de l'environnement du fromage et ses interactions avec les agents pathogènes. Les résultats du projet doivent servir à améliorer la filière et la maîtrise sanitaire des exploitations.

Matériel et méthodes. Une campagne de prélèvements a été réalisée en 2017 auprès des 14 fermes participantes, dont 7 fermes connaissant des contaminations récurrentes en agents pathogènes qui forment le groupe A et 7 autres généralement exemptes d'agents pathogènes qui forment le groupe B. Trois visites ont été effectuées en été et trois en hiver. Sept environnements ont été analysés : fèces, litière, air ambiant, surface des trayons, eau de la machine à traire, filtre et lait du tank. Au total, 546 échantillons ont été analysés par métagénomique. Le gène 16S a été ciblé pour caractériser la communauté bactérienne et la région ITS2 pour la communauté fongique. Après une analyse sur la diversité des microbiotes, nous avons pu sélectionner les facteurs structurant ces communautés bactériennes. J'ai ensuite adapté MAGMA pour qu'il puisse inférer les associations inter et intra communautés, fongiques et bactériennes (voir encadré ci-dessous).

Résultats. Dans l'attente des données fongiques, la distribution des OTUs bactériens a été étudiée en fonction des deux classes de fermes identifiées a priori. L'environnement des surfaces de trayons a montré une différence significative de diversité entre les deux groupes de fermes pour les diversités de Brays-Curtis et Unifrac pondéré. Les taxa bactériens associés aux classes de fermes ont été identifiés à l'aide de quatre méthodes d'analyse discriminante (DESeq2 : Love et al., 2014, PLS-DA : Barker et Rayens, 2003, metacoder : Z. S. L. Foster et al., 2017 et metagenomeSeq : Paulson et al., 2013). Le réseau d'associations microbiennes du microbiote des surfaces de trayons en hiver a été inféré avec MAGMA sur l'ensemble des 42 échantillons des 14 fermes, sans

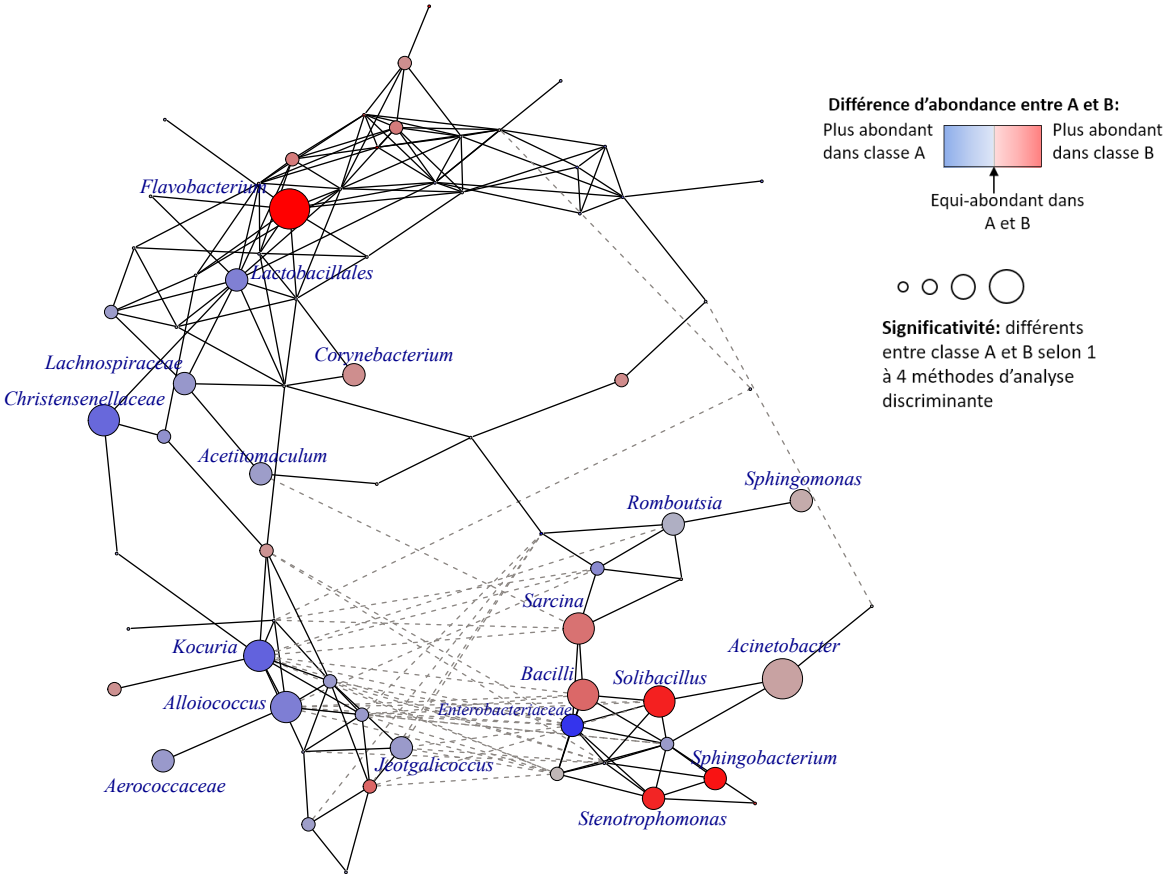


FIGURE 4.3 – Réseau d'associations entre genres bactériens d'échantillons prélevés sur les surfaces de trayons en hiver dans le cadre du projet Amont Saint-Nectaire ; représentation des genres différentiellement abondants dans les deux groupes de fermes. Seuls les OTUs avec une prévalence supérieure à 25% sont analysés ici. Les nœuds représentent les genres bactériens et les liens les associations détectées. Les liens en trait plein sont des associations positives et les liens en pointillé des associations négatives. La couleur du nœud dépend du log ratio des moyennes. La taille des nœuds est proportionnelle au nombre de fois où le genre bactérien a été retrouvé différentiellement abondant entre les classes de fermes pour quatre méthodes d'analyse discriminante (DESeq2, PLS-DA, metacoder, metagenomeSeq).

a priori sur les groupes de fermes. La place des taxa discriminants dans le réseaux d'associations a été étudiée (Figure 4.3). Les résultats obtenus montrent clairement un lien entre les résultats de l'analyse différentielle et les résultats du réseau d'associations. Il est possible d'identifier deux modules d'OTUs spécifiques aux groupes de fermes : un module d'OTUs représenté par *Kocuria* et *Alloiococcus* plus abondant dans le groupe A, et un module avec *Bacilli* et *Solibacillus* plus abondant dans le groupe B. Un grand nombre d'associations négatives sépare ces deux modules. Ces résultats pourrait révéler l'existence de sous-communautés microbiennes spécifiques aux fermes avec des contaminations récurrentes en agents pathogènes.

Dans l'attente de développements futurs, j'ai adapté MAGMA pour qu'il puisse inférer les associations inter et intra communautés, fongiques et bactériennes (voir encadré ci-dessous).

MAGMA : ADAPTATION POUR PLUSIEURS TABLEAUX

Soit Y les données observées tel que $Y = \left(Y^{(16S)} \mid Y^{(ITS)}\right)$ avec $Y^{(16S)}$ la matrice de données de communautés bactériennes et $Y^{(ITS2)}$ la matrice de données de communautés fongiques.

Les deux matrices $Y^{(16S)}$ et $Y^{(ITS2)}$ sont normalisées séparément en utilisant la transformation de l'outil MAGMA afin de prendre en compte les spécificités de chaque tableau (surdispersion, compositionalité, excès de zéros, éventuellement covariable). On obtient ainsi deux matrices de données $\hat{Z}^{(16S)}$ et $\hat{Z}^{(ITS)}$. La matrice concaténée $\hat{Z} = \left(\hat{Z}^{(16S)} \mid \hat{Z}^{(ITS)}\right)$ est une estimation de $Z \sim N(0, \Theta^{-1})$, la matrice latente aux données observées Y.

Le réseau d'association est ensuite inféré à partir de la matrice $\hat{Z}$ qui intègre les données bactériennes et fongiques.

Il sera important de comparer la sensibilité et la spécificité de détection d'associations au sein d'un tableau de données comparativement à la sensibilité et la spécificité de détection d'associations entre différents tableaux de données métagénomiques.

4.3 Adaptation de MAGMA pour variables supplémentaires de présence/absence : exemple du microbiote de l'abeille

Certains microorganismes peuvent être difficiles à étudier à l'aide de données métagénomiques de type barcoding. C'est notamment le cas des taxons bactériens rares, à l'image de certains pathogènes : si la couverture n'est pas assez importante, ils peuvent

ne pas être détectables par les analyses métagénomiques. De plus, les approches de barcoding sont difficiles à transposer aux virus, dont les génomes sont trop hétérogènes pour dégager un marqueur séquençable pour l'ensemble des espèces. La présence de ces taxons peut être validée par des approches de biologie moléculaire complémentaires au barcoding métagénomique. Dans ces conditions, il est intéressant d'ajouter la donnée de présence/absence de ces taxons au réseau d'association construit à partir des données de barcoding.

Un contact initié avec le Laboratoire Microorganismes : Génome et Environnement (LMGE) qui s'intéresse au microbiote de l'abeille nous a permis de travailler en collaboration avec Iris Eouzan et David Biron sur l'analyse statistique de leur jeu de données métagénomiques. Nous avons valorisé cette collaboration par un article en cours de soumission (annexe A). Ils ont acquis des données complémentaires sur les virus et parasites de l'abeille dont les présences ont été documentées par détection spécifique. Ces données sont des données de pseudo-abondance. Pour les inscrire dans le réseau, il a fallu adapter MAGMA pour que l'outil intègre ce type de données supplémentaires.

Ce travail s'inscrit dans le projet BeeHope (BioDIVERSA (H2020 EraNET)) qui vise à promouvoir l'abeille locale. Ce projet dirigé par le CNRS implique un dispositif citoyen et cinq partenaires scientifiques européens. Depuis une vingtaine d'années, les colonies d'abeilles s'effondrent massivement partout dans le monde. En Europe, l'abeille *Apis mellifera* peut subir dans certains pays des taux de mortalité supérieurs à 20%, contre 5% à 10% de mortalité naturelle hivernale selon le programme d'observation Epilobee. Plusieurs facteurs sont en cause : changement des environnements, perte de la biodiversité, diversité génétique de l'abeille, pesticides, virus, parasites tels que la loque, le varroa, des champignons comme *Nosema cerranae*. L'objectif du projet est notamment de comprendre la structuration du microbiote intestinal de la lignée évolutive M de l'abeille mellifère en fonction de facteurs biogéographiques, génétiques et temporels des colonies d'abeilles. Chacun de ces facteurs peut potentiellement influencer le microbiote de l'abeille et évolue avec elle. Les interactions entre microorganismes et agents pathogènes sont ici le sujet de notre application.

Matériel et méthodes. Des prélèvements d'abeilles sur sept conservatoires répartis au Portugal, en Espagne et en France ont été effectués sur trois mois : juillet, août et septembre. Pour chaque conservatoire, il a été prélevé sur six ruches, un échantillon de neufs intestins d'abeille par ruches. Un total de 192 échantillons a été analysé par métagénomique ciblé de l'ARNr 16s. L'article en annexe A présente en détail l'obtention de ces données et le contexte de l'étude. Deux variables climat et paysage ont été construites pour caractériser les observatoires.

L'humidité et la température en trois points de la ruche ont été analysées en continu. Sept virus (ABPV, BQCV, CBPV, DWV, IAPV, KBV et SBV) ont été analysés par RT-PCR ainsi que des microsporidies (*Nosema ceranae* et *Nosema apis*) et bactéries pathogènes (loques européenne et américaine) par PCR. Chaque PCR et RT-PCR détecte la présence/absence spécifique d'un parasite et s'effectue sur 3 réplicats biologiques. Ces 3 réplicats biologiques permettent d'estimer le taux de présence du parasite dans la ruche. Les variables utilisées pour les parasites dans les analyses statistiques sont les taux de présence identifiés sur les 3 réplicats ; e.g. 1/3 si le parasite a été trouvé dans un seul des 3 réplicats.

Nous avons mesuré les effets des facteurs biogéographiques sur la diversité bêta à l'aide de modèles NP-MANOVA. Deux types de diversité bêta ont été mesurés (Bray-Curtis et Unifrac pondéré). J'ai inféré le réseau d'associations avec MAGMA en ajoutant comme variable du réseau les données semi-quantitatives de présence de parasites et comme covariables les données de différents facteurs environnementaux : hygrométrie et température du centre de la ruche, conservatoire. J'ai dû adapter MAGMA pour qu'il puisse traiter ces données supplémentaires (voir encadré ci-dessous). Une transformation des données binaires, ordinales ou continues a été développée pour pouvoir ajouter ces données dans le réseau. Le modèle de copule graphique gaussien (Dobra et Lenkoski, 2011) sur lequel se base MAGMA s'accommode avec les données de différents types en modélisant marginalement les lois dans la copule : (i) les données 16S sont modélisées paramétriquement avec une loi ZIBN afin de modéliser le bruit associé aux données biologiques (taille de librairie des échantillons variable, présence de facteurs), (ii) les données supplémentaires sont modélisées non-paramétriquement par estimation empirique de la fonction de répartition.

MAGMA : DONNÉES SUPPLÉMENTAIRES BINAIRES OU ORDINALES

Soit $Y = \left(Y^{(16S)} \mid Y^{(sup)}\right)$ les données observées comprenant les données 16S et les données supplémentaires. $Y^{(sup)}$ est de dimension n le nombre d'échantillons et k le nombre de variables supplémentaires.
Nous cherchons à estimer $Z = \left(Z^{(16S)} \mid Z^{(sup)}\right)$ la matrice latente normale multivariée. Comme une estimation de $Z^{(16S)}$ est déjà donnée, il reste à estimer la matrice $Z^{(sup)}$ correspondant aux données supplémentaires.
Les fonctions marginales de répartition des données supplémentaires sont estimées empiriquement :

$$\hat{F}_j(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\left\{Y_{ij}^{(sup)} \leq y\right\},\ j = 1, \ldots, k$$

où $\mathbb{1}$ est la fonction indicatrice d'un ensemble.

Les données sont ensuite transformées en utilisant la médiane des valeurs possibles pour $Z^{(sup)}$ (se rapporter à la section inférence de l'article MAGMA).

$$\hat{Z}_{ij}^{(sup)} = \Phi^{-1}\left(\frac{\hat{F}_j\left(Y_{ij}^{(sup)} - 1\right) + \hat{F}_j\left(Y_{ij}^{(sup)}\right)}{2}\right),\ i = 1, \ldots, n,\ j = 1, \ldots, k$$

où Φ est la fonction de répartition d'une loi normale centrée réduite.

Les données 16S et les données supplémentaires ont été normalisées et sont concaténées dans une même table, $\hat{Z} = \left(\hat{Z}^{(16S)} \mid \hat{Z}^{(sup)}\right)$. Le réseau peut ensuite être inféré à partir de cette matrice $\hat{Z}$.

Résultats. Le conservatoire a un effet significatif sur la diversité bêta (17% pour la distance de Bray-Curtis, 27% pour UniFrac). Cet effet s'explique par la différence de climat et le type de paysage des conservatoires. Le mois a un effet faiblement explicatif (2%), ainsi que la diversité génétique (12% pour 10 modalités). Les paramètres

explicatifs de la diversité des échantillons ont été pris en compte dans MAGMA en tant que covariables : (i) une variable qualitative, le conservatoire qui était l'effet le plus important et qui prend en compte tous les aspects biogéographiques des données, (ii) deux variables continues, l'hygrométrie et la température qui sont connues pour structurer les communautés microbiennes.

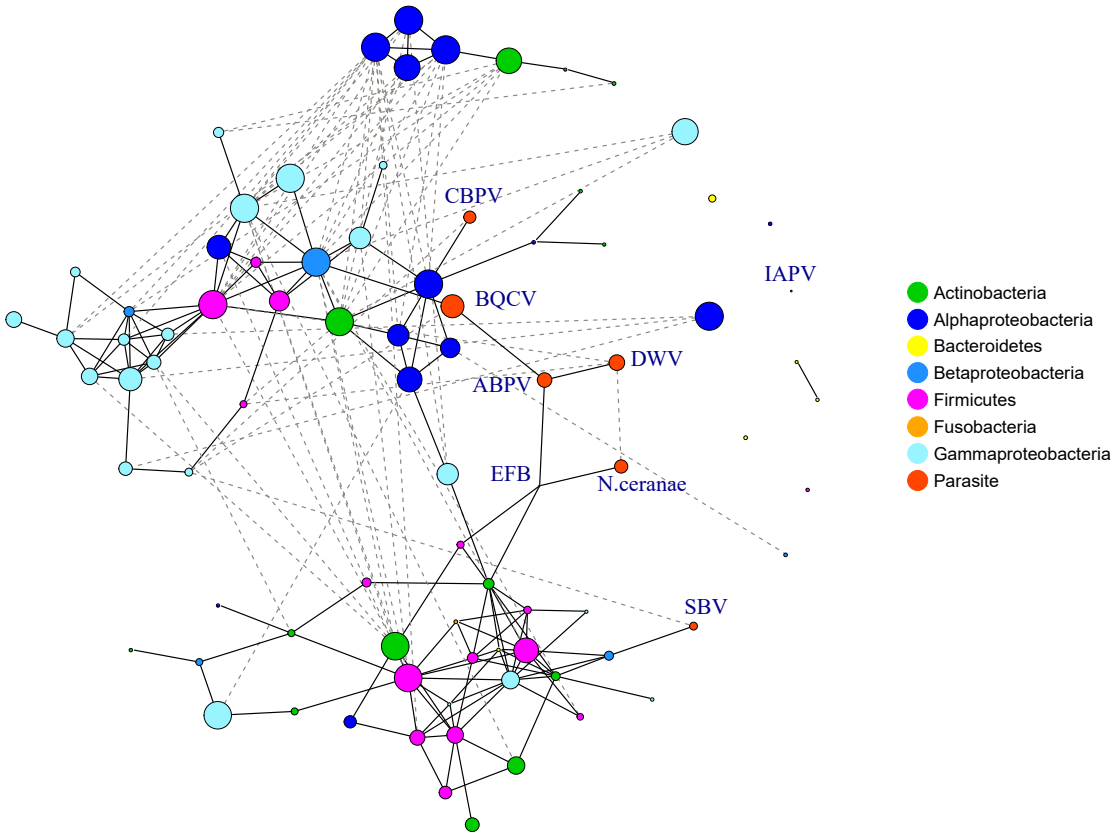


FIGURE 4.4 – Réseau d'associations au sein de microbiotes d'ouvrières de l'abeille mellifère échantillonnées dans le cadre du projet BeeHope (BioDIVERSA (H2020 Era-NET)).
Les parasites détectés par biomarqueurs spécifiques (amorces PCR) sont également ajoutés dans l'interactome. Les liens en trait plein sont des associations positives et les liens en pointillé des associations négatives. Les nœuds représentent les genres bactériens et les parasites, et la taille des nœuds est proportionnelle à leurs prévalences.

Le développement proposé permet d'obtenir, après avoir pris en compte l'effet des facteurs structurants, un réseau d'association incluant les données de présence/absence des virus et parasites (Figure 4.4). De fait, certains de ces éléments sont liés à des bactéries du réseau. Il conviendrait maintenant de tester la spécificité et la sensibilité de la détection d'associations impliquant les microorganismes caractérisés par des données

de présence/absence. Compte tenu des résultats présentés dans le chapitre 2 quant à l'importance des données de présence/absence sur la détection d'associations, il est probable que la méthode présente des propriétés intéressantes.

Chapitre 5

Discussion et perspectives

Sommaire

Mon travail de thèse a consisté à développer des méthodes d'analyse de réseaux d'associations entre OTUs. Ces réseaux offrent un angle de représentation et de compréhension de la structure globale des microbiotes qui est encore méconnue. Au-delà d'un portrait-robot des OTUs identifiables dans un microbiote, les réseaux proposent notamment une visualisation des associations statistiques entre OTUs observées sur le jeu de données. Cela permet en particulier de décrire les patrons de structure globale et les relations plus fines au sein des microbiotes échantillonnés. Dans les paragraphes suivants, je développe différentes idées pour affiner l'inférence d'associations et identifier celles qui peuvent représenter de réelles interactions biologiques.Enfin, l'observation de la distribution des relations observées entre OTUs peut fournir des informations sur les propriétés du microbiote étudié, comme l'identification d'OTUs clés de voûte (Berry et Widder, 2014 ; Banerjee et al., 2018). Je conclus la discussion sur la problématique de l'interprétation des réseaux d'associations.

5.1 Amélioration et généralisation de l'outil MAGMA

Un des résultats majeurs de ma thèse est le développement d'un outil d'étude de ces réseaux d'associations : MAGMA. Cet outil permet d'inférer un réseau d'associations microbiennes en prenant en compte la structure bruitée des données issues de la métagénomique. La méthode est basée sur le modèle graphique gaussien de copules dans lesquelles nous modélisons les distributions marginales avec des modèles linéaires généralisés (GLM) suivant une loi zero-inflated binomiale négative (ZIBN).

Son application à la détection d'interactions liées aux agents pathogènes est une question méthodologique complexe. La métagénomique est peu sensible pour détecter les agents pathogènes bactériens qui peuvent être en faible prévalence et/ou abondance. Nous avons vu dans le chapitre 2 que lors de l'étude des microorganismes de faibles prévalences, il est souvent difficile d'inférer des associations statistiques à tel point que les données quantitatives ne sont pas plus informatives que les données qualitatives.

Pour tester la présence ou l'absence d'agents pathogènes, des détections spécifiques sont effectuées en parallèle des analyses métagénomiques. En effet, la métagénomique ne permet pas toujours de déceler les agents pathogènes. De plus, la métagénomique ciblée reste spécifique aux bactéries ou éventuellement aux champignons alors que les agents pathogènes peuvent être aussi des virus ou des parasites. J'ai proposé une extension du modèle MAGMA dans la partie application afin d'intégrer des données de présence/absence comme celles issues des détections spécifiques. Cette méthode est adaptée pour intégrer les agents pathogènes non-bactériens à un réseau de données 16S. Comme indiqué dans le chapitre 4, la sensibilité et la spécificité de l'outil pour étudier les associations entre données métagénomiques et données de présence/absence doivent encore être étudiées en détail.

Pour aller plus loin, un problème méthodologique survient lorsque les données supplémentaires sont des données d'agents pathogènes bactériens. Dans ce cas, il est possible d'avoir la donnée de présence/absence d'une bactérie pathogène obtenue par la détection spécifique et une évaluation de son abondance issue de la métagénomique. Cela risque de poser problème lors du calcul des corrélations partielles : la donnée de

l'une des variables (qualitative ou quantitative) peut expliquer l'autre sans que cela ne soit maîtrisé. La solution la plus simple pour éviter ce problème pourrait-être de supprimer les données 16S des OTUs redondants afin de garder les données de détection spécifique qui sont plus précises. Une autre solution plus conservative serait d'utiliser les données de détections spécifiques pour estimer la probabilité de zéros structurels de la partie « zero-inflated » de la distribution modélisant l'agent pathogène pour lequel l'abondance serait estimée à l'aide des données métagénomiques. Ainsi, on pourrait combiner une variable de présence/absence précise mais non quantitative et une variable moins précise mais quantitative.

L'outil MAGMA, au-delà de la métagénomique pour laquelle il a été développé, peut être généralisé à d'autres contextes ou d'autres domaines. Les données d'abondances issues de la metagénomique globale « shotgun » ont des caractéristiques similaires d'un point de vue statistique, la généralisation de MAGMA est donc naturelle. Le modèle pourrait être utilisé pour l'identification de gènes associés dans les génomes d'un microbiote pour reconstituer des espèces métagénomiques (Nielsen et al., 2014), même si la quantité importante de données est à prendre en compte dans les possibilités d'application. La transcriptomique ou encore la métabolomique ont également des challenges méthodologiques communs (Gallopin et al., 2013 ; Krumsiek et al., 2011). MAGMA peut aussi être étendu à d'autres questions biologiques via l'utilisation de distributions statistiques adaptées. Par exemple, MAGMA a été utilisé sur des données de sérologie visant à caractériser le ou les sérovars de leptospirose présents dans des échantillons infectés. L'objectif était d'identifier les associations entre les données de détections de variants sérologiques pour mettre en évidence des co-infections ou des patrons de réactions croisées récurrents. Ces analyses permettent d'affiner l'interprétation des résultats et à plus long terme d'orienter les pratiques et les développements techniques de caractérisation des sérovars. Dans cette optique, il est déjà possible de ne pas prendre en compte la taille de librairie pour les données non compositionnelles. L'utilisation de lois plus adaptées à ce type de données, comme la loi « zero-inflated » Gamma serait nécessaire.

La normalisation de MAGMA pourrait également être utilisée dans d'autres contextes

que l'inférence de réseau d'associations. Cette transformation est prévue pour *gaussianiser* les données en se basant sur une estimation paramétrique de la fonction de distribution des abondances, sur la méthode de la transformée inverse, sur une imputation par la médiane et sur les copules gaussiennes. Cette normalisation pourrait servir pour l'analyse différentielle. Elle pourrait également être utilisée pour du clustering en calculant la distance euclidienne entre échantillons des données normalisées comme abordé dans la partie 4.1. Les distances inter-échantillons obtenues peuvent également être projetées à l'aide d'une analyse en coordonnées principales (PCoA, aussi appelée MDS : multidimensional scaling). Plus largement, cette normalisation peut servir en préalable aux nombreuses analyses statistiques dont la normalité des données est une hypothèse.

La modélisation de la distribution de l'abondance des OTUs est une étape importante de l'inférence du réseau d'associations par MAGMA. Je propose ici plusieurs pistes d'améliorations de cette modélisation faite à l'aide de GLMs utilisant une loi ZINB :

— Les facteurs structurants les échantillons sont actuellement pris en compte par l'ajout de covariables en effet simple. Des réplicats des échantillons sont souvent introduits dans l'analyse métagénomique pour pallier au bruit des données. La prise en charge de ces réplicas n'est souvent pas abordée par les outils existants. Elle peut se faire par l'ajout d'effets aléatoires dans les GLMs et cela reste à développer.

— Une comparaison des différents estimateurs de taille de librairie devrait être effectuée par une étude complète de simulations en évaluant la qualité d'inférence en fonction de l'estimateur. Le choix de cet estimateur a un effet important sur les résultats obtenus. L'inclusion d'autres estimateurs de taille de librairie comme ceux proposés par edgeR ou DESeq2 en plus de ceux déjà implémentés dans MAGMA : le ratio GMPR, le logarithme de la moyenne géométrique plus un pseudocount (proche de la transformation clr) et la somme totale permettrait de valider nos choix méthodologiques.

— L'effet des covariables est uniquement pris en compte sur la partie binomiale négative de la distribution ZIBN. L'effet des covariables sur la probabilité de

zéros structurels de la partie « zero-inflated » est difficile à évaluer. Les difficultés techniques proviennent probablement de la multiplication des paramètres à estimer. Des développements seraient nécessaires à ce niveau. En effet, un facteur de risque jouera plus probablement sur la présence ou l'absence d'un agent pathogène que sur son abondance.

— Actuellement, lorsqu'on choisit d'intégrer des covariables à la méthode MAGMA, les effets de celles-ci sont pris en compte pour l'ensemble des OTUs. Or ces covariables peuvent ne pas être pertinentes pour tous les OTUs. Il me parait intéressant de développer une méthode de sélection de ces covariables, par exemple avec des tests partiels. La possibilité d'ajouter les covariables en tant que variables supplémentaires dans MAGMA devrait également être évaluée.

— La prise en charge de l'effet des facteurs sur les liens, autrement dit sur les valeurs de la matrice de précisions, pourrait être envisagée durant l'inférence de celle-ci. La multiplication des paramètres du modèle que cela impliquerait doit être justifiée au regard de l'apport biologique de ce développement.

— Les covariables ne sont actuellement pas prises en compte pour les données supplémentaires comme par exemple pour les données de détections spécifiques. Des développements devraient être effectués afin de les intégrer dans l'estimation de la fonction de répartition de ces variables supplémentaires.

Afin d'améliorer l'efficacité et la qualité de l'inférence de réseaux de MAGMA, plusieurs possibilités se présentent :

— Pour l'inférence de la matrice de précision, l'utilisation d'algorithmes plus efficaces comme celui proposé dans le package R FastGGM (T. Wang et al., 2016) permettrait de gagner du temps de calcul.

— Le filtre par paires proposé dans l'article du chapitre 2 devrait être implémenté dans MAGMA et testé pour améliorer la qualité de l'inférence du réseau.

— Il est possible d'ajouter de l'information a priori phylogénétique ou sur les interactions connues dans le cadre de ce type d'analyse (Lo et Marculescu, 2017 ; Lo et Marculescu, 2018). Cette possibilité apportera sans doute une plus-value à MAGMA avec l'apparition de bases de données sur les interactions microbe-microbe.

— D'une manière plus générale, MAGMA pourrait être affiné en l'intégrant dans un modèle hiérarchique bayésien pour inclure davantage de facteurs à différents niveaux. Encore une fois, la complexité engendrée par un tel modèle et le taux de fausse découverte engendré par une surparamétrisation devrait être discuté au regard de l'apport biologique.

5.2 Utilisation des corrélations comme proxy des interactions

Les associations statistiques sont mesurées à partir de coefficients de corrélation, de coefficients de similarité/dissimilarité, de mesures de distance ou de coefficients découlant de la théorie de l'information. Ces associations expriment une observation faite sur un ensemble restreint d'échantillons. Ces réseaux sont parfois appelés « réseaux d'interactions », or à partir de l'analyse des associations statistiques, des hypothèses peuvent être posées mais il est impossible de conclure directement à une réelle interaction biologique. La détermination d'une interaction biologique doit être validée par une expérimentation *ad hoc* (Faust et Raes, 2012 ; Carr et al., 2019).

Bien qu'ils doivent être utilisés avec un regard critique, ces outils sont des générateurs d'hypothèses, indispensables pour accélérer la recherche sur les microbiotes en cours de développement. Les réseaux d'associations reflètent une observation de l'ensemble des forces qui dirigent l'assemblage des microbiotes. Ces forces sont multiples et complexes et les connaissances et techniques actuelles en microbiologie ne permettent pas encore une modélisation détaillée de celles-ci. De plus, il est difficile méthodologiquement de séparer les effets des causes.

Différentes sources de données peuvent toutefois faciliter l'identification d'interactions pertinentes. A. B. Hill (1965) liste dans le cadre d'études épidémiologiques différents critères qui peuvent être pris en compte pour évaluer la plausibilité d'une causalité. Ces critères peuvent être transposés et adaptés aux études d'interactions microbes-microbes dans les métagénomes et nous nous en servirons pour discuter notre propos.

Pertinence des dispositifs d'observation et présence de données expérimentales

L'observation de microbiotes « naturels » est indispensable pour étudier leur diversité et comprendre les liens entre le microbiote et son hôte, environnement ou organisme. Les connaissances acquises doivent nous permettre d'optimiser les études d'associations microbiennes dans les microbiotes.

Des choix adaptés dans le design de l'expérimentation permettraient de limiter les erreurs durant l'inférence du réseau d'associations. Le plan d'expérimentation doit être prévu pour maîtriser au mieux les différents biais expérimentaux en uniformisant les procédures. Lors de la construction du plan d'échantillonnage, il apparaît indispensable de maîtriser les facteurs environnementaux responsables de variations importantes du microbiote. Le but est de travailler avec des données les plus homogènes possible afin de limiter les facteurs confondants. Si l'objectif de l'étude est également d'étudier l'effet d'un facteur sur la diversité d'un microbiote, il est indispensable de prendre en compte cet effet dans l'inférence de réseau avec un modèle avec covariables ou en effectuant des analyses en sous-groupe. Le développement de tests pour mieux détecter les facteurs confondants et l'amélioration de l'intégration des covariables dans les modèles sont des perspectives de recherche pour améliorer la qualité de l'inférence de réseaux.

Preuve expérimentale

Il est devenu possible de construire des microbiotes synthétiques où les espèces initiales sont maîtrisées (Grosskopf et Soyer, 2014 ; Venturelli et al., 2018 ; Vázquez-Castellanos et al., 2019). Le développement actuel de ce type de microbiotes « jouets » permet d'en diminuer la complexité et de maîtriser les contingences historiques, par exemple la composition des inoculum ou l'exposition potentielles à d'autres microbes. L'arrivée des microbiotes synthétiques devrait permettre de vérifier de nombreuses hypothèses en écologie microbienne et de valider les modèles théoriques.

Plus couramment, des études entre paires de taxons peuvent être entreprises. Les connaissances acquises sur les interactions doivent nous permettre de nourrir d'autres critères envisagés par A. B. Hill : la plausibilité biologique, la cohérence biologique et l'analogie avec d'autres associations.

Plausibilité biologique, Cohérence biologique & Analogie

Dans les nombreuses applications d'inférence de réseaux que j'ai pu effectuer, les OTUs reliés sont proches phylogénétiquement. De fait, on sait que les OTUs phylogénétiquement proches ont également des fonctions similaires. Cette observation reflète la redondance fonctionnelle des microbiotes. L'utilisation de base de données a priori peut servir à valider et à déterminer le sens du lien de causalité de ce qui est fait dans le domaine des interactions proteine-proteine (Szklarczyk et Jensen, 2015 ; Keskin et al., 2016).

La cause précède l'effet

Il est possible d'inférer des relations de causalité en utilisant la temporalité des échantillons (Wunsch et al., 2010), donnée dans notre cas par des données longitudinales. Des relations de causalité de Granger ou des dynamiques de population de type Lotka-Volterra peuvent alors être inférées (Faust, Bauchinger et al., 2018 ; Mainali et al., 2019 ; Dohlman et Shen, 2019). Des probabilités conditionnelles peuvent également être inférées à l'aide des réseaux bayésiens (Lugo-Martinez et al., 2019). Les réseaux ainsi obtenus sont des réseaux avec des liens dirigés qui se rapprochent le plus des interactions biologiques. Cependant, la complexité des modèles augmente rapidement avec le nombre d'OTUs et le taux de fausses découvertes est encore plus important que pour l'inférence d'associations symétriques classiques. De plus, les modèles biologiques d'études ne permettent pas toujours un échantillonnage longitudinal (par exemple, la tique est broyée avant d'extraire son ADN et cette opération ne supporte pas la répétition). En plus de n'être pas toujours réalisables techniquement, les études sur des données longitudinales nécessitent d'avantage d'échantillons et des efforts financiers importants. Les réseaux d'associations sur les études transversales sont donc incontournables dans ces situations.

5.3 Interprétation du réseau

Différents modèles explicites d'interactions entre les microbes peuvent aboutir à la même forme d'associations. Les associations statistiques mises en évidence sont parfois contre-intuitives. Par exemple, le travail effectué sur le microbiote d'*Ixodes ricinus* met

en évidence une association positive entre *Spiroplasma* - *Wolbachia* – *Arsenophonus*, alors que les données biologiques suggèrent que les interactions entre *Spiroplasma* et les deux autres bactéries devraient être négatives. Une étude plus systématique des associations attendues pour différentes interactions biologiques serait bénéfique.

Cet exemple illustre également la complexité à interpréter des associations impliquant plus de deux OTUs. Nous avons vu à plusieurs reprises durant cette thèse que les réseaux d'associations pouvaient être structurés en sous-communautés d'OTUs ou modules. Il est possible (i) qu'un facteur structurant les données n'ait pas été pris en compte, ou (ii) que les interactions entre microbes puissent intrinsèquement mener à l'existence de différentes communautés d'équilibres (Gonze, Lahti et al., 2017). À partir des données de séquençage, les modèles à classes latentes sont une option pour identifier des facteurs structurants cachés (Chiquet, Smith et al., 2009). A partir du réseau, des outils sont nécessaires pour identifier des structures, d'autant plus que la majorité des algorithmes de clustering ne prennent pas en compte des natures de lien différentes. Au-delà, des travaux sont encore nécessaires pour améliorer notre capacité à distinguer les hypothèses sous-jacentes.

Un des objectifs de l'écologie des communautés est d'anticiper les changements au sein des écosystèmes, d'origine anthropique ou non. La perte de biodiversité à toutes les échelles de la vie, les perturbations des écosystèmes ou encore le déclin des services écosystémiques sont des questions universelles. A l'échelle des microbiotes, qui rendent de nombreux services aux écosystèmes dont l'homme, la détection de changement au sein de ces communautés est donc un indicateur du bouleversement des écosystèmes. De nombreux chercheurs tentent de contrôler ces changements à l'aide du biomonitoring (Bohan et al., 2017 ; Derocles et al., 2018). Dans ce contexte, la recherche d'indicateurs issus de la structure du réseau est également une question en suspens (Karimi et al., 2017). L'étude des changements observés au sein des réseaux d'associations, impliquant ou non des OTUs clés pour leur stabilité, pourrait servir à définir les états et les évolutions des écosystèmes microbiens.

Bibliographie

Abegaz, F. et E. C. Wit (2015). Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica* 69 (4), 419-441. doi : 10.1111/stan.12066.

Achtman, M. et M. Wagner (2008). Microbial diversity and the genetic nature of microbial species. *Nature Reviews Microbiology* 6 (6), 431-440. doi : 10.1038/nrmicro1872.

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)* 44 (2), 139-177. doi : 10.1007/978-94-009-4109-0.

Aitchison, J. (1986). The statistical analysis of compositional data : monographs in statistics and applied probability. *Chapman & Hall, London*. doi : 10.1007/978-94-009-4109-0.

Anders, S. et W. Huber (2010). Differential expression analysis for sequence count data. *Genome Biology* 11 (10), R106. doi : 10.1186/gb-2010-11-10-r106.

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26 (1), 32-46. doi : 10.1111/j.1442-9993.2001.01070.pp.x.

Anderson, M. J., P. de Valpine, A. Punnett et A. E. Miller (2019). A pathway for multivariate analysis of ecological communities using copulas. *Ecology and Evolution* 9 (6), 3276-3294. doi : 10.1002/ece3.4948.

Anderson, M. J. et D. C. I. Walsh (2013). PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions : What null hypothesis are you testing ? *Ecological Monographs* 83 (4), 557-574. doi : 10.1890/12-2010.1.

Bairey, E., E. D. Kelsic et R. Kishony (2016). High-order species interactions shape ecosystem iversity. *Nature Publishing Group* 7, 1-7. doi : 10.1038/ncomms12285.

Banerjee, S., K. Schlaeppi et M. G. A. V. D. Heijden (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews Microbiology* in press. doi : 10.1038/s41579-018-0024-1.

Barker, M. et W. Rayens (2003). Partial least squares for discrimination. *Journal of Chemometrics* 17 (3), 166-173. doi : 10.1002/cem.785.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, page 11.

Berry, D. et S. Widder (2014). Deciphering microbial interactions and detecting keystone species with co-occurrence networks. *Frontiers in Microbiology* 5 (MAY), 1-14. doi : 10.3389/fmicb.2014.00219.

Bishara, A. J. et J. B. Hittner (2012). Testing the significance of a correlation with nonnormal data : Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods* 17 (3), 399-417. doi : 10.1037/a0028087.

Biswas, S., M. McDonald, D. S. Lundberg, J. L. Dangl et V. Jojic (2016). Learning Microbial Interaction Networks from Metagenomic Count Data. *Journal of Computational Biology* 23 (6), 526-535. doi : 10.1089/cmb.2016.0061.

Björk, J. R., F. K. C. Hui, R. B. O'Hara et J. M. Montoya (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology* 27 (12), 2714-2724. doi : 10.1111/mec.14718.

Bohan, D. A., C. Vacher, A. Tamaddoni-Nezhad, A. Raybould, A. J. Dumbrell et G. Woodward (2017). Next-Generation Global Biomonitoring : Large-scale, Automated Reconstruction of Ecological Networks. *Trends in Ecology and Evolution* 32 (7), 477-487. doi : 10.1016/j.tree.2017.03.001.

Bordes, F. et S. Morand (2011). The impact of multiple infections on wild animal hosts : a review. *Infection Ecology & Epidemiology* 1 (0), 1-10. doi : 10.3402/iee.v1i0.7346.

Bray, J. R. et J. T. Curtis (1957). An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs*. doi : 10.2307/1942268.

Breitwieser, F. P., J. Lu et S. L. Salzberg (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics* (September), 1-15. doi : 10.1093/bib/bbx120.

Brooks, J. P. et al. (2015). The truth about metagenomics : Quantifying and counteracting bias in 16S rRNA studies Ecological and evolutionary microbiology. *BMC Microbiology* 15 (1), 1-14. doi : 10.1186/s12866-015-0351-6.

Bukin, Y. S., Y. P. Galachyants, I. V. Morozov, S. V. Bukin, A. S. Zakharenko et T. I. Zemskaya (2019). The effect of 16s rRNA region choice on bacterial community metabarcoding results. *Scientific Data* 6, 1-14. doi : 10.1038/sdata.2019.7.

Burkhardt, C., H. Insam, T. C. Hutchinson et H. H. Reber (1993). Biology and Fertility of S o i l Short communication Impact of heavy metals on the degradative capabilities of soil bacterial communities Key words : Heavy. *Biology and Fertility of Soils*, 154-156.

Caporaso, J. G. et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7 (5), 335-336. doi : 10.1038/nmeth.f.303.

Carr, A., C. Diener, N. S. Baliga et S. M. Gibbons (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal* (June). doi : 10.1038/s41396-019-0459-z.

Chen, E. Z. et H. Li (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics* 32 (17), 2611-2617. doi : 10.1093/bioinformatics/btw308.

Chen, L., J. Reeve, L. Zhang, S. Huang, X. Wang et J. Chen (2018). GMPR : A robust normalization method for zero-inflated count data with application to microbiome sequencing data. *PeerJ* 6, e4600. doi : 10.7717/peerj.4600.

Chiquet, J., M. Mariadassou et S. Robin (2018). Variational inference for sparse network reconstruction from count data, 1-30.

Chiquet, J., A. Smith, G. Grasseau, C. Matias et C. Ambroise (2009). SIMoNe : Statistical Inference for MOdular NEtworks. *Bioinformatics* 25 (3), 417-418. doi : 10.1093/bioinformatics/btn637.

Cho, I. et M. J. Blaser (2012). The human microbiome : At the interface of health and disease. *Nature Reviews Genetics* 13 (4), 260-270. doi : 10.1038/nrg3182.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Austral Ecology* 18 (1), 117-143. doi : 10.1111/j.1442-9993.1993.tb00438.x.

Cole, J. R., Q. Wang, J. A. Fish, B. Chai, D. M. McGarrell, Y. Sun, C. T. Brown, A. Porras-Alfaro, C. R. Kuske et J. M. Tiedje (2014). Ribosomal Database Project : Data and tools for high throughput rRNA analysis. *Nucleic Acids Research* 42 (D1), 633-642. doi : 10.1093/nar/gkt1244.

Coly, S., A.-F. Yao, D. Abrial et M. Charras-Garrido (2016). Distributions to model overdispersed count data. *Journal de la Société Française de Statistique* 154 (2), 39-63.

Costea, P. I., G. Zeller, S. Sunagawa et P. Bork (2014). A fair comparison. *Nature Methods* 11 (4), 359-359. doi : 10.1038/nmeth.2897.

Coyte, K. Z., J. Schluter et K. R. Foster (2015). The ecology of the microbiome : Networks, competition, and stability. *Science* 350 (6261), 663-666. doi : 10.1126/science.aad2602.

Cunningham, R. B. et D. B. Lindenmyer (2005). Modeling Count Data of Rare Species. *Ecology* 86 (5), 1135-1142. doi : 10.1890/04-0589.

Derocles, S. A., D. A. Bohan, A. J. Dumbrell, J. J. Kitson, F. Massol, C. Pauvert, M. Plantegenest, C. Vacher et D. M. Evans (2018). Biomonitoring for the 21st Century : Integrating Next-Generation Sequencing Into Ecological Network Analysis. *Advances in Ecological Research*. Tome 58. Academic Press Inc., p. 1-62. doi : 10.1016/bs.aecr.2017.12.001.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu et G. L. Andersen (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* 72 (7), 5069-5072. doi : 10.1128/AEM.03006-05.

Dobra, A. et A. Lenkoski (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Annals of Applied Statistics* 5 (2 A), 969-993. doi : 10.1214/10-AOAS397.

Dohlman, A. B. et X. Shen (2019). Mapping the microbial interactome : Statistical and experimental approaches for microbiome network inference. *Experimental Biology and Medicine*, 153537021983677. doi : 10.1177/1535370219836771.

Dominguez-Bello, M. G., F. Godoy-Vitorino, R. Knight et M. J. Blaser (2019). Role of the microbiome in human development. *Gut* 68 (6), 1108-1114. doi : 10.1136/gutjnl-2018-317503.

Duff, I. S., A. M. Erisman et J. K. Reid (1986). *Direct methods for sparse matrices*, page 429.

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics* 34 (14), 2371-2375. doi : 10.1093/bioinformatics/bty113.

Fan, J. et J. Lv (2010). A Selective Overview of Variable Selection in High Dimensional Feature Space. *Statistica Sinica* 20 (1), 101-148. doi : 10.1021/nl061786n.Core-Shell.

Fang, H., C. Huang, H. Zhao et M. Deng (2015). CCLasso : Correlation inference for compositional data through Lasso. *Bioinformatics* 31 (19), 3172-3180. doi : 10.1093/bioinformatics/btv349.

Fang, H., C. Huang, H. Zhao et M. Deng (2017). gCoda : Conditional Dependence Network Inference for Compositional Data. *Journal of Computational Biology* 24 (0), cmb.2017.0054. doi : 10.1089/cmb.2017.0054.

Fang, R., B. D. Wagner, J. K. Harris et S. A. Fillon (2016). Zero-inflated negative binomial mixed model : an application to two microbial organisms important in oesophagitis. *Epidemiology and Infection* 144 (11), 2447-2455. doi : 10.1017/S0950268816000662.

Faust, K., F. Bauchinger, B. Laroche, S. de Buyl, L. Lahti, A. D. Washburne, D. Gonze et S. Widder (2018). Signatures of ecological processes in microbial community time series. *Microbiome* 6 (1), 1-13. doi : 10.1186/s40168-018-0496-2.

Faust, K. et J. Raes (2012). Microbial interactions : from networks to models. *Nature Reviews Microbiology* 10 (8), 538-550. doi : 10.1038/nrmicro2832.

Faust, K. et J. Raes (2016). CoNet app : inference of biological association networks using Cytoscape. *F1000Research* 5, 1519. doi : 10.12688/f1000research.9050.1.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes et C. Huttenhower (2012). Microbial co-occurrence relationships in the Human Microbiome. *PLoS Computational Biology* 8 (7). doi : 10.1371/journal.pcbi.1002606.

Feizi, S., D. Marbach, M. Médard et M. Kellis (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature biotechnology* 31 (8), 726-33. doi : 10.1038/nbt.2635.

Foster, K. R., J. Schluter, K. Z. Coyte et S. Rakoff-Nahoum (2017). The evolution of the host microbiome as an ecosystem on a leash. *Nature* 548 (7665), 43-51. doi : 10.1038/nature23292.

Foster, Z. S. L., T. J. Sharpton et N. J. Grünwald (2017). Metacoder : An R package for visualization and manipulation of community taxonomic diversity data. *PLOS Computational Biology* 13 (2). Sous la direction de T. Poisot, e1005404. doi : 10.1371/journal.pcbi.1005404.

Friedman, J., T. Hastie et R. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 (3), 432-441. doi : 10.1093/biostatistics/kxm045.

Friedman, J. et E. J. Alm (2012). Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology* 8 (9). Sous la direction de C. von Mering, e1002687. doi : 10.1371/journal.pcbi.1002687.

Galan, M. et al. (2016). 16S rRNA amplicon sequencing for epidemiological surveys of bacteria in wildlife : the importance of cleaning post-sequencing data before estimating positivity , prevalence and co-infection. *mSystems* 1 (4), e00032-16. doi : http://dx.doi.org/10.1101/039826.

Gallopin, M., A. Rau et F. Jaffrézic (2013). A Hierarchical Poisson Log-Normal Model for Network Inference from RNA Sequencing Data. *PLoS ONE* 8 (10). doi : 10.1371/journal.pone.0077503.

Gevers, D. et al. (2005). Re-evaluating prokaryotic species. *Nature Reviews Microbiology* 3 (9), 733-739. doi : 10.1038/nrmicro1236.

Ghoul, M. et S. Mitri (2016). The Ecology and Evolution of Microbial Competition. *Trends in Microbiology* 24 (10), 833-845. doi : 10.1016/j.tim.2016.06.011.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn et J. J. Egozcue (2017). Microbiome Datasets Are Compositional : And This Is Not Optional. *Frontiers in Microbiology* 8 (November), 1-6. doi : 10.3389/fmicb.2017.02224.

Gloor, G. B., J. R. Wu, V. Pawlowsky-Glahn et J. J. Egozcue (2016). It's all relative : analyzing microbiome data as compositions. *Annals of Epidemiology* 26 (5), 322-329. doi : 10.1016/j.annepidem.2016.03.003.

Gonzales-Barron, U., M. Kerr, J. J. Sheridan et F. Butler (2010). Count data distributions and their zero-modified equivalents as a framework for modelling microbial data with a relatively high occurrence of zero counts. *International Journal of Food Microbiology* 136 (3), 268-277. doi : 10.1016/j.ijfoodmicro.2009.10.016.

Gonze, D., K. Z. Coyte, L. Lahti et K. Faust (2018). Microbial communities as dynamical systems. *Current Opinion in Microbiology* 44, 41-49. doi : 10.1016/j.mib.2018.07.004.

Gonze, D., L. Lahti, J. Raes et K. Faust (2017). Multi-stability and the origin of microbial community types. *The ISME Journal* 11 (10), 2159-2166. doi : 10.1038/ismej.2017.60.

Goslee, S. C. et D. L. Urban (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software* 22 (7), 1-19. doi : 10.18637/jss.v022.i07.

Gotelli, N. J. et R. K. Colwell (2001). Quantifying biodiversity : procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4 (4), 379-391. doi : 10.1046/j.1461-0248.2001.00230.x.

Grosskopf, T. et O. S. Soyer (2014). Synthetic microbial communities. *Current Opinion in Microbiology* 18 (1), 72-77. doi : 10.1016/j.mib.2014.02.002.

Hill, A. B. (1965). The Environment and Disease : Association or Causation ? *Proceedings of the Royal Society of Medicine* 58 (2), 295-300.

Hill, T. C., K. A. Walsh, J. A. Harris et B. F. Moffett (2003). Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology* 43 (1), 1-11. doi : 10.1111/j.1574-6941.2003.tb01040.x.

Hoek, T. A., K. Axelrod, T. Biancalani, E. A. Yurtsev, J. Liu et J. Gore (2016). Resource Availability Modulates the Cooperative and Competitive Nature of a Microbial Cross-Feeding Mutualism. *PLOS Biology* 14 (8), e1002540. doi : 10.1371/journal.pbio.1002540.

Horner-Devine, M. C., M. Lage, J. B. Hughes et B. J. M. Bohannan (2004). A taxa–area relationship for bacteria. *Nature* 432 (7018), 750-753. doi : 10.1038/nature03073.

Huffnagle, G. B. et M. C. Noverr (2013). The emerging world of the fungal microbiome. *Trends in Microbiology* 21 (7), 334-341. doi : 10.1016/j.tim.2013.04.002.

Hughes, J. B. et J. J. Hellmann (2005). The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity. *Methods in Enzymology*. Tome 397. (1995), p. 292-308. doi : 10.1016/S0076-6879(05)97017-1.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*. doi : 10.5169/seals-266450.

Jiang, X. et X. Hu (2016). Microbiome data mining for microbial interactions and relationships. *Big Data Analytics : Methods and Applications*. Springer India, p. 221-235. doi : 10.1007/978-81-322-3628-3_12.

Johnson, K. V.-A. et K. R. Foster (2018). Why does the microbiome affect behaviour ? *Nature Reviews Microbiology* 16 (10), 647-655. doi : 10.1038/s41579-018-0014-3.

Jonsson, V., T. Österlund, O. Nerman et E. Kristiansson (2018). Modelling of zero-inflation improves inference of metagenomic gene count data. *Statistical Methods in Medical Research*, 096228021881135. doi : 10.1177/0962280218811354.

Jousset, A. et al. (2017). Where less may be more : how the rare biosphere pulls ecosystems strings. *The ISME Journal* 11 (4), 853-862. doi : 10.1038/ismej.2016.174.

Jovel, J. et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Frontiers in Microbiology* 7 (APR), 1-17. doi : 10.3389/fmicb.2016.00459.

Karimi, B., P. A. Maron, N. Chemidlin-Prevost Boure, N. Bernard, D. Gilbert et L. Ranjard (2017). Microbial diversity and ecological networks as indicators of environmental quality. *Environmental Chemistry Letters* 15 (2), 265-281. doi : 10.1007/s10311-017-0614-6.

Keskin, O., N. Tuncbag et A. Gursoy (2016). Predicting Protein-Protein Interactions from the Molecular to the Proteome Level. doi : 10.1021/acs.chemrev.5b00683.

Kho, Z. Y. et S. K. Lal (2018). The human gut microbiome - A potential controller of wellness and disease. *Frontiers in Microbiology* 9 (AUG), 1-23. doi : 10.3389/fmicb.2018.01835.

Klassen, J. L. (2018). Defining microbiome function. *Nature Microbiology* 3 (8), 864-869. doi : 10.1038/s41564-018-0189-4.

Knight, R. et al. (2018). Best practices for analysing microbiomes. *Nature Reviews Microbiology* 16 (July), 1-13. doi : 10.1038/s41579-018-0029-9.

Konopka, A. (2009). What is microbial community ecology ? *The ISME Journal* 3 (11), 1223-1230. doi : 10.1038/ismej.2009.88.

Krumsiek, J., K. Suhre, T. Illig, J. Adamski et F. J. Theis (2011). Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Systems Biology* 5 (1), 21. doi : 10.1186/1752-0509-5-21.

Kumar, M. S., E. V. Slud, K. Okrah, S. C. Hicks, S. Hannenhalli et H. Corrada Bravo (2018). Analysis and correction of compositional bias in sparse sequencing count data. *BMC genomics* 19 (1), 799. doi : 10.1186/s12864-018-5160-5.

Kurtz, Z. D., C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser et R. A. Bonneau (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology* 11 (5). Sous la direction de C. von Mering, e1004226. doi : 10.1371/journal.pcbi.1004226.

Layeghifard, M., D. M. Hwang et D. S. Guttman (2017). Disentangling Interactions in the Microbiome : A Network Perspective. *Trends in Microbiology* 25 (3), 217-228. doi : 10.1016/j.tim.2016.11.008.

Li, C., K. M. K. Lim, K. R. Chng et N. Nagarajan (2016). Predicting microbial interactions through computational approaches. *Methods* 102, 12-19. doi : 10.1016/j.ymeth.2016.02.019.

Lidicker, W. Z. (1979). A Clarification of Interactions in Ecological Systems. *BioScience* 29 (8), 475-477. doi : 10.2307/1307540.

Lindén, A. et S. Mäntyniemi (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology* 92 (7), 1414-1421. doi : 10.1890/10-1831.1.

Lo, C. et R. Marculescu (2017). MPLasso : Inferring microbial association networks using prior microbial knowledge. *PLOS Computational Biology* 13 (12). Sous la direction de N. Segata, e1005915. doi : 10.1371/journal.pcbi.1005915.

Lo, C. et R. Marculescu (2018). PGLasso : Microbial Community Detection through Phylogenetic Graphical Lasso. *CEUR Workshop Proceedings* 2065, 54-57.

Locey, K. J. et J. T. Lennon (2016). Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences* 113 (21), 5970-5975. doi : 10.1073/pnas.1521291113.

Love, M. I., W. Huber et S. Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15 (12), 1-21. doi : 10.1186/s13059-014-0550-8.

Lozupone, C. et R. Knight (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology* 71 (12), 8228-8235. doi : 10.1128/AEM.71.12.8228-8235.2005.

Lozupone, C. A., M. Hamady, S. T. Kelley et R. Knight (2007). Quantitative and qualitative β diversity measures lead to different insights into factors that structure microbial communities. *Applied and Environmental Microbiology* 73 (5), 1576-1585. doi : 10.1128/AEM.01996-06.

Lozupone, C. A., J. I. Stombaugh, J. I. Gordon, J. K. Jansson et R. Knight (2012). Diversity, stability and resilience of the human gut microbiota. *Nature* 489 (7415), 220-230. doi : 10.1038/nature11550.

Lugo-Martinez, J., D. Ruiz-Perez, G. Narasimhan et Z. Bar-Joseph (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* 7 (1), 54. doi : 10.1186/s40168-019-0660-3.

Lunn, D. J., A. Thomas, N. Best et D. Spiegelhalter (2000). WinBUGS - A Bayesian modelling framework : Concepts, structure, and extensibility. *Statistics and Computing* 10 (4), 325-337. doi : 10.1023/A:1008929526011.

Lynch, M. D. J. et J. D. Neufeld (2015). Ecology and exploration of the rare biosphere. *Nature Reviews Microbiology* 13 (4), 217-229. doi : 10.1038/nrmicro3400.

Mainali, K., S. Bewick, B. Vecchio-Pagan, D. Karig et W. F. Fagan (2019). Detecting interaction networks in the human microbiome with conditional Granger causality. *PLOS Computational Biology* 15 (5), e1007037. doi : 10.1371/journal.pcbi.1007037.

Mandakovic, D. et al. (2018). Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. *Scientific Reports* 8 (1), 1-12. doi : 10.1038/s41598-018-23931-0.

Martiny, A. C. (2019). High proportions of bacteria are culturable across major biomes. *The ISME Journal* 13 (8), 2125-2128. doi : 10.1038/s41396-019-0410-3.

McKnight, D. T., R. Huerlimann, D. S. Bower, L. Schwarzkopf, R. A. Alford et K. R. Zenger (2019). Methods for normalizing microbiome data : An ecological perspective. *Methods in Ecology and Evolution* 10 (3), 389-400. doi : 10.1111/2041-210X.13115.

McMurdie, P. J. et S. Holmes (2014). Waste Not, Want Not : Why Rarefying Microbiome Data Is Inadmissible. *PLoS Computational Biology* 10 (4). Sous la direction d'A. C. McHardy, e1003531. doi : 10.1371/journal.pcbi.1003531.

Meinshausen, N. et P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics* 34 (3), 1436-1462. doi : 10.1214/009053606000000281.

Mohajeri, M. H., R. J. Brummer, R. A. Rastall, R. K. Weersma, H. J. Harmsen, M. Faas et M. Eggersdorfer (2018). The role of the microbiome for human health : from basic science to clinical applications. *European Journal of Nutrition* 57 (1), 1-14. doi : 10.1007/s00394-018-1703-4.

Navas-Molina, J. A. et al. (2013). Advancing Our Understanding of the Human Microbiome Using QIIME. *Methods in Enzymology*. 1re édition. Tome 531. Elsevier Inc., p. 371-444. doi : 10.1016/B978-0-12-407863-5.00019-8.

Nemergut, D. R. et al. (2013). Patterns and processes of microbial community assembly. *Microbiology and Molecular Biology Reviews* 77 (3), 342-356. doi : 10.1128/MMBR.00051-12.

Nielsen, H. B. et al. (2014). Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nature Biotechnology* 32 (8), 822-828. doi : 10.1038/nbt.2939.

Oksanen, J. et al. (2013). vegan : Community Ecology Package. R package.

Oulas, A., C. Pavloudi, P. Polymenakou, G. A. Pavlopoulos, N. Papanikolaou, G. Kotoulas, C. Arvanitidis et L. Iliopoulos (2015). Metagenomics : Tools and Insights for Analyzing Next-Generation Sequencing Data Derived from Biodiversity Studies. *Bioinformatics and Biology Insights* 9, 75-88. doi : 10.4137/BBI.S12462.

Ovaskainen, O., G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin et N. Abrego (2017). How to make more out of community data ? A conceptual framework and its implementation as models and software. *Ecology Letters* 20 (5). Sous la direction de J. Chave, 561-576. doi : 10.1111/ele.12757.

Pacheco, A. R. et D. Segrè (2019). A multidimensional perspective on microbial interactions. *FEMS Microbiology Letters* 366 (11). doi : 10.1093/femsle/fnz125.

Paulson, J. N., O. C. Stine, H. C. Bravo et M. Pop (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods* 10 (12), 1200-1202. doi : 10.1038/nmeth.2658.

Pearson, K. (1897). Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London* 60 (359-367), 489-498. doi : 10.1098/rspl.1896.0076.

Pereira, M. B., M. Wallroth, V. Jonsson et E. Kristiansson (2018). Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics* 19 (1), 1-17. doi : 10.1186/s12864-018-4637-6.

Plummer, M. (2003). JAGS : A program for analysis of Bayesian graphical models using Gibbs sampling. *Workshop on Distributed Statistical Computing*. Vienna.

Pollock, J., L. Glendinning, T. Wisedchanwet et M. Watson (2018). The Madness of Microbiome : Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology* 84 (7). Sous la direction de S.-J. Liu, e02627-17. doi : 10.1128/AEM.02627-17.

Popovic, G. C., D. I. Warton, F. J. Thomson, F. K. C. Hui et A. T. Moles (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*. Sous la direction de D. Murrell, 2041-210X.13247. doi : 10.1111/2041-210X.13247.

Qin, J. et al. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464 (7285), 59-65. doi : 10.1038/nature08821.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies et F. O. Glöckner (2013). The SILVA ribosomal RNA gene database project : Improved data processing and web-based tools. *Nucleic Acids Research* 41 (D1), 590-596. doi : 10.1093/nar/gks1219.

Quinn, T. P., I. Erb, M. F. Richardson et T. M. Crowley (2018). Understanding sequencing data as compositions : an outlook and review. *Bioinformatics* 34 (March), 2870-2878. doi : 10.1093/bioinformatics/bty175.

Razzauti, M., M. Galan, M. Bernard, S. Maman, C. Klopp, N. Charbonnel, M. Vayssier-Taussat, M. Eloit et J.-F. Cosson (2015). A Comparison between Transcriptome Sequencing and 16S Metagenomics for Detection of Bacterial Pathogens in Wildlife. *PLOS Neglected Tropical Diseases* 9 (8). Sous la direction de P. L. C. Small, e0003929. doi : 10.1371/journal.pntd.0003929.

Reshef, D. N., Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher et P. C. Sabeti (2011). Detecting Novel Associations in Large Data Sets. *Science* 334 (6062), 1518-1524. doi : 10.1126/science.1205438.

Ridout, M., C. G. B. Demétrio et J. Hinde (1998). Models for count data with many zeros. *International Biometric Conference*. Cape Town.

Robinson, M. D. et A. Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* 11 (3), R25. doi : 10.1186/gb-2010-11-3-r25.

Röttjers, L. et K. Faust (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews* 42 (6), 761-780. doi : 10.1093/femsre/fuy030.

Sanders, H. L. (1968). Marine Benthic Diversity : A Comparative Study. *The American Naturalist* 102 (925), 243-282. doi : 10.1086/282541.

Schloss, P. D. et J. Handelsman (2004). Status of the Microbial Census. *Microbiology and Molecular Biology Reviews* 68 (4), 686-691. doi : 10.1128/MMBR.68.4.686-691.2004.

Schloss, P. D., S. L. Westcott et al. (2009). Introducing mothur : Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology* 75 (23), 7537-7541. doi : 10.1128/AEM.01541-09.

Schmidt, T. S. B., J. F. Matias Rodrigues et C. von Mering (2017). A family of interaction-adjusted indices of community similarity. *The ISME Journal* 11 (3), 791-807. doi : 10.1038/ismej.2016.139.

Shah, A. A., F. Hasan, A. Hameed et S. Ahmed (2008). Biological degradation of plastics : A comprehensive review. *Biotechnology Advances* 26 (3), 246-265. doi : 10.1016/j.biotechadv.2007.12.005.

Simon, J. C., J. R. Marchesi, C. Mougel et M. A. Selosse (2019). Host-microbiota interactions : From holobiont theory to analysis. *Microbiome* 7 (1), 1-5. doi : 10.1186/s40168-019-0619-4.

Sohn, M. B. et H. Li (2018). A GLM-based latent variable ordination method for microbiome samples. *Biometrics* 74 (2), 448-457. doi : 10.1111/biom.12775.

Stackebrandt, E. et B. M. Goebel (1994). Taxonomic Note : A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology* 44 (4), 846-849. doi : 10.1099/00207713-44-4-846.

Steen, A. D., A. Crits-Christoph, P. Carini, K. M. DeAngelis, N. Fierer, K. G. Lloyd et J. Cameron Thrash (2019). High proportions of bacteria and archaea across most biomes remain uncultured. *The ISME Journal*. doi : 10.1038/s41396-019-0484-y.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 64 (3), 479-498. doi : 10.1111/1467-9868.00346.

Szklarczyk, D. et L. J. Jensen (2015). Protein-protein interaction databases. *Methods in Molecular Biology* 1278, 39-56. doi : 10.1007/978-1-4939-2425-7_3.

Tan, P. N., V. Kumar et J. Srivastava (2004). Selecting the right objective measure for association analysis. *Information Systems* 29 (4), 293-313. doi : 10.1016/S0306-4379(03)00072-3.

Thorsen, J., A. Brejnrod, M. Mortensen, M. A. Rasmussen, J. Stokholm, W. A. Al-Soud, S. Sørensen, H. Bisgaard et J. Waage (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4 (1), 62. doi : 10.1186/s40168-016-0208-8.

Vaumourin, E., G. Vourc'h, P. Gasqui et M. Vayssier-Taussat (2015). The importance of multiparasitism : examining the consequences of co-infections for human and animal health. *Parasites & vectors* 8, 545. doi : 10.1186/s13071-015-1167-9.

Vayssier-Taussat, M., E. Albina et al. (2014). Shifting the paradigm from pathogens to pathobiome : new concepts in the light of meta-omics. *Frontiers in Cellular and Infection Microbiology* 4 (March), 1-7. doi : 10.3389/fcimb.2014.00029.

Vayssier-Taussat, M., M. Kazimirova et al. (2015). Emerging horizons for tick-borne pathogens : from the 'one pathogen–one disease' vision to the pathobiome paradigm. *Future Microbiology* (November), fmb.15.114. doi : 10.2217/fmb.15.114.

Vázquez-Castellanos, J. F., A. Biclot, G. Vrancken, G. R. Huys et J. Raes (2019). Design of synthetic microbial consortia for gut microbiota modulation. *Current Opinion in Pharmacology* 49, 52-59. doi : 10.1016/j.coph.2019.07.005.

Vellend, M. (2010). Conceptual Synthesis in Community Ecology. *The Quarterly Review of Biology* 85 (2), 183-206. doi : 10.1086/652373.

Venter, J. C. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304 (5667), 66-74. doi : 10.1126/science.1093857.

Venturelli, O. S., A. C. Carr, G. Fisher, R. H. Hsu, R. Lau, B. P. Bowen, S. Hromada, T. Northen et A. P. Arkin (2018). Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular Systems Biology* 14 (6), e8157. doi : 10.15252/msb.20178157.

Wang, T., Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang et W. Chen (2016). FastGGM : An Efficient Algorithm for the Inference of Gaussian Graphical Model in Biological Networks. *PLoS Computational Biology* 12 (2), 1-16. doi : 10.1371/journal.pcbi.1004755.

Wang, Y., U. Naumann, S. T. Wright et D. I. Warton (2012). mvabund - an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution* 3 (3), 471-474. doi : 10.1111/j.2041-210X.2012.00190.x.

Warton, D. I., S. T. Wright et Y. Wang (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution* 3 (1), 89-101. doi : 10.1111/j.2041-210X.2011.00127.x.

Weiss, S., W. Van Treuren et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME Journal* 10 (7), 1669-1681. doi : 10.1038/ismej.2015.235.

Weiss, S., Z. Z. Xu et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5 (1), 27. doi : 10.1186/s40168-017-0237-y.

Whitman, W. B., D. C. Coleman et W. J. Wiebe (1998). Prokaryotes : The unseen majority. *Proceedings of the National Academy of Sciences* 95 (12), 6578-6583. doi : 10.1073/pnas.95.12.6578.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley Publishing.

Wunsch, G., F. Russo et M. Mouchart (2010). Do We Necessarily Need Longitudinal Data to Infer Causal Relations ? *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 106 (1), 5-18. doi : 10.1177/0759106309360114.

Xia, Y. et J. Sun (2017). Hypothesis testing and statistical analysis of microbiome. *Genes and Diseases* 4 (3), 138-148. doi : 10.1016/j.gendis.2017.06.001.

Xu, L., A. D. Paterson, W. Turpin et W. Xu (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLOS ONE* 10 (7). Sous la direction d'Y. Xia, e0129606. doi : 10.1371/journal.pone.0129606.

Yang, Y., N. Chen et T. Chen (2017). Inference of Environmental Factor-Microbe and Microbe-Microbe Associations from Metagenomic Data Using a Hierarchical Bayesian Statistical Model. *Cell Systems* 4 (1), 129-137.e5. doi : 10.1016/j.cels.2016.12.012.

Yuan, H., S. He et M. Deng (2019). Compositional data network analysis via lasso penalized D-trace loss. *Bioinformatics*. Sous la direction de B. Berger. doi : 10.1093/bioinformatics/btz098.

Yule, G. U. (1912). On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society* 75 (6), 579. doi : 10.2307/2340126.

Annexe A

Biogéographie, génétique et temps : quel impact sur la structuration du microbiote des abeilles ?

The role of biogeography, genetics and time in shaping diversity of honeybee gut microbiota

Iris Eouzan[1,*], Ana Marta Muñoz[2], Arnaud Cougoul[3], Lionel Garnery[4,5], Maria Alice Pinto[6], Damien Delalande[5], Sylvie Houte[7], Andone Estonba[2], Iratxe Montes[2], Patrick Gasqui[3], Xavier Bailly[3], Télesphore Sime-Ngando[1], David G. Biron[1]

[1]Laboratoire Microorganismes : Génome et Environnement, UMR CNRS 6023, Université Clermont Auvergne, Campus Universitaire des Cézeaux, 1 Impasse Amélie Murat, 63178 Aubière cedex, France; [2]Department of Genetics, Physical Anthropology and Animal Physiology, University of the Basque Country (UPV/EHU), Barrio Sarriena s/n, 48940 Leioa (Bizkaia), Spain; [3]Institut National de la Recherche Agronomique, UMR 346 Épidémiologie des maladies animales et zoonotiques, Saint Genès Champanelle, France; [4]Laboratoire Evolution, Génomes et Spéciation, UMR CNRS 9191, Bâtiment 13, Avenue de la Terrasse, 91198 Gif-sur-Yvette, France; [5]Saint Quentin en Yvelines, Université de Versailles, 45 Avenue des Etats-Unis, 78000 Versailles, France; [6]Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal; [7]Centre d'Etudes Biologique de Chizé, UMR CNRS 7372, Université de la Rochelle, 79360 Villiers-en-Bois, France.

5.1 Abstract

Many studies have been conducted recently to understand the effect of different eco-ethological, environmental or genetic factors on the microbiota communities of many animals. They showed that factors such as diet or environment have a significant effect and contribute through the microbiota to the host health. Here, a metagenomic survey (16s rRNA) was done on the gut microbiota for 36 beehives of the M evolutive lineage (i.e. the black bee and the Iberian bee) along a geographic gradient from southern Portugal to northern France. We analyzed the effect of genetic (i.e. beehive haplotype on DNAmt), environmental (i.e. country, climate, apiary, landscape) and temporal (i.e. three months during summer) factors to decipher which one(s) contribute(s) to the structure and diversity of the bacterial populations present in these honeybees gut. Our samples contained bacteria mainly belonging to the proteobacteria phylum, represented in large quantities by the genera *Sphingomonas* and *Snodgrassella*, and the French conservatories appeared to have the biggest alpha diversity. Although all factors had a significant effect on bacterial communities' structure, the conservatory, beehive DNAmt haplotype and landscape had the biggest effect. However, when we considered the interactions, the time appeared to increase greatly the effect of the

conservatory alone. Our data suggest that diet is also an important factor. Moreover, these data provide a better understanding of the effect of several factors such as genetics, which are not well addressed in the literature, and open new perspectives to better understand the relationship between the host's environment and its microbiota.

Key words: biogeography, genetics, DNAmt, gut microbiota, black bee, Iberian bee

5.2 Introduction

Animal guts contain diverse microbial communities that are often dominated by bacteria but also include archaea, viruses, protozoa and fungi (Sommer and Bäckhed, 2013). This microbiota has an important role in different processes such as digestion, detoxification of harmful molecules, immunity or resistance to infectious diseases (Engel and Moran, 2013; Flint et al., 2012; Hooper et al., 2012). Studies have been multiplied to understand what influences the structuration of gut microbiota. They have so far shown the importance of various factors on guts microbial diversity. In mammals, many studies showed that host's genetics, diet but also geography has a strong influence on ecological and diversity patterns of microbial gut communities in humans (Gupta et al., 2017; Ley et al., 2005, 2006) but also other mammals (Linnenbrink et al., 2013; Ochman et al., 2010; Phillips et al., 2012). It is also the case in animal groups like reptiles (Lankau et al., 2012) or even fishes, in which microbiota patterns also show variations according to the water salinity or the trophic level of the host (Sullam et al., 2012).

Regarding insects, the host's ecology, species, diet, or even sex plays a key role in shaping ecological and diversity patterns of gut microbiota (Jones et al., 2018; Kwong et al., 2017; Santo Domingo et al., 1998; Van Treuren et al., 2015; Yun et al., 2014). Insects are easy to study, as they contain relatively few microbial species compared to mammalians (Engel and Moran, 2013). Besides, the microbiota of eusocial insects like honeybees, *Apis mellifera* (Linnaeus 1758) (Hymenoptera, Apidae), has some similarities with that of mammals, while being of simpler composition (Kwong and Moran, 2016). These important pollinators have a digestive tract microbiota composed of highly specialized bacteria including Firmicutes, Actinobacteria, α- and γ-Proteobacteria, that are mainly transmitted by contact between

workers in the hive (Anjum et al., 2018; Engel et al., 2012; Kwong and Moran, 2016; Kwong et al., 2017; Martinson et al., 2011; Moran et al., 2012).

Some rare studies have shown that bees' genetic diversity can directly influence the diversity patterns of their microbial gut communities (Mattila et al., 2012), just like geography (Hroncova et al., 2015). To our knowledge, no studies comparing the effects of host's geographic position and its genetic origin along a geographic gradient have been carried out to understand which of these factors most impacts the shaping of the diversity patterns of honeybee microbiota. In our survey, worker honeybees were sampled from 36 different beehives (i.e. six hives by apiary) along a geographic gradient from southern Portugal to northern France, and their microbiota were analyzed thanks to Illumina 16S rRNA sequencing. Genetic analysis on bee workers DNAmt were done to determine the evolutive lineage of each beehive and their haplotype, in order to assess the effect of genetic (i.e. beehive DNAmt haplotype) and environmental (i.e. country, climate, apiary, landscape) factors. Those factors were analyzed in order to understand which one(s) contribute(s) the most to the structure and diversity of the bacterial populations observed along a geographic gradient of the M evolutive lineage and for two honeybee subspecies, the black bee *Apis apis mellifera* and the Iberian bee *Apis apis iberiensis*.

5.3 Material and methods

5.3.1 Sampling

The study was conducted in six conservation centers in West-Europe to study *A. m. mellifera* in France and Spain, and *A. m. iberiensis* in Portugal. The French conservation centers are located at Champart in the region of "Ile-de-France", Pontaumur in the region of "Auvergne-Rhône-Alpes" and Chizé in the region of "Nouvelle-Aquitaine". The Spanish conservation center is located at Bilbao, in the region of "Bizkaia". The Portuguese conservation centers are located at two latitudinal extremes of the country, Gimonde and Zavial, in the regions of "Trás-os-Montes" and "Algarve", respectively (Figure 5.1).

In each conservation center, six healthy beehives were randomly chosen to be used for this study. The bees were collected directly on a frame near the brood in July, August and September 2016, in a 50 ml falcon tube. About fifty bees were collected in each sampling. The

tubes were then frozen at -80°C until DNA extraction, and only one bee per sample was putted in 90° ethanol for DNAmt genetic analysis.

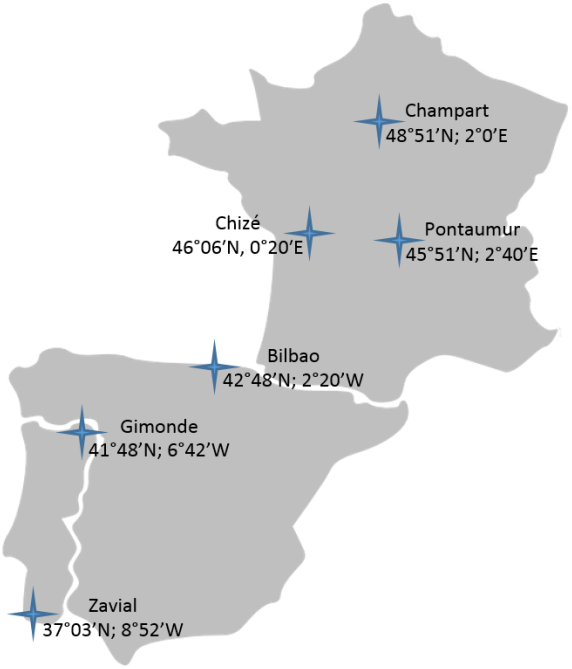


Figure 5.1: sampling locations across France, Spain and Portugal.

5.3.2 Environment

We classified our colonies according to three environmental parameters: beekeeping's landscape, climate and country. Pontaumur, Bilbao and Gimonde are "mountain beekeeping", whereas Chizé and Zavial are "plain beekeeping". The last one, Champart, was classified as "forest beekeeping". The climate was defined according to the region where the beehives are. Thus, Champart, Chizé and Bilbao have an oceanic climate, Pontaumur has a continental climate and Gimonde and Zavial are in a Mediterranean climate.

5.3.3. Genetic analysis

According to morphometric, genetic, physiological and behavioral studies, honeybee subspecies are divided into four evolutionary lineages: A (African), M (West-Mediterranean), C (North-Mediterranean) and O (Turkey and Caucasus) (Miguel et al., 2007, 2011; Ruttner, 1988). Lineage A corresponds to subspecies found in Africa, M to Western European subspecies (i.e. the black bee *A. m. mellifera* and to the Iberian bee *A. m iberiensis*), group C to the subspecies living in Eastern Europe (e.g. *A. m. carnica*, *A. m. ligustica*), and group O

includes subspecies present in Turkey and the Caucasus (Han et al., 2012). In total, there are currently 26 subspecies of *A. mellifera* (Miguel et al., 2011).

DNAmt is a circular molecule contained in the mitochondria of cells. Unlike nuclear DNA, the transmission of this molecule is only maternal. In the honeybee, the DNAmt has therefore a very strong colony marker power since all workers of a same colony have the same mother, the queen. Thus, the study of a single bee is sufficient to characterize the colony. The maternal transmission of the molecule makes it a particularly suitable marker for the determination of the maternal origin of a colony and of the queen. This test permits to characterize on the one hand the evolutionary lineage (M, A, C and O) of each colony, but also to study the intra-lineage polymorphism among all the haplotypes observed, those that may possibly correspond to local variants.

The intergenic COI-COII region of DNAmt was studied according to the protocol described by (Garnery et al., 1993) to identify the evolutive lineage. The intergenic region, COI-COII is amplified by PCR (Polymerase Chain Reaction) using two primers (E2: 5'-GGCAGAATAAGTGCATTG-3', H2: 5'-CAATATCATTGATGACC-3') developed by (Garnery et al., 1992). The PCR products are then deposited on 1.4% agarose gel and electrophoresed. Finally, to characterize the beehive haplotype, the amplified products obtained are subjected to a digestion by the restriction enzyme DraI. Restricted DNA fragments were separated on 5% and 10% acrylamide gels and stained with ethidium bromide.

5.3.4 Metagenomic 16s RNA analysis

5.3.4.1 DNA extraction

Only the bees' guts were used for DNA extraction. For each sample, three replicates of nine bees each were dissected: on ice, the sting was pulled out of the bee with tweezers, and the whole gut was putted in a Tenbroek Potter homogenizer. Each replicate was crushed about twenty times in 1.5 ml phosphate buffer. The obtained supernatant was centrifuged at 8 000 rcf, at 4°C, during 10 min. The Tenbroek Potter homogenizers were washed between each use under water, filled with 2% steranios solution for 30min, rinsed with 90% ethanol and dried before next use. The DNA extraction was made using QIAamp DNA Mini Kit following the manufacturers' guidelines.

5.3.4.2 Pyrosequencing of the 16S rRNA gene and sequence processing

The 515F-806R primer pair spanning the V4 region of the bacterial 16S rRNA gene, located in the SSU, was used to determine the bacterial community present in the samples. These primers contained in the forward primers a 12 bp barcode sequence to identify each sample ("Earth Microbiome Project"; http://www.earthmicrobiome.org/). These primers are used to amplify prokaryotes (but they also amplify mitochondrial and chloroplast DNA), and give a ~291 bp amplicon. The following conditions were used for the V4 region amplification PCR: 0.3 µL Taq polymerase (Go Flexi Promega Taq), 5 µL of Taq Buffer (5x), 2 µL of MgCl2+ (50mM), 2 µL of 2.5 mM dNTPs, 0.5 µL of each primer (10 µM), 1 µL of extracted DNA (15-60 ng/µL), and bidistilled water to reach a total volume of 25 µL. The PCR program was the following: initial denaturing step at 95 ºC for 4 min; 35 cycles of 15 secs at 95 ºC, 30 sec at 50 ºC, 30 sec at 72 ºC; and a final elongation step at 72°C for 2 min. The PCR products were then checked on a 1.5% agarose gel stained with ethidium bromide. The CleanPCR kit (Cleanna) was used using magnetic beads for DNA purification of the samples. The quantification of purified DNA was made using Qubit™ v2.0 (ThermoFisher Scientific) and the samples were normalized into an 8 pM pool. The paired-end sequencing of the pool was carried out on an Illumina MiSeq sequencer at the Sequencing and Genotyping Unit of the University of the Basque country (SGIKER) using the kit v2 PE 2 x 150 bp (300 cycles); 10% of PhiX were added as external control of the sequencing process.

5.3.4.3 Data analysis

The paired-end reads were trimmed with sickle-quality-based-trimming (Sickle v1.33; Joshi & Fass, 2011), and merged using PEAR v0.9.10 (Zhang et al., 2014). Hereafter, the sequences were processed by QIIME v1.9 pipeline (Caporaso et al., 2010). The following scripts were then sequentially run: split_libraries_fastq.py for demultiplexing the samples, maximum number of errors in barcode = 0, and quality filtering with a Phred quality score ≥ 20. Taxonomy was assigned using pick_open_reference_otus.py, with default settings, for 97% OTU (operational taxonomic unit) clustering. The chloroplast and mitochondrial sequences and OTUs with frequency lower than 10 were filtered, as well as the samples with less than 5000 reads.

Shannon and Chao1 indexes for the alpha diversity were calculated with R v.3.5.1 software (*phyloseq* package). The OTU-table was normalized using DESeq2 command, before calculating the Bray-Curtis and Unifrac distance matrix at the OTU scale. The beta diversity was represented by a non-metric multidimensional scaling (NMDS) projection based on the resultant Bray-Curtis matrix (R; *phyloseq* package). Finally, Bray-Curtis and UniFrac distances were used in order to do a NPMANOVA statistical analysis. The Non-parametric MANOVA (NPMANOVA, (Anderson, 2001), also called Adonis in the vegan R package and QIIME) is a non-parametric analyses of variance that has been used to test for differences in microbial community composition. The tested factors for this NPMANOVA analysis were: conservatory, beehive DNAmt haplotype, landscape, country, climate and month. Different models were calculated for these two distances, to combine the effects of some parameters on the microbiota diversity.

5.4 Results

5.4.1. Genetic analysis

Thanks to the mitochondrial DNA analysis, we were able to classify our bee colonies into ten different DNAmt haplotypes, belonging to three lineages: A, C and M (Table 1). Five bee colonies (in Champart and Chizé) appeared to be C lineage, whereas Gimonde and Zavial have only A lineage. The Portugal is a South-North cline with unparalleled levels of haplotype diversity and complexity specific to this country and to the M lineage resulting from the coexistence of African (A) and western European (M) lineages (Cánovas et al., 2011; Franck et al., 1998). Some haplotypes were found in more than one conservatory (C1, M4, A1), whereas the others only appeared in one sanctuary and for some of them only in one colony (M11, M17, A2).

Table 1: Information available for each conservatory, and used for the statistical analysis.

Conservatory	Country	Climate	Landscape	Haplotypes					
				Hive 1	Hive 2	Hive 3	Hive 4	Hive 5	Hive 6
Champart	**France**	Oceanic	Forest	C1	C1	C1	C1	M4	M11
Pontaumur	**France**	Continental	Mountain	M66'	M4	M4'	M4'	M4'	M66'
Chizé	**France**	Oceanic	Plain	/	M4	C1	M4	M4	M17
Bilbao	**Spain**	Oceanic	Mountain	M7	M4	M4	M4	M4	M7
Gimonde	**Portugal**	Mediterranean	Mountain	A2	A11	A1	A11	A1	A1
Zavial	**Portugal**	Mediterranean	Plain	A1	A1	A1	A1	A1	A1

5.4.2 Composition of the honeybee gut microbiota for the geographic gradient studied

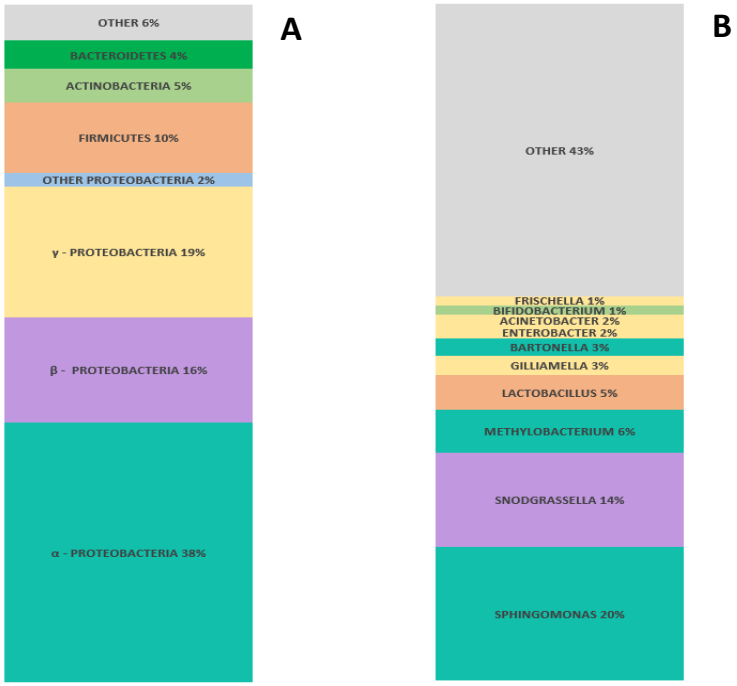


Figure 5.2: Classification phyla-wise (A) and genera-wise (B) of the bacteria found in *A.mellifera* guts of the M evolutive lineage along the geographic gradient from southern Portugal to northern France. The "other" category is represented by phyla and genus that represent less than 1% of the samples.

The sequencing allowed us to isolate and identify 3 267 OTUs belonging to 24 bacteria phyla and 473 distinct genera. The identified bacteria belong for the great majority to the proteobacteria (75%), firmicutes (10%), actinobacteria (5%) and bacteroidetes (4%) phylum, the other phyla being represented by less than 1% of the bacteria (Figure 5.2). Among those phyla, the proteobacteria are represented in majority by *Sphingomonas* (20%), *Snodgrassella* (14%), *Methylobacterium* (6%), *Gilliamella* (3%), *Bartonella* (3%), *Enterobacter* (2%), *Acinetobacter* (2%) and *Frischella* (1%), whereas the firmicutes and the actinobacteria are mostly represented by *Lactobacillus* (5%) and *Bifidobacterium* (1%), respectively (Figure 5.2).

5.4.3 Alpha and beta diversity

The alpha diversity was calculated using Shannon and Chao1 indexes (**Figure 5.3**). The French conservatories (Champart, Pontaumur and Chizé) have a bigger OTU diversity than the

Spanish and Portuguese colonies according to Shannon index. When the singletons are taken into account, the estimated number of species appears lower for Champart, according to Chao1 index.

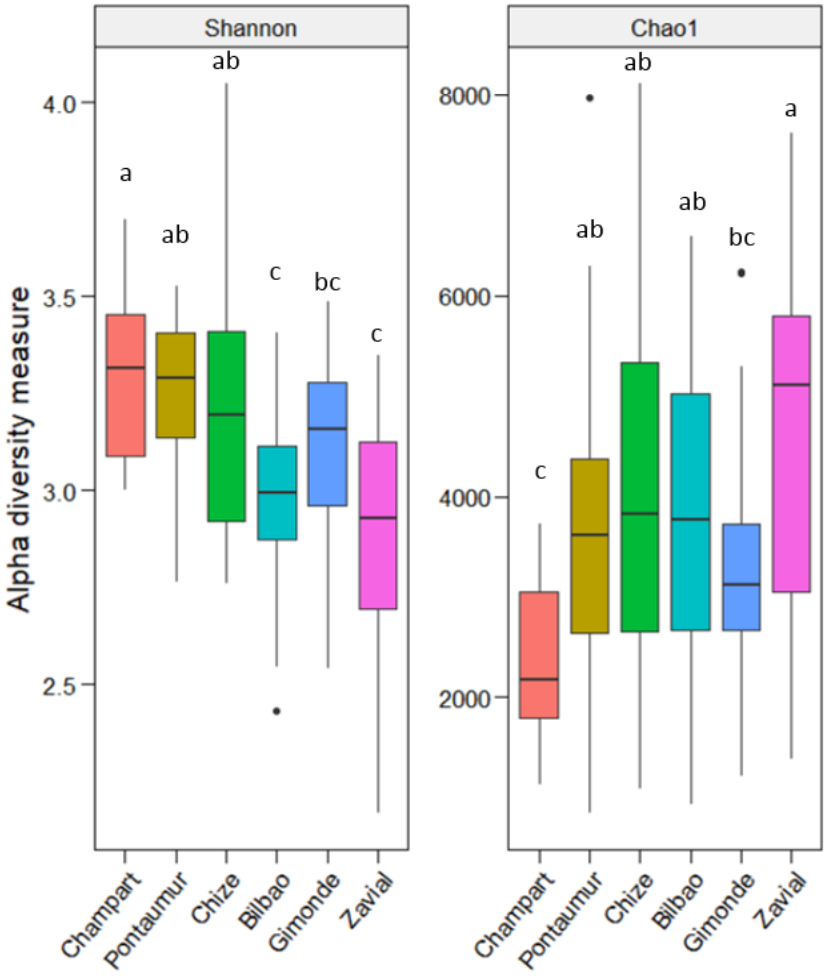


Figure 5.3: Boxplots for the Shannon and Chao1 indexes for the six apiaries (sites) studied and belonging to the M evolutive lineage of the European bee along the geographic gradient from southern Portugal to northern France. For each apiary, data of all six beehives taken one time during three months were putted together. The letters indicate the significance of the differences, plots without letters in common being significantly different (Tukey HSD, p<0.05).

The beta diversity was calculated with the Bray-Curtis matrix and represented with a NMDS projection in Figure 5.4. Chizé appears to have the biggest diversity variation, whereas Champart has the lowest.

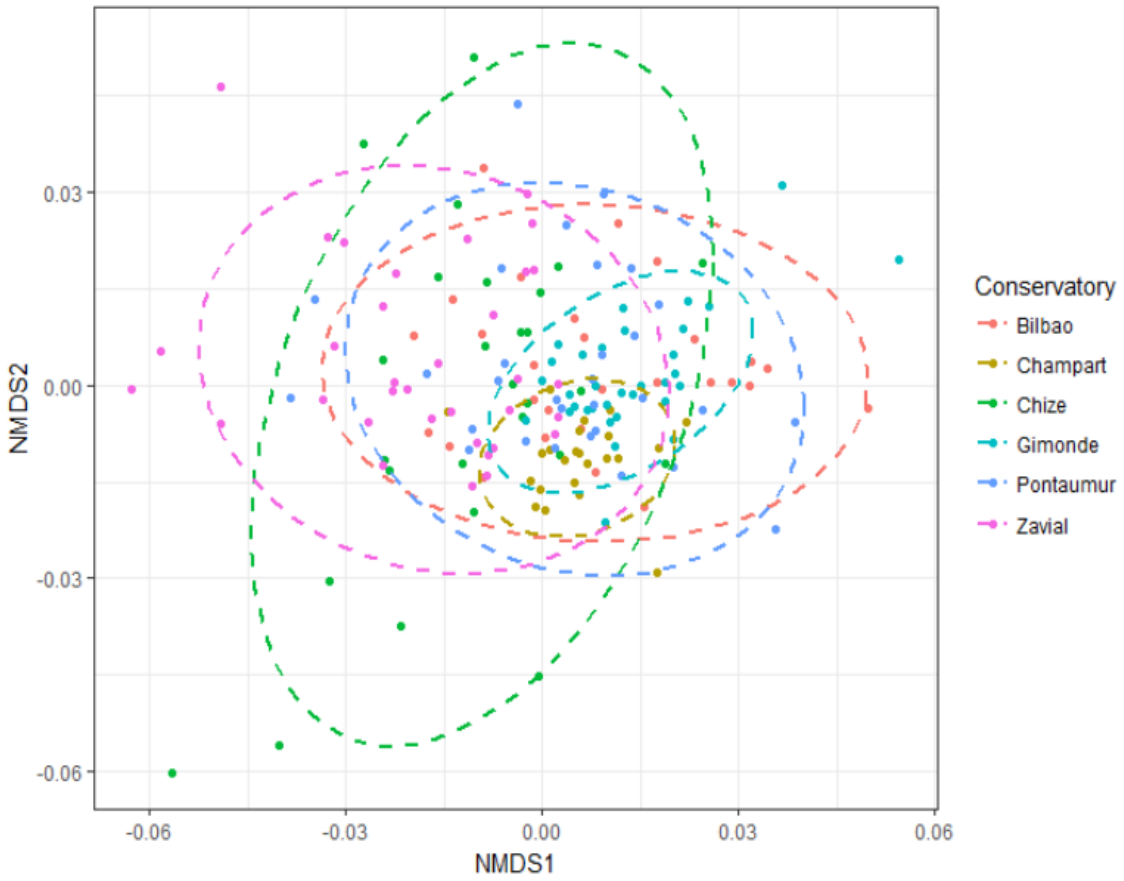


Figure 5.4: NMDS projection of the Bray-Curtis matrix. For each studied apiary (site), the NMDS analysis was done by putting together the data of all six beehives taken one time during three months.

5.4.4. Comparison of the effects of biogeography and genetic on the microbiota diversity

The NPMANOVA analysis was done using Bray-Curtis and UniFrac distances. The global data were explained for 28% (Bray-Curtis) and 45% (UniFrac) by two axes. The results, calculated on the whole matrixes, are given in Table 2.

Table 2: NPMANOVA results for unifactorial and multifactorial analysis made with Bray-Curtis and UniFrac distances. The significance is indicated when p < 0.05 (*), p < 0.01 (**) or p < 0.001 (***).

		Bray-Curtis distance R2 (percent and significance)	*UniFrac distance R2 (percent and significance)*
Unifactorial analysis	Conservatory (6 modalities)	17% ***	27% ***
	DNAmt haplotype (10 modalities)	12% ***	15% ***

	Landscape (3 modalities)	10% ***		17% ***	
	Country (3 modalities)	6% ***		7% ***	
	Climate (3 modalities)	5% ***		5% ***	
	Month (3 modalities)	2% **		2% *	
model 1:	Conservatory	17.5% ***	Global 19.4% ***	27.4% ***	Global 29.6% ***
	+ Month	2.3% ***		3.0% ***	
model 2:	Conservatory	17.5% ***		27.4% ***	
	+ Month	2.3% ***	Global 29.6% ***	3.0% ***	Global 40.9% ***
	+ Month x Conservatory	10,1% ***		11.3% ***	
model 3:	Conservatory	10.5% ***		17.2% ***	
	+ Month	2.3% ***	Global 24.7% ***	3.0% ***	Global 34.9% ***
	+ DNAmt haplotype	5.3% **		5.3% **	
model 4:	Landscape	11.1% ***		18.9% ***	
	+ Climate	5.0% ***	Global 17.7% ***	5.7% ***	Global 26.2% ***
	+ Month	2.4% ***		3.2% **	
model 5:	Landscape	11.1% ***		18.9% ***	
	+ Climate	5.0% ***		5.7% ***	
	+ Month	2.4% ***	Global 26.5% ***	3.2% ***	Global 35.8% ***
	+ Month x Landscape	4,0% ***		4.2% ***	
	+ Month x Climate	4.7% ***		5.1% ***	
model 6:	Landscape	8.5% ***		14.0% ***	
	+ Climate	0.7%	Global 23.5% ***	0.4%	Global 32.2% ***
	+ Month	2.4% ***		3.2% ***	
	+ DNAmt haplotype	5.7% ***		6.0% **	

The unifactorial analyzes show that the conservatory explains most the variability of the bacterial communities (17% for Bray-Curtis, 27% for UniFrac). Only climate, country and month explain less than 10% of variability. However, all the factors are significant (NPMANOVA, p <0.05%, R function Adonis).

Some confounding factors could not be putted together with conservatory: landscape, country and climate. Thus, different models were created, to take them into account separately and to test the interactions. Three models were created with the conservatory, and three with landscape, climate and country instead. By adding the interactions, we better explain the diversity, especially the interaction month * conservatory which explains about 30% of the total variability in the case of Bray-Curtis and 41% for UniFrac. Globally, analyzes realized with UniFrac distances explain the greatest variability regarding to Bray-Curtis.

5.5 Discussion

In this study, many factors were taken into account: conservatory, DNAmt haplotype, landscape, country, climate and time (three consecutive months during summer). The DNAmt genetic data provide a first information on the diversity and purity of honeybee colonies inside the conservatories in which local bee imports are prohibited, as well as transhumance. This DNAmt analysis should be completed by microsatellite DNA analysis to determine the population structure in each conservatory, and also the introgression level by non-native honeybee sub-species for the geographic area studied (Miguel et al., 2007). Thus, the high proportion of C1 haplotypes in the Champart conservatory, and less in the conservatory of Chizé, suggest the presence of apiaries with C lineage around a preservation centers, for instance *A. mellifera ligustica* or *A. mellifera carnica*, honeybees whose import into France are common (Chávez-Galarza et al., 2017; Franck et al., 2000).

In addition, the strong presence of the C lineage in France has been described many times, as well as that of lineage A in the Iberian region (Franck et al., 1998). The haplotype A1, very present in our conservatories of Portugal, is present in the populations of the subspecies *A. m. sicula*, a hybrid bee with a strong presence in Sicily (Franck et al., 2000). This haplotype is also the only one detected in the hives of Zavial, a haplotype very present in the south of the country where there are only two haplotypes: A1 and A2 (Chávez-Galarza et al., 2017). On the contrary, the northern part of the country has much more genetic diversity, a phenomenon that we see in the conservatory of Gimonde (Chávez-Galarza et al., 2017). This low diversity in Zavial is also reflected by the diversity of microbial communities, since this conservatory, as the one of Bilbao, has a relatively low specific diversity compared to others. However, diversity increases when the OTUs that are weakly present (i.e singletons) are taken into account. This phenomenon explains the classification obtained, which includes a large number of genera representing less than 1% of the samples.

Among the genera present in larger quantities, we found bacteria belonging to the core microbiota of social bees, namely *Bartonella*, *Frischella*, *Snodgrassella*, *Gilliamella*, *Bifidobacterium*, *Lactobacillus* and *Sphingomonas* (Graystock et al., 2017; Kwong and Moran, 2016). These bacterial groups are particularly necessary for various functions such as immune defense against pathogens, decomposition of carbohydrates or degradation of pollen walls (Engel et al., 2012). According to our results, the shaping of these populations is related to all

analyzed factors, the main one being their location at the conservatory scale, the DNAmt haplotype, the landscape and the time alone (i.e. month) having a less structuring effect. The environment close to the honeybees seems therefore primordial, a phenomenon reinforced by the weak effect of the climate in our models, and which could be reflected by the food. Indeed, the feeding of bees has a direct effect on their microbiota, either by nectar or by pollen (Anderson et al., 2013; Colman et al., 2012; Saraiva et al., 2015; Yun et al., 2014). Other factors such as altitude or the presence of parasites could also have an impact. Indeed, numerous studies have shown the significant impact of parasites such as *Nosema* (Corby-Harris et al., 2016; Cordes et al., 2012; El Khoury et al., 2018; Hubert et al., 2016), *Varroa destructor* (Hubert et al., 2016, 2017), or other pathogens (Cariveau et al., 2014) on the bacterial communities of the intestinal tract of the honeybee.

Regarding the time (i.e. the month), when the interactions between the factors are taken into account, the month greatly increases the effect of the conservatory alone. These two parameters are therefore closely related despite the chosen period, which takes into account only three consecutive months. July, August and September are important months for the colony's structure, which prepares for the winter by storing reserves and giving birth to the winter bees in September (Clément, 2015). These reserves vary according to the availability of the plants, which vary according to the geographical location but according to the months because of the blooms in particular which vary very quickly. In our data, the evolution of microbial communities is strongly linked to their immediate environment, namely the conservatory and the landscape: the diet could therefore have a great influence on the structuring of the microbiota. These data are of significant health interest, as the displacement of colonies over a short distance or the development of colonies carrying a DNAmt haplotype of interest could favor certain microbiota communities, a key factor of immunity (Fagundes et al., 2012; Koch and Schmid-Hempel, 2011), thereby increasing resistance to certain pathogens.

5.6 Conclusion

Many studies have recently been conducted to understand the effect of different eco-ethological, environmental or genetic factors on the microbiota communities of many animals (Colman et al., 2012; Gupta et al., 2017; Lankau et al., 2012; Ley et al., 2005, 2006; Phillips et

al., 2012; Santo Domingo et al., 1998; Sullam et al., 2012; Yun et al., 2014). These publications agree that factors such as diet and environment have a significant effect and contribute via the microbiota to the health of the host (Hamdi et al., 2011; Hooper et al., 2012; Ley et al., 2005). Regarding our results, although all factors had a significant effect on bacterial communities' structure, the conservatory, beehive DNAmt haplotype and landscape had the biggest effect. Furthermore, the time appeared to increase greatly the effect of the conservatory, bringing these two factors to be the most significant when putted together. Our results bring new elements for social bees, whose microbiota is very close to that of bees (Kwong et al., 2017). They make it possible to better understand the effect of several factors such as genetics, which is not much discussed in the literature, but also time. Our results also suggest that many other factors are involved in the structuring of microbial communities, like diet.

5.7 Data accessibility

5.8 Authors' contributions

D.G.B. and L.G. conducted the whole BEEHOPE project. D.G.B., I.E. and J.L.S. collected the data in Pontaumur and conducted the data analysis for all sites studied. T.S.N., I.M., A.E. and S.H. worked on the BEEHOPE project conception and set up. F.F. worked in 2015 on the project and prepared the data analyses that were then used in 2016. M.A.P set up the apiaries in Zavial and Gimonde with C.J.N. and collected the data monthly in these two conservatories. D.D. and L.G. actively participated to the project by taking care of the bees and collecting the data in Rochefort. A.C., P.G. and X.B. helped with the statistical analysis. I.E. and D.G.B. wrote the manuscript. All authors revised and approved the manuscript.

5.9 Ethic statement

The study was conducted in six conservation centers created in France, Spain and Portugal to preserve the two native M-lineage subspecies. The conservation apiaries were deployed in private lands after obtaining permission from the owners, and to conduct experiments by them on their sites. This field study granted by a European program, BioDIVERSA ERANET, in favor of preservation and protection of biodiversity, did not involve

endangered or protected species. For all locations in France, Spain and Portugal, no specific permission was required, as the apiaries were outside of Natural Parks. We only had to comply for Portugal with the general regulations about distance between apiaries, which is 800 m from the closest apiary (Decree Law n.o 203/2005).

5.10 Competing interests

We have no competing interests.

5.11 Funding

This work was supported in part by the research project BEEHOPE funded by the European call for projects 2013-2014 BiodivERsA / FACCE-JPI from research agencies of France (ANR-14-EBID-0001), Spain (PCIN-2014-090) and Portugal (BiodivERsA / 0002/2014). I. Eouzan is financed by a doctoral grant from the Ministry of National Education, Higher Education and Research (France).

5.12 Acknowledgements

We thank Noel Mallet, Claude Grenier, Jean-Charles Labat, Céline Robert, Paulo Ventura, Miguel Vilas-Boas, Jonathan Gaboulaud, Cécile Ribout, Jean-François Odoux, Egoitz Galarza and Hélène Legout, who all helped us in the BEEHOPE project.

5.13 References

Anderson, M.J. (2001). A new method for non-parametric multivariate analysis of variance: NON-PARAMETRIC MANOVA FOR ECOLOGY. Austral Ecology *26*, 32–46.

Anderson, K.E., Sheehan, T.H., Mott, B.M., Maes, P., Snyder, L., Schwan, M.R., Walton, A., Jones, B.M., and Corby-Harris, V. (2013). Microbial Ecology of the Hive and Pollination Landscape: Bacterial Associates from Floral Nectar, the Alimentary Tract and Stored Food of Honey Bees (Apis mellifera). PLoS ONE *8*, e83125.

Anjum, S.I., Shah, A.H., Aurongzeb, M., Kori, J., Azim, M.K., Ansari, M.J., and Bin, L. (2018). Characterization of gut bacterial flora of Apis mellifera from north-west Pakistan. Saudi Journal of Biological Sciences *25*, 388–392.

Cánovas, F., de la Rúa, P., Serrano, J., and Galián, J. (2011). Microsatellite variability reveals beekeeping influences on Iberian honeybee populations. Apidologie *42*, 235–251.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nature Methods *7*, 335–336.

Cariveau, D.P., Powell, J.E., Koch, H., Winfree, R., and Moran, N.A. (2014). Variation in gut microbial communities and its association with pathogen infection in wild bumble bees (Bombus). The ISME Journal *8*, 2369–2379.

Chávez-Galarza, J., Garnery, L., Henriques, D., Neves, C.J., Loucif-Ayad, W., Jonhston, J.. S., and Pinto, M.. A. (2017). Mitochondrial DNA variation of Apis mellifera iberiensis: further insights from a large-scale study using sequence data of the tRNAleu-cox2 intergenic region. Apidologie *48*, 533–544.

Clément, H. (2015). Le traité rustica de l'apiculture (Rustica).

Colman, D.R., Toolson, E.C., and Takacs-Vesbach, C.D. (2012). Do diet and taxonomy influence insect gut bacterial communities? Molecular Ecology *21*, 5124–5137.

Corby-Harris, V., Snyder, L., Meador, C.A.D., Naldo, R., Mott, B., and Anderson, K.E. (2016). Parasaccharibacter apium, gen. nov., sp. nov., Improves Honey Bee (Hymenoptera: Apidae) Resistance to Nosema. Journal of Economic Entomology *109*, 537–543.

Cordes, N., Huang, W.-F., Strange, J.P., Cameron, S.A., Griswold, T.L., Lozier, J.D., and Solter, L.F. (2012). Interspecific geographic distribution and variation of the pathogens Nosema bombi and Crithidia species in United States bumble bee populations. Journal of Invertebrate Pathology *109*, 209–216.

El Khoury, S., Rousseau, A., Lecoeur, A., Cheaib, B., Bouslama, S., Mercier, P.-L., Demey, V., Castex, M., Giovenazzo, P., and Derome, N. (2018). Deleterious Interaction Between Honeybees (Apis mellifera) and its Microsporidian Intracellular Parasite Nosema ceranae Was Mitigated by Administrating Either Endogenous or Allochthonous Gut Microbiota Strains. Frontiers in Ecology and Evolution *6*.

Engel, P., and Moran, N.A. (2013). The gut microbiota of insects – diversity in structure and function. FEMS Microbiology Reviews *37*, 699–735.

Engel, P., Martinson, V.G., and Moran, N.A. (2012). Functional diversity within the simple gut microbiota of the honey bee. Proceedings of the National Academy of Sciences *109*, 11002–11007.

Fagundes, C.T., Amaral, F.A., Teixeira, A.L., Souza, D.G., and Teixeira, M.M. (2012). Adapting to environmental stresses: the role of the microbiota in controlling innate immunity and behavioral responses: Microbiota and response to ambient stresses. Immunological Reviews *245*, 250–264.

Flint, H.J., Scott, K.P., Louis, P., and Duncan, S.H. (2012). The role of the gut microbiota in nutrition and health. Nature Reviews Gastroenterology & Hepatology *9*, 577–589.

Franck, P., Garnery, L., Solignac, M., and Cornuet, J.-M. (1998). The origin of West European subspecies of honeybees (Apis mellifera): new insights from microsatellite and mitochondrial data. Evolution *52*, 1119–1134.

Franck, P., Garnery, L., Celebrano, G., Solignac, M., and Cornuet, J.-M. (2000). Hybrid origins of honeybees from Italy (Apis mellifera ligustica) and Sicily (A. m. sicula). Molecular Ecology *9*, 907–921.

Garnery, L., Cornuet, J.-M., and Solignac, M. (1992). Evolutionary history of the honey bee Apis mellifera inferred from mitochondrial DNA analysis. Molecular Ecology *1*, 145–154.

Garnery, L., Solignac, M., Celebrano, G., and Cornuet, J.-M. (1993). A simple test using restricted PCR-amplified mitochondrial DNA to study the genetic structure of Apis mellifera L. Experientia 1016–1021.

Graystock, P., Rehan, S.M., and McFrederick, Q.S. (2017). Hunting for healthy microbiomes: determining the core microbiomes of Ceratina, Megalopta, and Apis bees and how they associate with microbes in bee collected pollen. Conservation Genetics *18*, 701–711.

Gupta, V.K., Paul, S., and Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. Frontiers in Microbiology *8*.

Hamdi, C., Balloi, A., Essanaa, J., Crotti, E., Gonella, E., Raddadi, N., Ricci, I., Boudabous, A., Borin, S., Manino, A., et al. (2011). Gut microbiome dysbiosis and honeybee health: Gut microbiome dysbiosis and honeybee health. Journal of Applied Entomology *135*, 524–533.

Han, F., Wallberg, A., and Webster, M.T. (2012). From where did the Western honeybee (Apis mellifera) originate? Ecology and Evolution *2*, 1949–1957.

Hooper, L.V., Littman, D.R., and Macpherson, A.J. (2012). Interactions Between the Microbiota and the Immune System. Science *336*, 1268–1273.

Hroncova, Z., Havlik, J., Killer, J., Doskocil, I., Tyl, J., Kamler, M., Titera, D., Hakl, J., Mrazek, J., Bunesova, V., et al. (2015). Variation in Honey Bee Gut Microbial Diversity Affected by Ontogenetic Stage, Age and Geographic Location. PLOS ONE *10*, e0118707.

Hubert, J., Kamler, M., Nesvorna, M., Ledvinka, O., Kopecky, J., and Erban, T. (2016). Comparison of Varroa destructor and Worker Honeybee Microbiota Within Hives Indicates Shared Bacteria. Microbial Ecology *72*, 448–459.

Hubert, J., Bicianova, M., Ledvinka, O., Kamler, M., Lester, P.J., Nesvorna, M., Kopecky, J., and Erban, T. (2017). Changes in the Bacteriome of Honey Bees Associated with the Parasite Varroa destructor, and Pathogens Nosema and Lotmaria passim. Microbial Ecology *73*, 685–698.

Jones, J.C., Fruciano, C., Hildebrand, F., Al Toufalilia, H., Balfour, N.J., Bork, P., Engel, P., Ratnieks, F.L., and Hughes, W.O. (2018). Gut microbiota composition is associated with environmental landscape in honey bees. Ecology and Evolution *8*, 441–451.

Koch, H., and Schmid-Hempel, P. (2011). Socially transmitted gut microbiota protect bumble bees against an intestinal parasite. Proceedings of the National Academy of Sciences *108*, 19288–19292.

Kwong, W.K., and Moran, N.A. (2016). Gut microbial communities of social bees. Nature Reviews Microbiology *14*, 374–384.

Kwong, W.K., Medina, L.A., Koch, H., Sing, K.-W., Soh, E.J.Y., Ascher, J.S., Jaffé, R., and Moran, N.A. (2017). Dynamic microbiome evolution in social bees. Science Advances *3*, e1600513.

Lankau, E.W., Hong, P.-Y., and Mackie, R.I. (2012). Ecological drift and local exposures drive enteric bacterial community differences within species of Galápagos iguanas: Galapagos iguana enteric community drivers. Molecular Ecology *21*, 1779–1788.

Ley, R.E., Backhed, F., Turnbaugh, P., Lozupone, C.A., Knight, R.D., and Gordon, J.I. (2005). Obesity alters gut microbial ecology. Proceedings of the National Academy of Sciences *102*, 11070–11075.

Ley, R.E., Peterson, D.A., and Gordon, J.I. (2006). Ecological and Evolutionary Forces Shaping Microbial Diversity in the Human Intestine. Cell *124*, 837–848.

Linnenbrink, M., Wang, J., Hardouin, E.A., Künzel, S., Metzler, D., and Baines, J.F. (2013). The role of biogeography in shaping diversity of the intestinal microbiota in house mice. Molecular Ecology *22*, 1904–1916.

Martinson, V.G., Danforth, B.N., Minckley, R.L., Rueppell, O., Tingek, S., and Moran, N.A. (2011). A simple and distinctive microbiota associated with honey bees and bumble bees: the microbiota of honey bees and bumble bees. Molecular Ecology *20*, 619–628.

Mattila, H.R., Rios, D., Walker-Sperling, V.E., Roeselers, G., and Newton, I.L.G. (2012). Characterization of the Active Microbiotas Associated with Honey Bees Reveals Healthier and Broader Communities when Colonies are Genetically Diverse. PLoS ONE *7*, e32962.

Miguel, I., Iriondo, M., Garnery, L., Sheppard, W.S., and Estonba, A. (2007). Gene flow within the M evolutionary lineage of Apis mellifera: role of the Pyrenees, isolation by distance and post-glacial re-colonization routes in the western Europe. Apidologie *38*, 141–155.

Miguel, I., Baylac, M., Iriondo, M., Manzano, C., Garnery, L., and Estonba, A. (2011). Both geometric morphometric and microsatellite data consistently support the differentiation of the M evolutionary branch.pdf. Apidologie *42*, 150–161.

Moran, N.A., Hansen, A.K., Powell, J.E., and Sabree, Z.L. (2012). Distinctive Gut Microbiota of Honey Bees Assessed Using Deep Sampling from Individual Worker Bees. PLoS ONE *7*, e36393.

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B.N., Peeters, M., Hahn, B.H., and Hugenholtz, P. (2010). Evolutionary Relationships of Wild Hominids Recapitulated by Gut Microbial Communities. PLoS Biology *8*, e1000546.

Phillips, C.D., Phelan, G., Dowd, S.E., McDonough, M.M., Ferguson, A.W., Delton Hanson, J., Siles, L., OrdóñEz-Garza, N., San Francisco, M., and Baker, R.J. (2012). Microbiome analysis among bats describes influences of host phylogeny, life history, physiology and geography: microbiome analysis among bats. Molecular Ecology *21*, 2617–2627.

Ruttner, F. (1988). Biogeography and Taxonomy of Honeybees (Berlin: Springer Verlag).

Santo Domingo, J.W., Kaufman, M.G., Klug, M.J., Holben, W.E., Harris, D., and Tiedje, J.M. (1998). Influence of diet on the structure and function of the bacterial hindgut community of crickets. Molecular Ecology *7*, 761–767.

Saraiva, M.A., Zemolin, A.P.P., Franco, J.L., Boldo, J.T., Stefenon, V.M., Triplett, E.W., de Oliveira Camargo, F.A., and Roesch, L.F.W. (2015). Relationship between honeybee nutrition and their microbial communities. Antonie van Leeuwenhoek *107*, 921–933.

Sommer, F., and Bäckhed, F. (2013). The gut microbiota — masters of host development and physiology. Nature Reviews Microbiology *11*, 227–238.

Sullam, K.E., Essinger, S.D., Lozupone, C.A., O'Connor, M.P., Rosen, G.L., Knight, R., Kilham, S.S., and Russell, J.A. (2012). Environmental and ecological factors that shape the gut bacterial communities of fish: a meta-analysis: fish gut bacterial communities. Molecular Ecology *21*, 3363–3378.

Van Treuren, W., Ponnusamy, L., Brinkerhoff, R.J., Gonzalez, A., Parobek, C.M., Juliano, J.J., Andreadis, T.G., Falco, R.C., Ziegler, L.B., Hathaway, N., et al. (2015). Variation in the Microbiota of Ixodes Ticks with Regard to Geography, Species, and Sex. Applied and Environmental Microbiology *81*, 6200–6209.

Yun, J.-H., Roh, S.W., Whon, T.W., Jung, M.-J., Kim, M.-S., Park, D.-S., Yoon, C., Nam, Y.-D., Kim, Y.-J., Choi, J.-H., et al. (2014). Insect Gut Bacterial Diversity Determined by Environmental Habitat, Diet, Developmental Stage, and Phylogeny of Host. Applied and Environmental Microbiology *80*, 5254–5264.

Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics *30*, 614–620.