Characterizing and modeling the evolution of host-microbiota interactions

Benoît Perez-Lamarque

PSL UNIVERSITÉ PARIS

MUSÉUM NAL HIST NATURELLE

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École normale supérieure
Dans le cadre d'une cotutelle avec Muséum national d'Histoire Naturelle

Caractérisation et modélisation de l'évolution des interactions hôtes-microbiotes

Characterizing and modeling the evolution of host-microbiota interactions

Soutenue par
Benoît Perez-Lamarque
Le 29 juin 2021

École doctorale n° 474
Frontières de l'Innovation en Recherche et Éducation

Spécialité
Écologie, évolution et biologie environnementale

ENS

Composition du jury :

Philippe VANDENKOORNHUYSE PR, Université de Rennes	*Président du jury*
Frédéric DELSUC DR CNRS, Université de Montpellier	*Rapporteur*
Mélanie ROY MC, Université de Toulouse	*Rapporteure*
Damien DE VIENNE CR CNRS, Université de Lyon	*Examinateur*
Élisa THÉBAULT CR CNRS, Sorbonne Université	*Examinatrice*
Hélène MORLON DR CNRS, ENS	*Directrice de thèse*
Marc-André SELOSSE PR, MNHN	*Co-directeur de thèse*
Florent MARTOS MC, MNHN	*Co-encadrant de thèse*

Acknowledgments

Je souhaite tout d'abord remercier les membres de mon jury de thèse, Frédéric Delsuc, Mélanie Roy, Damien de Vienne, Élisa Thébault et Philippe Vandenkoornhuyse d'avoir accepté de relire et d'évaluer ce manuscrit de thèse. Un grand merci aussi à Marianne Elias et Damien de Vienne d'avoir fait partie de mon comité de thèse et de m'avoir suivi au cours de ces trois années.

Je tiens ensuite à remercier chaleureusement mes directeurs de thèse qui m'ont accompagné avec attention et bienveillance lors de ces dernières années tout en me laissant beaucoup de libertés quant à mes projets. Hélène, merci de m'avoir conseillé et guidé depuis mon arrivée à l'ENS en 2014. Merci de m'avoir permis de découvrir le monde de la modélisation à compter de mon stage de M2. Merci pour ton encadrement sans faille, ton soutien et ta rigueur de travail qui m'ont beaucoup inspiré pendant ma thèse. Marc-André, merci pour l'émerveillement que tu as suscité au cours de mes années d'études et m'a donné l'envie de me plonger dans le monde des symbioses et du microbien. Merci pour les discussions enrichissantes, ainsi que pour les nombreuses opportunités de projets et collaborations, de Tartu à Strasbourg en passant par la Pologne. Puis enfin Florent, merci de ton encadrement attentionné et tes nombreux conseils. Merci de m'avoir permis de réaliser du terrain et du labo pendant ma thèse et de m'avoir ainsi fait un peu sortir de mon écran d'ordinateur.

Ma thèse n'aurait pas été la même sans les deux superbes équipes dans lesquelles j'ai été intégré. À l'IBENS comme à l'ISYEB, un grand merci aux autres doctorants, pour tous ces moments partagés, en particulier Sophia, Rémi, Odile, Marc, Félix, Laure, Tomas, Jakub, Liam, et Géromine, aux post-doctorants, Ana, Amélia, Carmelo, Guilhem, Ignacio, Isaac, Julien, et Leandro, ainsi qu'aux stagiaires qui ont passé un bout de chemin dans les labos, Rémy, Benoît, Loréna, Bastien et Thomas. Merci aussi aux autres membres permanents de l'ISYEB, notamment Chantal et Céline pour vos nombreux conseils, ainsi qu'à nos collaborateurs scientifiques, en particulier Maarja, Christine, Richard, Naomi, Henrik et Rosie. Je tiens aussi à remercier les enseignants du département de Biologie de l'ENS, de m'avoir confié des missions de monitorat pendant ces trois années de thèse, qui ont toutes été si enrichissantes. Je souhaite également remercier les équipes administratives et techniques, en particulier le service informatique de l'IBENS en charge du cluster de calcul et la plateforme de systématique moléculaire du Muséum.

J'en profite aussi pour remercier tous les enseignants et chercheurs que j'ai eu la chance de croiser au cours de ma formation scolaire puis universitaire qui m'ont donné le goût pour la recherche et les sciences fondamentales.

Merci ensuite à mes proches parisiens, toulousains, limouxins ou d'ailleurs, pour ces soirées, brunchs, footings, week-ends, randonnées, vacances ou autres qui m'ont per-

mis de m'évader pendant cette thèse, notamment au cours de la dernière année qui s'est déroulée dans ces conditions si spéciales. Merci en particulier Clément, pour m'avoir supporté au quotidien de manière indéfectible.

Et finalement, je souhaite remercier ma famille, ma sœur et mes parents. De mon premier microscope à 10 ans à mes dernières années d'étude, un grand merci pour m'avoir apporté autant de soutien et d'encouragements, sans lesquels je n'aurais jamais pu effectuer ces études et autant m'y épanouir.

Contents

Chapter 0.

Introduction:

Since the pioneer experiments of Louis Pasteur and Robert Koch in the 19th century, the idea that microbes are responsible for many human and animal diseases has been largely popularized (Koch, 1876; Pasteur, 1880). In addition, the works of Anton de Bary clarified the causal roles of microbes in major crop diseases (De Bary, 1853). A war against infectious diseases caused by the plethora of animal- or plant-associated microbes therefore started (Duffy & Dowling, 1979). Human infectious diseases have been largely reduced thanks to the improvement of hygiene standards, the discovery of new antimicrobial compounds (Fleming, 1929), and the development of prophylactic and therapeutic strategies (Pasteur, 1885a; D'Hérelle, 1917). Similarly, in agriculture, the use of synthetic antimicrobial pesticides have significantly contributed to the regulation of plant diseases (Howard, 1996). For more than a century, bacteria, fungi, and other types of microbes associated with animals or plants have thus widely been considered pathogenic (Smith, 2012).

Nevertheless, in parallel with the experiments of Pasteur, Koch, and de Bary, microbes were observed in abundance in healthy animals and plants: Ernst Hallier reported bacteria in the feces of healthy humans (Hallier, 1869) and Albert Bernhard Frank observed filamentous fungi colonizing the roots of trees (Frank, 1885), suggesting that some microbes might not be harmful (Pasteur, 1885b). These observations under microscopes indeed were rapidly reinforced by experiments showing that these microbes were beneficial to their hosts: at the eve of the 19th century, Theodor Escherich showed the beneficial role of bacteria in the digestion of food in children (Escherich, 1886) and Frank demonstrated that root-associated fungi improved the growth of *Pinus sylvestris* (Frank, 1892). In 1905, Henry Tissier administered bacteria to humans and successfully cured a gastrointestinal disease (Tissier, 1905), therefore being the first therapeutic-aware use of probiotic bacteria. Probiotic bacteria were increasingly used in the following decades through the consumption of fermented milk (Farré-Maduell & Casals-Pascual, 2019).

However, a conceptual shift occurred only at the end of the second part of the 20th century thanks to the advances in molecular biology and DNA sequencing technics (Heather & Chain, 2016). These methods have offered a precise and large-scale characteriza-

tion of the 'invisible' microbial communities hosted by healthy animals and plants. Combined with experimental works, they have shed light on the diversity of these microbes and improved our understanding of their functioning and their consequences on host phenotypes. It is now widely recognized that these microbial communities are normal constituents of healthy hosts, plants and animals, and are mostly beneficial to them (Margulis, 1970; Selosse *et al.*, 2004; Berendsen *et al.*, 2012; McFall-Ngai *et al.*, 2013).

The accumulation of host-associated microbiota studies has helped to better understand their functioning and ecology, and they now allow us to start answering questions about their evolution. Hereafter, we introduce the ecology of host-associated microbiota (section 1), discuss the potential mechanisms driving their evolution (section 2), and present the available methodological tools to study them (section 3). We mainly focus our work on two of the most studied host-microbiota interactions: the bacterial gut microbiota of animals and the root fungal microbiota of plants.

Contents of Chapter 0

1. The ecology of host-microbiota interactions

1.1. The ubiquity and diversity of host-associated microbiota

Most organisms interact with individuals from other species. This can happen through brief interactions or through long-term and intimate associations, referred to as **symbiosis** (De Bary, 1879). These biotic interactions can be beneficial for both interacting species (mutualism), harmful for both (competition), or beneficial for one and harmful for the other (antagonism). Less frequently, the outcome of the interaction is almost neutral for one species but beneficial (resp. harmful) for the other, and we referred to it as commensalism (resp. amensalism). Though these main categories of biotic interactions have been mainly proposed for interactions between macroorganisms (*e.g.* animals and plants), they also apply for animal-microbe or plant-microbe interactions (Figure 0.1.1). However, rather than forming pairs of interacting species, a single plant or animal macroorganism often hosts diverse microbial communities composed of a multitude of microorganisms, referred to as the **microbiota**. The word '**microbiome**' then refers to the assemblage of microbes as well as their surrounding environmental conditions (Marchesi & Ravel, 2015).

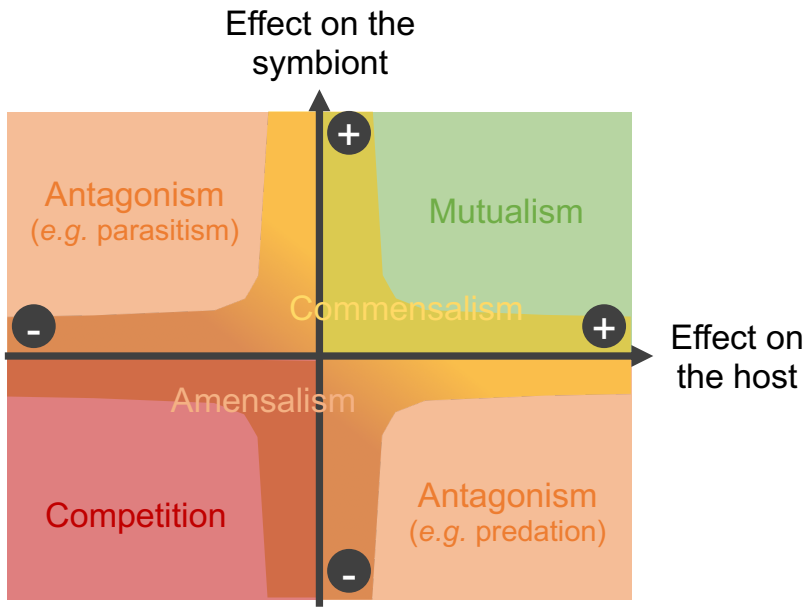


Figure 0.1.1: Classification of the different natures of host-symbiont interspecific interactions based on the (positive or negative) effects on the interactions on the host (x-axis) or the symbiont (y-axis).

Animals and plants are indeed generally colonized by species-rich and complex microbiota. They tend to associate with various bacteria, archaea, viruses, fungi, and other eukaryotic microbes. The composition of these 'invisible' communities has been characterized thanks to microscopic observations and the development of molecular biology. In particular, metabarcoding technics, which consist in amplifying and sequencing a targeted marker gene of the microbial community, like the small subunit ribosomal ri-

bonucleic acid (SSU rRNA) gene or the internal transcribed spacer (ITS), enable precise identification of the microbes present in a given microbiota (see section 3.1).

1.1.1. The gut microbiota of animals

The **animal kingdom** (Metazoa) is composed of heterotrophic multicellular organisms getting their energy and their organic matter from the consumption and transformation of preexisting organic matter. The Bilaterians, including most animal species (but sponges and cnidarians), share a common body organization: they present a bilateral symmetry, an anterior-posterior axis, and are traversed by a digestive tract (the gut) with a separate mouth and anus, where food from the environment is digested and assimilated by the organism. In addition, animal organisms are protected from the external environment by tissues that form a continuous, mono- or pluricellular layer, like the skin or the intestinal epithelium, and that prevents the entrance of most microbes into the organism. Animal-associated microbes are thus mainly colonizing the surface of skin or gut epithelium. Guts have therefore a dual function: they (i) digest and assimilate nutrients from the food, and (ii) allow the development of microbes, while preventing their entrance inside the organism.

Humans are colonized by as many bacteria as they have cells, and 99% of these bacteria are present in the large intestine, or colon (Sender *et al.*, 2016). Per gram of intestinal content, there are on average 10^{12} bacterial cells that belong to 300 to 1,000 different species (Guarner & Malagelada, 2003). The gut microbiota of humans, like most mammal species, are dominated by mostly anaerobic bacteria from the phyla Firmicutes and Bacteroidetes, and the phyla Actinobacteria, Proteobacteria, Verrucomicrobia, and Tenericutes to a lesser extent (Guarner & Malagelada, 2003; Ley *et al.*, 2008; Delsuc *et al.*, 2014; Nishida & Ochman, 2018; Figure 0.1.2). Most of these microbial lineages are specific to a mammalian order (Song *et al.*, 2020) and are not found in other environments. Nishida & Ochman (2018) found a positive association between microbial diversity and mammalian gut capacity, as the larger ones represent a larger ecological niche. The composition of the mammalian microbiota significantly varies according to their diet, in particular between herbivorous, omnivorous, and carnivorous species (Muegge *et al.*, 2011). Microbiota of herbivorous mammals are in particular more diverse, with significant differences between foregut fermenters (including ruminants) and hindgut fermenters (*e.g.* rabbits or horses; Ley *et al.*, 2008).

Comparative studies of thousands of human microbiota have clarified the environmental factors influencing the composition of the gut microbiota: variations are mainly associated with age (Yatsunenko *et al.*, 2012), disease status (Guarner & Malagelada, 2003), geography (Suzuki & Worobey, 2014), and cultural traditions including diets (David *et al.*, 2014). In particular, Arumugam *et al.* (2011) proposed the existence of three different types of human gut microbiota based on the relative abundances of the main phyla influenced by long-term diet. The three enterotypes are respectively enriched in *Bac-*

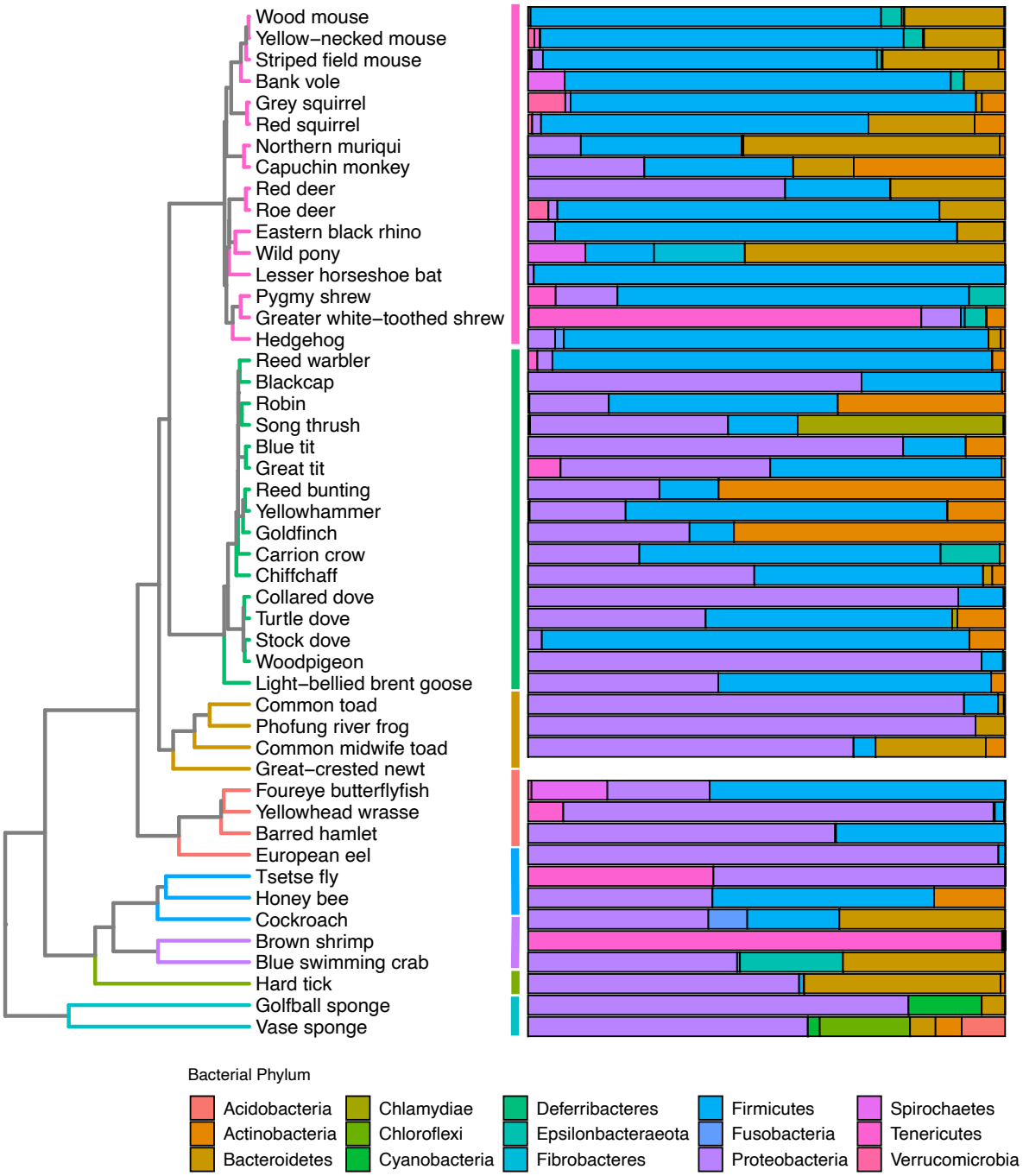


Figure 0.1.2: Example of the composition of the bacterial gut microbiota of different animal species belonging to different classes (including mammals, birds, insects, and spiders). For each host species, the bar plot indicates the relative proportion of main bacterial phyla present in gut microbiota. The figure is derived from Harrison *et al.* (2020).

teroides (Bacteroidetes), *Prevotella* (Bacteroidetes), and *Ruminococcus* (Firmicutes); the former being associated with animal-based diet whereas the two latter with plant-based diets. Although the presence of similar enterotypes has been observed in chimpanzees (Moeller *et al.*, 2012), the discrete delineation into enterotypes have been questioned, and more recent studies rather suggest the existence of continuous variations rather than clear types (Knights *et al.*, 2014; Costea *et al.*, 2017).

In contrast, the microbiota of other vertebrates, such as fishes, reptiles, and birds, tend to be mainly dominated by Proteobacteria (Hacquard *et al.*, 2015; Hird *et al.*, 2015; Song *et al.*, 2020; Figure 0.1.2). Similarly, because they have a microbiota relatively enriched in Proteobacteria, bats are more similar to birds than to any non-flying mammals (Song *et al.*, 2020). Both bird- and bat-associated microbes also have a limited diversity, harbor little host specificity, and do not significantly vary according to host diet (Nishida & Ochman, 2018; Song *et al.*, 2020).

Similar bacterial phyla colonize the midgut of arthropods, including insects and spiders, but their relative composition is much more heterogeneous (Colman *et al.*, 2012; Engel & Moran, 2013; Yun *et al.*, 2014; Sanders *et al.*, 2014; Hu *et al.*, 2019; Figure 0.1.2). The small arthropods are generally colonized by far less bacterial taxa than vertebrate hosts and arthropod-associated microbes present a large range of specificity: for instance, symbiotic bacteria associated with soil-feeding termites are extremely specific (Brune, 2014), whereas bacteria found in some ant species seem to be very transient and non-specific (Russell *et al.*, 2017), and the caterpillar *Manduca sexta* even lacks a gut microbiota (Hammer *et al.*, 2017, 2019).

Besides bacteria, fungi and archaea are also abundant components of the animal gut microbiota (Harrison *et al.*, 2020; Youngblut *et al.*, 2020). Animal gut microbiota seem to be mainly composed of the fungal phyla Ascomycota and Basidiomycota, except for a few exceptions, like some herbivorous mammals that are mainly colonized by the phylum Neocallimastigomycota (Harrison *et al.*, 2020). Concerning archaea, vertebrate gut microbiota appear to be vastly dominated by the methanogenetic lineages from the phylum Euryarchaeota (Youngblut *et al.*, 2020). However, studies specifically characterizing fungal and archaeal communities currently remain very scarce: for instance, >90% archaeal lineages are generally not detected by standard metabarcoding approaches targeting the 16S SSU rRNA gene.

1.1.2. The root microbiota of plants

Conversely, **land plants** (Embryophyta) are autotrophic multicellular organisms: thanks to solar energy, they get their energy and convert inorganic matter (water and carbon dioxide) into organic matter. This reaction (photosynthesis) mediated by chlorophyll mostly occurs in their green leaves, whereas roots ensure the supply of water and mineral nutrients (*e.g.* phosphate, nitrogen, or potassium). Contrary to animals, plants have cell walls allowing intercellular spaces, and are therefore internally colonized by a plethora of microbes including bacteria and filamentous fungi, in the tissues of their leaves, reproductive tissues, and roots.

In particular, observations under microscope often report fungal colonizations. This

includes fungi inducing particular structural changes in the roots of healthy plants, referred to as **mycorrhiza** (Smith & Read, 2008). Such mycorrhizas are present in more than 90% of plant species (Brundrett & Tedersoo, 2018) and in almost all ecosystems, including agrosystems (Read, 1991). Traditionally, mycorrhizas have been classified into 4 main categories based on their morphology and on the taxonomy of the plants and fungi involved (Smith & Read, 2008; van der Heijden *et al.*, 2015; Brundrett & Tedersoo, 2018).

First, **arbuscular mycorrhiza** is the most recurrent association between land plants and fungi from the Glomeromycotina subphylum (Mucoromycota phylum; Davison *et al.*, 2015). Glomeromycotina (also called arbuscular mycorrhizal fungi) form microscopic aseptate filaments that penetrate plant roots and cell walls to form arbuscular structures that invaginate the plant cell membrane. Glomeromycotina are obligate symbionts that cannot survive or be cultivated without plants (Bago & Bécard, 2002; Figure 0.1.3). They reproduce by producing microscopic spores that are used to delineate species. However, less than 400 Glomeromycotina species have been morphologically described (Stefani *et al.*, 2020), and most of the Glomeromycotina are rather identified using metabarcoding technics: thanks to this, between 300 and 2,700 Glomeromycotina 'species' are estimated (Öpik *et al.*, 2014; van der Heijden *et al.*, 2015; Stefani *et al.*, 2020; but see section 3.1). Given than Glomeromycotina associate with >72% of the >200,000 land plant species, they are mostly highly generalists, associating with a large range of plant species. However, despite this low host specificity, plant-Glomeromycotina interactions overall reflect non-random assemblages (Vandenkoornhuyse *et al.*, 2003; Sepp *et al.*, 2019; Kokkoris *et al.*, 2020). They have been documented in the roots of most plant lineages, including angiosperms, gymnosperms, ferns, lycopods, and in the thalli of liverworts; but they lack in hornwort thalli and in the roots of some lineages of flowering plants where they are often replaced by other fungi (*e.g.* the orchids - see below; Hoysted *et al.*, 2018). In addition, they are particularly abundant in the tropics and in temperate grasslands (Read, 1991).

Second, **ectomycorrhizas** are found in ∼2% of the plant species (Brundrett & Tedersoo, 2018). Ectomycorrhizal fungi form a mantel at the surface of the root and penetrate between root cells, but without crossing the cell wall (Hartig net; Figure 0.1.3). These fungal lineages mainly belong to Pezizomycetes (Ascomycota) and Agaricomycetes (Basidiomycota), which form macroscopic fruiting bodies, like the black truffle (*Tuber melanosporum*) or the fly agaric (*Amanita muscaria*). Like arbuscular mycorrhizal fungi, many ectomycorrhizal fungi are obligate symbionts (Miyauchi *et al.*, 2020). Ectomycorrhizas are particularly found in the roots of trees or shrubs under temperate latitudes (Read, 1991), but new ectomycorrhizal plants and fungi are also being found in tropical regions (Roy *et al.*, 2009).

Lastly, **orchid mycorrhizas** and **ericoid mycorrhizas** have been described in plant species from the families Orchidaceae and Ericaceae, respectively (Brundrett & Teder-

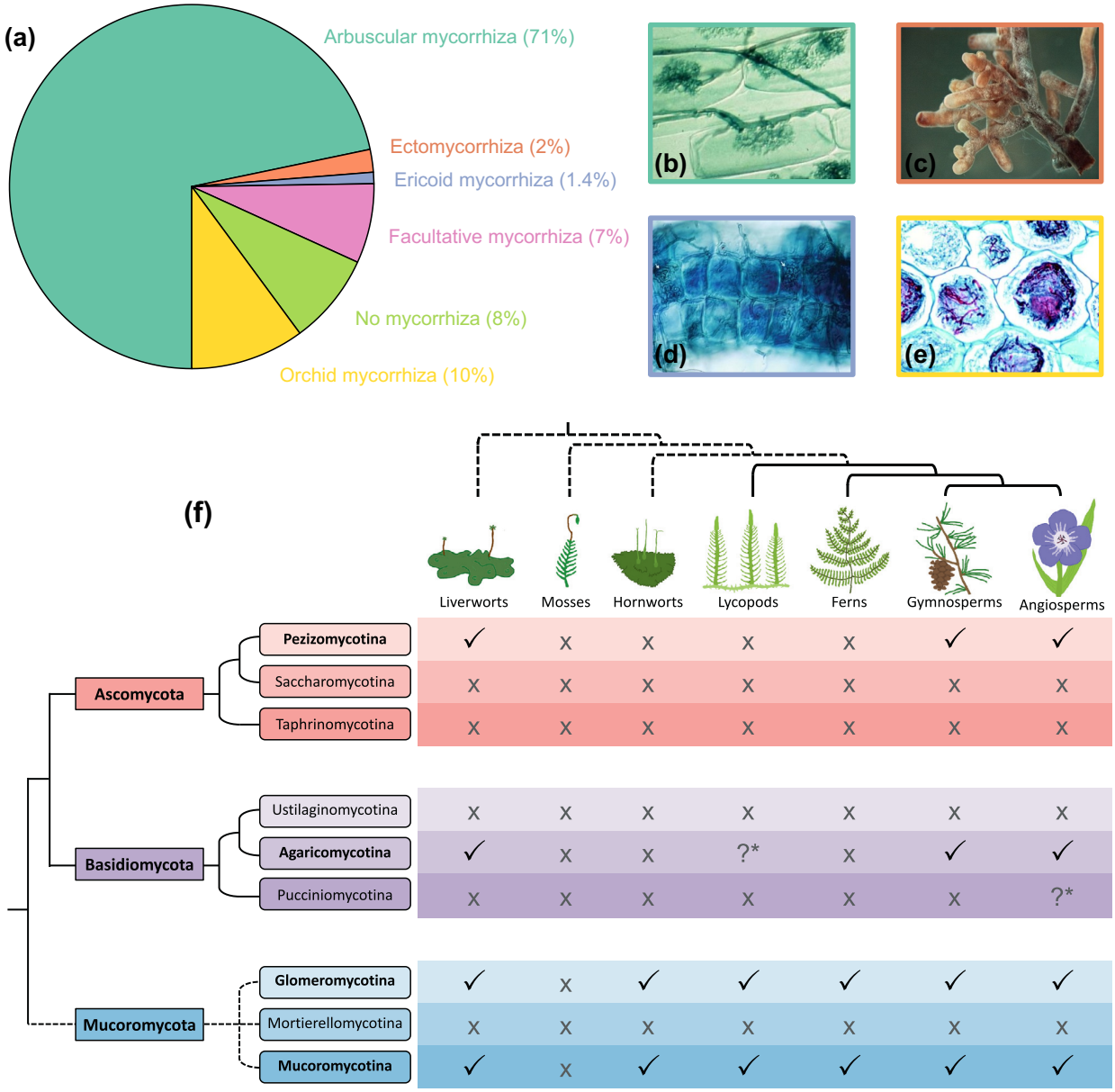


Figure 0.1.3: Proportion of the different types of mycorrhizas among land plants. (a) The pie chart indicates the proportion of land plants having each type of mycorrhiza according to Brundrett & Tedersoo (2018). (b-e) Microscopic photos representing the morphological specificity of each type of mycorrhiza: (b) arbuscular mycorrhiza (Larry Peterson): filamentous fungi penetrate the root and form 'intracellular' arbuscular structures; (c) ectomycorrhizal fungi (Ellen Larsson): ectomycorrhizal fungi forming a dense mantel around the root tip; (d) ericoid mycorrhiza (Jesse Sadowsky): dark septate fungi colonize the root and form intracellular hyphal coils; and (e) orchid mycorrhiza (Nancy Collins Johnson): orchid mycorrhizal fungi similarly form hyphal coils within plant cells. (f) Phylogenetic trees of the main clades of fungi (right) and land plants (top – dashed lines indicate uncertain phylogenetic relationships) with indications concerning the mycorrhizal status of each pair of clades. Question marks represent uncertain mycorrhizal status. Figure modified from Hoysted *et al.* (2018).

soo, 2018). In both symbioses, fungi penetrate the plant cell walls, evaginate the cell membranes and form large pelotons (coils). Orchid mycorrhizas involve different fungal lineages from Basidiomycota, especially from the families Tulasnellaceae, Ceratobasidi-

aceae, and Serendipitaceae (Dearnaley *et al.*, 2012), whereas ericoid mycorrhizas involve diverse Ascomycota and Basidiomycota lineages, including from the order Helotiales and the family Serendipitaceae (Selosse *et al.*, 2009; Toju *et al.*, 2016). Contrary to other mycorrhizal fungi, these lineages are often facultative symbionts, that can live without plants thanks to saprophytic lifestyle (Dearnaley *et al.*, 2012).

Besides these typical mycorrhizas, recent molecular advances in metabarcoding technics have enabled the characterizations of new undocumented associations: for instance, Endogonales fungi from the Mucoromycotina subphylum (a sister clade of the Glomeromycotina) were recently found to colonize many plant species from angiosperms, gymnosperms, lycopods, hornworts, and liverworts, and can even form arbuscular-like mycorrhizal structures (Bidartondo *et al.*, 2011; Desirò *et al.*, 2013; Rimington *et al.*, 2015).

Although most plants always host mycorrhizal fungi, 8% of the plant species do not have mycorrhizas or present very limited rudimentary mycorrhizal fungal colonizations and 7% of the plant species are only facultatively mycorrhizal (Cosme *et al.*, 2018; Brundrett & Tedersoo, 2018). Non-mycorrhizal plants correspond in particular to parasitic or carnivorous plants or plants that have proteoid roots (cluster roots; Brundrett & Tedersoo, 2018). Compared to obligate mycorrhizal plants, facultative mycorrhizal plants tend to present wider geographic ranges and broader ecological niches (Hempel *et al.*, 2013), and mycorrhizal colonization often depends on environmental conditions: for instance, some plant species stop hosting mycorrhizal fungi when growing on soil with high phosphorus concentrations (Thomson *et al.*, 1986; Mujica *et al.*, 2020).

In addition, a plethora of fungi also colonizes plant roots less densely, without forming any specialized morphological structures such as the mycorrhiza (Wilson, 1995). These fungi, referred to as **endophytes**, colonize healthy plants without impacting or damaging plant tissues (Selosse *et al.*, 2018). Among them, many fungi, like *Tuber* or *Sebacina*, that form mycorrhizas in some plant species can colonize as endophytes the roots of other surrounding plant species (Selosse *et al.*, 2009; Schneider-Maunoury *et al.*, 2020). These endophytes also frequently colonize non-mycorrhizal plants (Almario *et al.*, 2017).

Finally, plants also host a large diversity of bacteria in their tissues and in their surrounding soil (the rhizosphere; Berendsen *et al.*, 2012). Root-associated bacteria are particularly well studied in model organisms such as *Arabidopsis thaliana* and in crops that do not present mycorrhizas (Hacquard *et al.*, 2015). These bacterial microbiota are particularly enriched in Proteobacteria, Actinobacteria, Acidobacteria, and Bacteroidetes (Yeoh *et al.*, 2017; Vannier *et al.*, 2018; Benucci *et al.*, 2020). In addition, some plant lineages, like the Fabaceae family, form root nodules that host bacteria, referred to as rhizobia (Young & Haukka, 1996). However, plant-associated bacterial and archaeal communities remain overall less frequently characterized than fungal ones. The same applies for viruses that are abundant but rarely characterized in plant microbiomes (Roossinck, 2019).

1.1.3. Comparisons between animal and plant microbiota

Contrary to animals that have internalized their gut microbiota in a more controlled environment (in particular, anaerobic), for plants, the hyphae of mycorrhizal fungi freely explore the surrounding soil. The same fungal individual can then form mycorrhizal associations with different plants: as a consequence, plants are usually interconnected by mycorrhizal mycelial networks, referred to as '**wood-wide-webs**' (Simard *et al.*, 1997). A mycorrhizal network therefore consists in fungus and plant individuals interacting with multiple partners, whereas a gut microbiota system is one host individual with multiple microbial partners.

Only a small fraction of the vast diversity of bacteria and fungi is regularly found in animal and plant microbiota. For instance, less than 10 of the >90 bacterial phyla are frequently present in the gut microbiota of healthy animals (Hug *et al.*, 2016) and only 0.5 to 10% of the total number of fungal species are estimated to be mycorrhizal or endophytic (Taylor *et al.*, 2014), belonging to a few classes only (Hoysted *et al.*, 2018). Host-associated microbiota are therefore non-random assemblages of the microbes found in the biosphere and are rather dominated numerically by a few microbial clades that can colonize the host niche, which suggests the existence of widespread host filtering.

Host-associated microbiota therefore present very diverse compositions, with a mix of resident microbes (forming durable symbiotic associations with their host) and more transient ones. Usually, the microbiota of host individuals from the same species tend to be more similar than the microbiota of other host species, and more similar than expected by chance, if microbiota were randomly assembled from microbes found in the host's environment. Although some hosts have developed a high specificity toward a very limited number of microbes, most animals and plants host species-rich microbial community: for a given host species, we can separate the core microbes (that are present in all the microbiomes) or the flexible microbes (that are only present in a fraction of the microbiomes; Shapira, 2016). Besides host identity, the composition of host-associated microbiota depends on the availability of the microbes and may also be modulated by environmental factors, a phenomenon referred to as ecological specificity (Molina *et al.*, 1992). Regarding the associated microbes, they can present a range of host specificity: while some microbes can live freely in the host's environment, other microbes are restricted to the host niche. Among them, some microbes have been uniquely found associated with a particular host (*i.e.* specialist microbes), whereas others can be associated with multiple host lineages (*i.e.* generalist microbes). These multiple-partners interactions involving a large number of host and microbe species are often represented using bipartite networks (see section 3.3).

1.2. The functions of host-microbiota interactions:

The ubiquity of host-associated microbiota has thus rapidly questioned the functions that such interactions can have. Following by a few decades the works of Escherich and Franck on the beneficial roles of gut bacteria and mycorrhizal fungi to humans and plants respectively, many studies have investigated the functions of host-associated microbiota (Figure 0.1.4). These studies generally compare the fitness and phenotypes of hosts colonized by normal microbiota with hosts whose microbiota have been experimentally eliminated (Smith *et al.*, 2007). Such germ-free hosts (also referred to as axenic) are obtained by applying antimicrobial compounds or by growth in sterile environments.

1.2.1. The roles of microbiota in host nutrition, protection, and development

First, microbiota improve host growth thanks to their effects on its **nutrition**. For instance, 80%-100% of the mineral resources needed by plants can come from their mycorrhizal fungi (Li *et al.*, 1991). Indeed, mycorrhizal fungi form dense filamentous hyphal networks in the soil where they efficiently gather water and mineral nutrients, including poorly soluble ones like phosphorous, that they trade with their associated plants through the mycorrhiza in exchange for organic matter produced by plant photosynthesis (Smith & Read, 2008). In some cases, mycorrhizal fungi also deliver organic matter to the plants: for instance, orchid seeds lack nutritional reserve and thus rely on their mycorrhizal fungi to get organic matter (Merckx, 2013). In addition, given that a mycorrhizal fungus often interacts simultaneously with multiple plants, nutrients can thus transit between plants: for instance, transfers of organic matter have frequently been observed from mature plants to juvenile ones (Simard *et al.*, 1997; Selosse *et al.*, 2006).

In animals, and in mammals in particular, gut microbiota actively contribute to host digestion. For instance, many animal lineages do not have the enzymes to digest the dietary fibers (polysaccharides) that particularly abound in plant-based diets: they therefore rely on a plethora of gut microbes, such as the genera *Ruminococus* or *Prevotella*, to ferment fibers into short-chain fatty acids, which are then absorbed by the host (Moran *et al.*, 2019): for instance, more than 80% of the maintenance energy of ruminants is ensured by these short-chain fatty acids produced by their microbial symbionts (Bergman 1990). Conversely, animal-based diets tend to decrease the proportion of fibrolytic bacteria and increase that of bile-tolerant bacteria, such as *Bacteroides*, useful for protein digestion (Arumugam *et al.*, 2011; David *et al.*, 2014). The presence of different enterotypes according to the herbivorous *versus* carnivorous diets in humans and mammals thus reflects a trade-off between the gut bacteria responsible for carbohydrate and protein fermentations (David *et al.*, 2014). Gut bacteria are also assisted by fungi in the digestion of carbohydrates, like the phylum Neocallimastigomycota that is abundant in mammalian guts, especially herbivores (Akin & Borneman, 1990).

Besides helping the mineral assimilation or the digestion of organic matter, host-

associated microbiota can also synthesize new compounds. In mammals, many essential amino acids or vitamins are synthesized by gut microbes (Smith *et al.*, 2007). In addition, bacteria associated with some plant lineages, like Fabaceae, or some animals, like wood-feeding termites, ensure the fixation of inorganic nitrogen and provide a reliable source of organic matter containing nitrogen (*e.g.* amino acids; Kneip *et al.*, 2007). Microbiota can thus ensure a role of complementation of the diets, which is particularly useful when the hosts rely on diets with unbalanced nutrient uptakes, especially in insects with highly specialized trophic type (Engel & Moran, 2013).

Second, microbiota enhance the **protection** of their host from abiotic or biotic stresses (Ubeda *et al.*, 2017; Begum *et al.*, 2019). Microbes can modulate the effect of the environment on the host: for instance, gut microbes can neutralize dietary toxins (Berasategui *et al.*, 2017; Moeller & Sanders, 2020) and mycorrhizal fungi can help the plants to tolerate high calcium concentration in the soil (Lapeyrie, 1990). In addition, microbes can protect their hosts from pathogens. First, by occupying the host niches, they limit the establishment of invading pathogens (Bauer *et al.*, 2018). Second, they can produce antimicrobial compounds that prevent the growth of pathogens (Ubeda *et al.*, 2017; Begum *et al.*, 2019).

Third, host-associated microbiota play a role in the **development** of the hosts. Animals and plants 'dwell in a microbial world' (McFall-Ngai *et al.*, 2013), and for the first

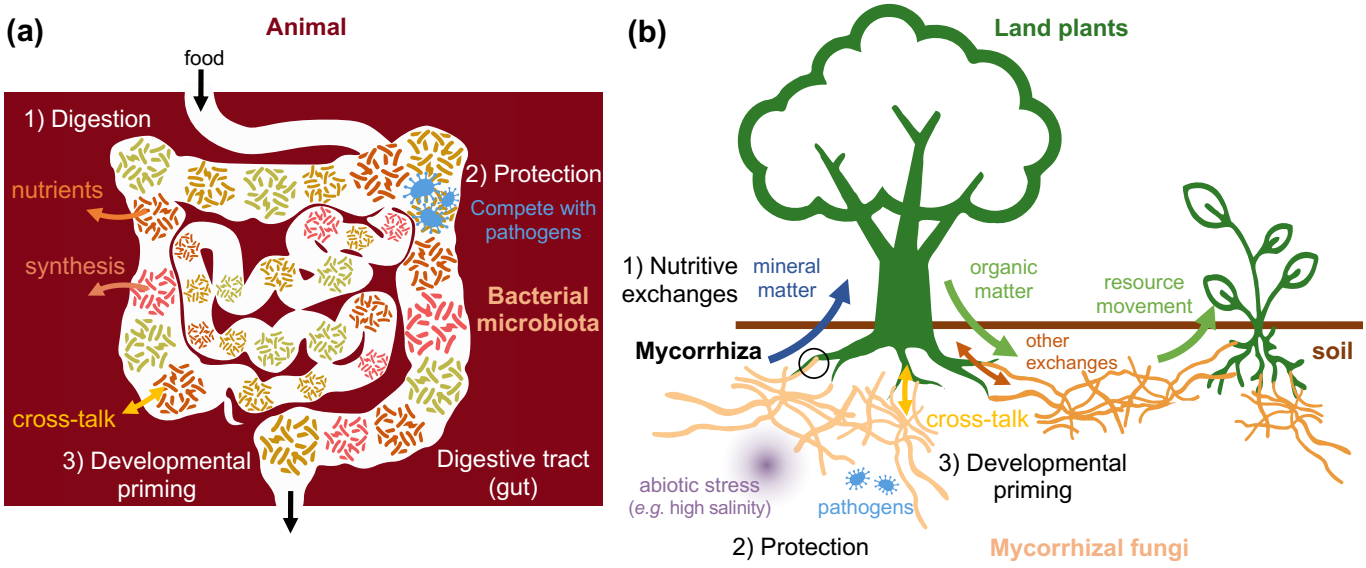


Figure 0.1.4: Functions of host-microbiota interactions. Examples of the bacterial microbiota of animal guts (a) and the fungal microbiota of plant roots (b). Both microbiota participate in 1) host nutrition (through nutritive exchanges or by helping food digestion, or by synthesizing new molecules), 2) host protection (by competing with pathogens or by reducing abiotic stresses), and 3) host development (by priming the maturation of the host immunity). Note that while gut bacterial microbiota are internalized within the animal organism, mycorrhizal fungi are partially (if not largely) external symbionts exploring the soil, and can interact with multiple plants, forming common mycelial networks (wood-wide-webs), enabling resource movement between plants and communication.

stages of their lives, they are in contact with these microbes. The correct development and maturation of the immune system of animals and plants therefore rely on microbes to be initiated: axenic plants or animals usually badly performed against pathogens compared with individuals that grew with a normal microbiota (McFall-Ngai, 2002; Conrath *et al.*, 2006; Smith *et al.*, 2007). The adaptive immune systems of vertebrates are continuously learning beneficial microbes from non-beneficial ones, that therefore maintain homeostasis in the gut microbiota (Hooper *et al.*, 2012). Besides immunity priming, the brain development seems to be modulated by gut microbes, which can therefore influence animal behavior (Heijtz *et al.*, 2011) and shifts in microbiota composition can be used as cues for phenotypic changes, like cold acclimation (Moeller & Sanders, 2020). The mycorrhizal network connecting plants can also enable communication between plant individuals through the emission of warning signals that stimulate plant response against herbivory (Babikova *et al.*, 2013).

1.2.2. The advantages of relying on microbes

Microbiota play thus essential passive or active roles in host functioning. By improving nutrition and protection, microbiota can thus enable colonization of new niches and thus expand host ecological ranges (Moran *et al.*, 2019; Suzuki & Ley, 2020). For instance, Ericaceae plants are particularly successful in acidic and nutrient-poor soils thanks to the buffering effect of their mycorrhizal fungi (Shaw *et al.*, 1990). Similarly, the adaptation of human populations to cold climates may be helped by the enrichment in their gut microbiota of bacteria with efficient energy extraction favoring fat storage (Suzuki & Worobey, 2014). Because microbes are rapidly evolving, especially thanks to frequent horizontal gene transfers, host-associated microbes can provide to the host a very rapid way of adaptation: for instance, *Bacteroides* in the gut microbiota of Japanese humans have acquired enzymes that enable them to ferment algal fibers frequently consumed by these individuals (Hehemann *et al.*, 2010).

In addition, the fact that host-microbiota interactions can be facultative may be by itself advantageous for the hosts. For instance, mycorrhizal fungi are beneficial for the plants when the phosphorous concentration in the soil is low; but they are no longer beneficial when phosphorous is abundantly available and many plants are thus not mycorrhizal anymore in these conditions (Thomson *et al.*, 1986). Similarly, the digestion of lactose is often limited in human adults as they do not produce enough lactase: they thus have to rely on gut microbes, like *Bifidobacterium*, to partially digest lactose. However, some human populations have acquired a mutation that ensures higher lactase production and therefore guarantees autonomous and efficient digestion of lactose (Ségurel & Bon, 2017): in these individuals, the abundance of lactose-digesting bacteria is significantly lower than in individuals that do not have this mutation (Suzuki & Ley, 2020). Thus, host-microbiota interactions can be modulated according to the needs of the host, which depends on both its genotype and its environment (Box 1).

Because host-associated microbiomes often contain a plethora of microbial species, competition between microbes is important in these niches (Box 1). However, the different microbes may also provide different benefits to their host, which ensures their coexistence (Batstone *et al.*, 2018). For instance, such a complementarity of the symbionts in plant nutrition has been demonstrated for Mucoromycotina-Glomeromycotina dual symbiosis with liverworts (Field *et al.*, 2016) and the same applies among gut bacteria associated with bees, which partition their niches by specializing on different pollen-derived resources (Brochet *et al.*, 2021).

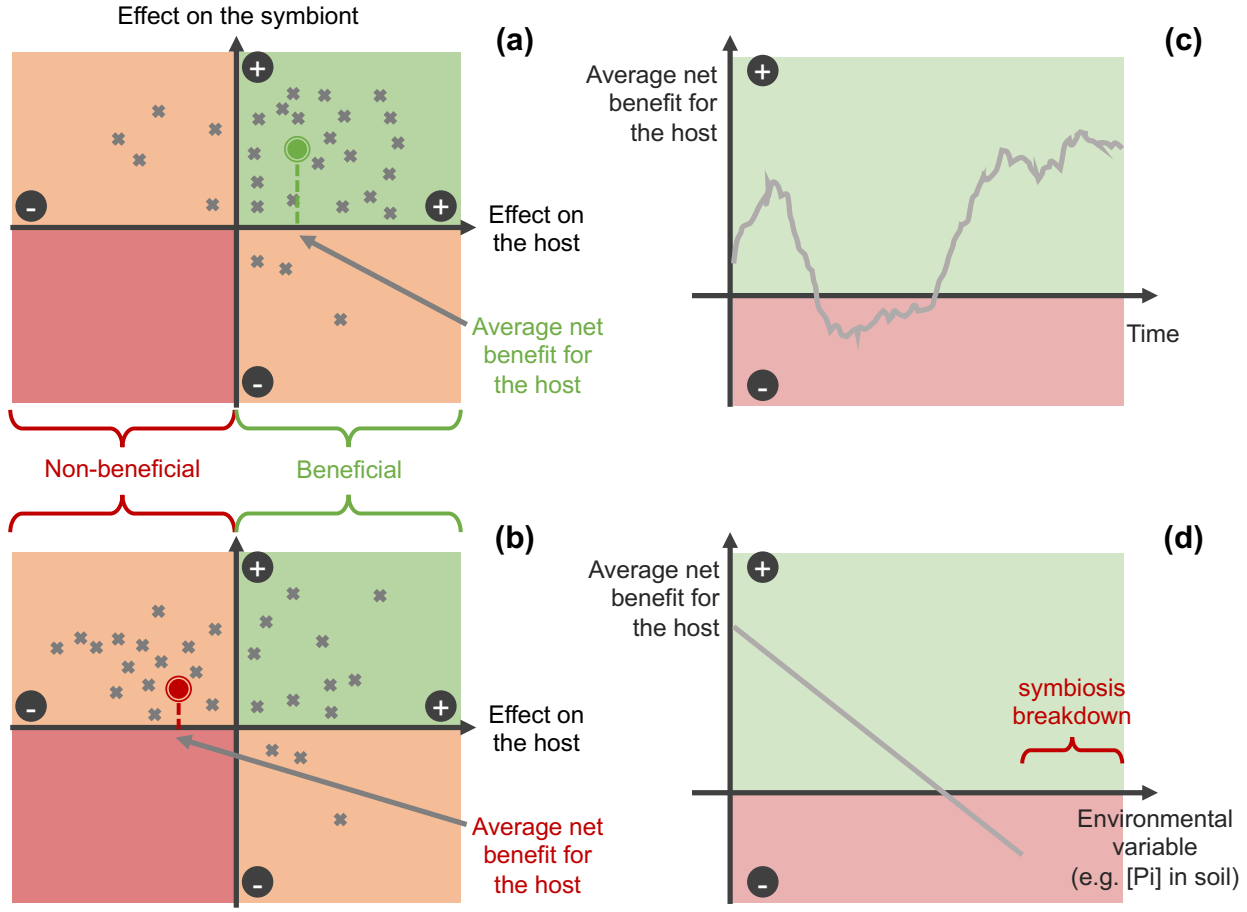


Figure 0.1.5: Host-microbiota systems form complex ecological communities: (a-b) Diagram indicated the nature of each host-microbe interaction within the microbiota and its effect on the host and the microbe. While these microbe-specific effects are difficult to experimentally quantify, it is rather easy to measure the average effect of the whole microbiota on the host. Thus, large dots indicate the average effect of the microbial symbionts on the host: it can be beneficial (green; a) or non-beneficial (red; b). (c-d) Examples of variation of the average benefit of the microbiota for the host as a function of time (c, *e.g.* the average effect varies according to the time of the year or over evolutionary timescales) or as a function of environmental conditions (d; *e.g.* it varies according to the available nutrient, like the available phosphorous in the soil [Pi] for mycorrhizas).

Box 1: A true mutualism?

All the functions mediated by the microbiota illustrate that most host-microbe interactions are largely beneficial to the hosts. But are all the microbes beneficial to the host? And reciprocally, are hosts always beneficial to their microbes?
Because microbiota are composed of a multitude of microbe lineages, measuring the effect of each microbe on the host (and *vice versa*) is challenging. However, experimental works suggest that microbiota are constituted by microbes ranging from beneficial to neutral and to non-beneficial (Johnson *et al.*, 1997). Instead of the effect of the individual microbes, it is the overall beneficial effect of the whole microbiota on the host that matters, and that is generally positive (Figure 0.1.5a-b).
Concerning the effect of the host on the microbes, for mycorrhizal fungi, mineral resources gathered by the fungi are generally traded against organic matter produced by plant photosynthesis: plants generally spend more than 10% of their organic matter into these symbiotic exchanges (Leake *et al.*, 2004; van der Heijden *et al.*, 2015). In addition, by hosting fungi in their roots, plants provide shelter to their mycorrhizal symbionts: for instance, arbuscular mycorrhizal fungi store lipid reserves in vesicles inside the plant roots. Mycorrhizal symbioses are mainly mutualistic as both partners generally benefit from the interactions (Figure 0.1.5a), although plants can sometimes have uncooperative strategies (see section 2.4). The same likely applies to gut microbes that are fed and protected by animals. However, in some cases, like in the foregut of the ruminants, the host maintains the fiber-fermenting microbes in the rumen, but once the fermentation is done, the host directly digests the microbes: whether such "farming" is a true mutualistic interaction can therefore be questioned (Mushegian & Ebert, 2016).
In addition, the benefits of an interaction are not static: for instance, according to the timing (Figure 0.1.5c) or the environmental conditions (Figure 0.1.5d), the overall benefit of the whole microbiota on the host can shift from positive to negative. It can vary on short time-scales during the life of the individuals or also over long time-scales (see section 2.2). In this case, it can exist a range of strategies to avoid hosting costly symbionts (see section 2.4).

Besides host-microbe interactions, the different microbial lineages composing the microbiota are also interacting with each other. They are in particular competing for space and resources (Bauer *et al.*, 2018) or can interact antagonistically (*e.g.* predation), but facilitation or mutualistic interactions (*e.g.* syntrophy) can also occur between specific strains (Morris *et al.*, 2013).

Therefore, host-microbiota interactions form a complex set of interlinked interactions involving multiple partners and can be strongly modulated by environmental conditions. They are thus far less simple to study and characterize than classical textbook examples of mutualism between a pair species (Selosse, 2000).

1.3. The assembly of host-associated microbiota

Microbiota transplantation experiments from one host species to another in mammals and arthropods often found that such interspecific transfers lead to a fitness decrease for the receiving hosts (Chung *et al.*, 2012; van Opstal & Bordenstein, 2019; Moeller *et al.*, 2019). This therefore highlights that some microbes might be specific to their host species and that this specificity is important for the good functioning of the host organism (Moeller & Sanders, 2020). Therefore, at each host generation, some mechanisms should exist to ensure that the microbiota assembly of the new generation includes species-specific microbes. The assembly of a new microbiota depends on the heritability of the microbes (Bright & Bulgheresi, 2010; Figure 0.1.6). On one hand, microbes could be **transmitted** directly by the parents to the newborn or may colonize from conspecific hosts. On the other hand, microbes can be directly **acquired from the environment** at each generation, independently of the microbes colonizing the parents or other conspecific hosts. Note that the terms "vertical" *versus* "horizontal transmissions" are gener-

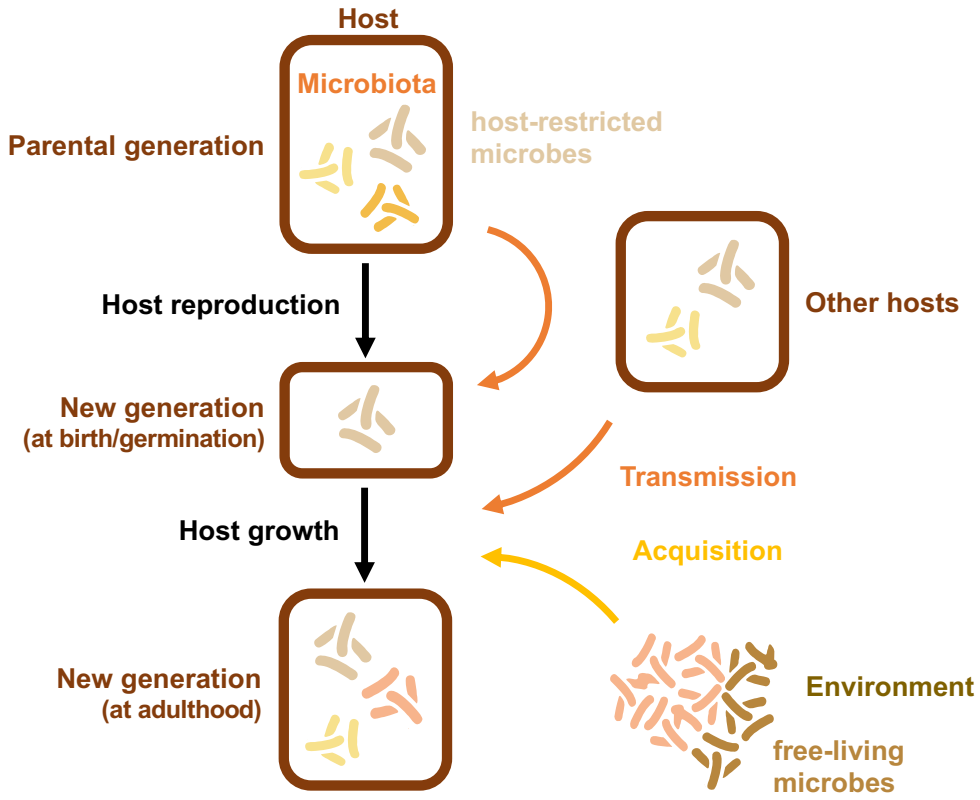


Figure 0.1.6: Illustration of the different modes of assembly of the host microbiota. Hosts (brown rectangles) reproduce and the parental generation can directly transmit its microbial symbionts to the next generation by transmission through the germline, at birth/germination, or through parental care. Microbes can also be transmitted from other hosts that share the same niche. Alternatively, after birth/germination, microbes can be acquired from the environment among the pool of available microbes, including free-living ones. Note that transmission is particularly expected for host-restricted microbes, whereas free-living microbes in the environment can easily be acquired. In any case, these microbes can durably colonize the microbiota niche (resident microbes) or just be present in the short-term (transient microbes).

ally used in the literature to refer to these two modes of assembly (Bright & Bulgheresi, 2010). In this manuscript, we rather use the terms "transmissions" *versus* "environment acquisition" and restrict the use of "vertical" and "horizontal transmissions" to refer to processes over long-time scale (see section 2.1).

1.3.1. Microbiota assembly in animals

Microbes heritability can be investigated thanks to experiments looking at microbiota composition between generations. In mammals, newborns are sterile at birth but there is plenty of evidence of parent-to-offspring transmission, in particular during the delivery (Dominguez-Bello *et al.*, 2010). Following delivery in humans, breast milk enriched in complex polysaccharides favors the establishment of *Bifidobacterium* symbionts in the gut of the newborn. During the lifetime of an individual, microbes are also transmitted between conspecifics sharing the same environment or through social contact (Moeller *et al.*, 2013). Besides transmitted microbes, the resident gut microbiota is also composed of microbes acquired from the environment (David *et al.*, 2014), and continuously faces a plethora of transient microbes (Zhang *et al.*, 2016). For instance, in mammals, many bacteria found in predator guts are also found in their preys (Moeller *et al.*, 2017), suggesting that a lot of (possibly transient) microbes transit thought the trophic webs. However, recent experiments with different mouse lines (*Mus musculus*) sharing the same cage demonstrated that new generations tend to conserve some ancestral bacterial symbionts (Moeller *et al.*, 2018), highlighting the primordial role of parent-to-offspring transmission. In particular, obligate anaerobes were more frequently transmitted than aerobic microbes (Moeller *et al.*, 2018), suggesting that microbial traits (in particular those related with host restrictiveness) and not only host traits contribute to parent-to-offspring transmission. Whatever their origins, microbes that colonized the mammalian gut start a complex cross-talk with their host that is essential to the homeostasis of the whole gut microbiota and the host (Kelly *et al.*, 2005).

Other vertebrates show less frequent cases of faithful parent-to-offspring transmission (Burns & Guillemin, 2017), probably because they generally present less parental care than mammals. The same applies to invertebrates that have often even fewer social contacts. Nevertheless, some arthropods have developed alternative ways of direct transmissions: contrary to vertebrates, arthropod newborns may not be sterile and some symbionts can be directly transmitted with the maternal germline or symbiont-containing secretions can be deposited onto the eggs (Bright & Bulgheresi, 2010; Engel & Moran, 2013). These mechanisms are particularly frequent when gut microbes play essential roles in insect nutrition. Conversely, when the evidence of host functions relying on microbes is scarce, it has been observed that microbes are rather acquired from the environment (Hammer *et al.*, 2017, 2019). For instance, the microbiota of *Drosophila melanogaster* in wild populations is highly variable between individuals and mainly constituted by microbes common in their environment (Blum *et al.*, 2013). In addition, a feeding experiment of a spider (*Badumna longinqua*) demonstrated that their gut micro-

biota directly derived from that of their insect prey (Kennedy *et al.*, 2020).

1.3.2. Microbiota assembly in plants

In plants, microbial heritability also range from generation-to-generation transmissions to environmental acquisitions (Hacquard *et al.*, 2015). Although plants mainly acquire their microbiota from their environment at germination, plant spores and seeds are not always sterile (Truyens *et al.*, 2015) and frequent cases of parent-to-offspring transmissions have been described: some parental endophytes have been found to directly colonize seeds (Shade *et al.*, 2017) and spores of mycorrhizal fungi (which cannot colonize plant seeds) have been characterized in the fruit tissues surrounding the seeds, guarantying the presence of the symbionts at germination (Séne *et al.*, 2018). When sterile, seedlings are quickly colonized after germination by mycorrhizal fungi from the surrounding environment: if seedlings germinate close to their parents, they likely encounter the same symbionts (Vannier *et al.*, 2018), whereas in the case of long-distance dispersal, they might face different pools of symbionts. In both cases, the encounter of microbial symbionts in the soil is often nonrandom (Shade *et al.*, 2017). For instance, mycorrhizal fungi are oriented toward the plant thanks to the emission by the plant of substances in the soil, such as strigolactones (Bonfante & Genre, 2015), starting a cross-talk between the interacting plant and fungus that is essential for the establishment of the symbiosis (Gadkar *et al.*, 2001). Microbiota assembly thus depends on the host species and on the microbes available, but also on the order of arrival of the microbial species (priority effect; Werner & Kiers, 2015; Leopold & Busby, 2020; Kohl, 2020).

Therefore, modes of microbiota assembly range from strict parent-to-offspring transmissions to random environmental acquisitions: most host microbiota assembly lies actually somewhere in the middle (Shapira, 2016). These modes of assembly can also have a drastic impact on the host-associated microbial symbionts, depending on whether they are host-restricted or whether they can live freely in the host's environment (Moran *et al.*, 2019). In the latter case, microbial symbionts are not much affected by microbiota assembly, whereas in the former case, they strongly rely on the host niches and thus depend on mechanisms favoring microbiota conservatism: they benefit from faithful transmissions from parent-to-offspring or through close contact between generations.

To conclude, host-associated microbes are ubiquitous in particular in animal guts and plant roots. In addition, microbiota often participate in essential host functions. They can provide novel ecological niches and confer to their host a rapid adaptability. It therefore raises the question of the evolution of such host-microbiota interactions: How do host-associated microbiota evolve? What is the impact of microbiota on the evolution of their hosts? What guarantees the stability of such interactions over long time scales?

2. The evolution of host-microbiota interactions

The Modern Synthesis of evolutionary biology, which unites the theories of Charles Darwin and Gregor Mendel on natural selection and heredity respectively, recognizes the role of four processes in the changes thought time (*i.e.* evolution) of populations (Huxley, 1942): mutation (*e.g.* the heritable changes of the genetic material that lead to phenotypic variability), selection (*e.g.* the reproductive advantages of the organisms having some particular phenotypic traits), drift (*e.g.* the random transmission of the genetic material to the next generation), and dispersal (*e.g.* the mixing with organisms from different populations). Therefore, through these processes, two evolving populations can accumulate divergences (microevolution), such that two organisms from each of these populations may become too different to efficiently reproduce, therefore forming two species (speciation). If all the individuals of a species die, species can disappear (extinction), and over long timescales (macroevolution), the balance between speciation and extinction (net diversification) determines the diversity of a clade of organisms. These mechanisms apply as well to organisms in interactions: host organisms and their associated microbes are (separately and/or jointly) experiencing mutation, selection, drift, and dispersal, altogether shaping their evolutions.

Here, we will see that host-microbiota interactions are rather conserved over evolutionary time scales and that this pattern can be due to diverse ecological and evolutionary processes, including vertical transmissions or host-mediated environmental acquisitions (section 2.1). Besides such conservatism, microbiota can also experience major shifts during host evolution, which can be linked to the acquisition of new functions for the host (section 2.2). Therefore, hosts and their associated microbes can reciprocally influence their macro-evolutionary histories (section 2.3). Finally, we will see that the stability of such mutualistic interactions can be challenged and that cheating strategies often emerge (section 2.4).

2.1. The evolutionary conservatism of host-microbiota interactions

Studies that investigated the microbiota composition of a given clade of animals or plants often reported that closely related hosts tend to have more similar microbiota composition than distantly related ones, suggesting that host-associated microbiota are rather conserved over evolutionary timescales. Indeed, although diets and soil properties have a major impact on gut and root microbiota respectively, the host evolutionary history often additionally contributes to microbiota composition (Wehner *et al.*, 2014; Groussin *et al.*, 2017; Yeoh *et al.*, 2017). Consequently, the microbiota differentiations between host species (*i.e.* their dissimilarities in microbiota composition) often recapitulate the host phylogeny. This pattern of phylogenetic signal in microbiota composition reflects rather slow and continuous changes of the microbiota along the host phylogeny; it is referred to as **phylosymbiosis** (Lim & Bordenstein, 2020 ; see section 3.3 for the different ways to quantify it). Phylosymbiosis has been widely observed across animals and plants, like in

the microbiota of mammal guts, including great apes (Ochman *et al.*, 2010), arthropods (Brucker & Bordenstein, 2013; Armstrong *et al.*, 2020), and also in plant roots, including in grasses (Bouffaud *et al.*, 2014), willows (Tedersoo *et al.*, 2013), orchids (Jacquemyn *et al.*, 2011), or lycopods (Benucci *et al.*, 2020), as well as in their leaves (Donald *et al.*, 2020).

However, phylosymbiosis is not a generality: for instance, no phylosymbiosis has been reported in birds (Hird *et al.*, 2015), bats (Song *et al.*, 2020), or flowering plants (Erlandson *et al.*, 2018). The absence of phylosymbiosis is especially observed when host microbiota are composed of transient microbes acquired from the environment. In addition, despite the presence of a phylosymbiosis at the host species level, intra-specific host-microbiota differentiation can be rather scarce, as illustrated by the apes that tend to present more similar microbiota when they share the same area, irrespectively of their genealogical relationships (Degnan *et al.*, 2012). As a matter of fact, phylosymbiosis is only a pattern and not a process. To understand the origins of microbiota conservatism and phylosymbiosis over long-time scales, we have to investigate how host-associated microbiota are assembled and evolve over short-time scales.

2.1.1. Conservatism driven by microbial transmission

First, the **transmission of microbes** in a host lineage can explain phylosymbiosis. Indeed, host-associated microbes are often transmitted from the parents to the offspring or colonize from other conspecific hosts (see section 1.3). If these transmissions are stable and faithful, host-microbe interactions are conserved in the host lineage over long-time scales (we refer to this process as vertical transmission). At host speciation, transmitted microbes are then isolated from each other in the two daughter host lineages: by accumulated divergences in each host lineage, the phylogenetic tree of these transmitted microbes therefore tends to mirror that of their hosts (a pattern referred to as **cophylogeny**; Figure 0.2.7). By default, we assume that transmitted microbes passively undergo host speciations (phylogenetic tracking), although they might also actively promote host speciations in some rare cases (see section 2.3). Such cophylogenetic patterns have been frequently observed in the intracellular endosymbionts of invertebrates, like the bacteria *Buchnera* of aphids (Moran *et al.*, 2008), or in the dominant gut bacteria in stinkbugs (Hosokawa *et al.*, 2006), both maternally inherited. In species-rich microbiota, such cophylogenetic patterns are more difficult to investigate because the classical metabarcoding markers, like the 16S SSU rRNA gene, often do not have enough resolution to reconstruct robust phylogenies of the transmitted microbes (see section 3.1), but several studies amplifying more resolutive markers or taking into account phylogenetic uncertainty have demonstrated that transmitted bacteria are present in the gut microbiota of insects (Kwong *et al.*, 2017) and mammals (Groussin *et al.*, 2017; Youngblut *et al.*, 2019), including great apes (Moeller *et al.*, 2016). It also appeared that some of these vertically transmitted microbes have punctually experienced horizontal transfers (host-switching) between different host lineages, resulting in non-perfect congruency between the host and the mi-

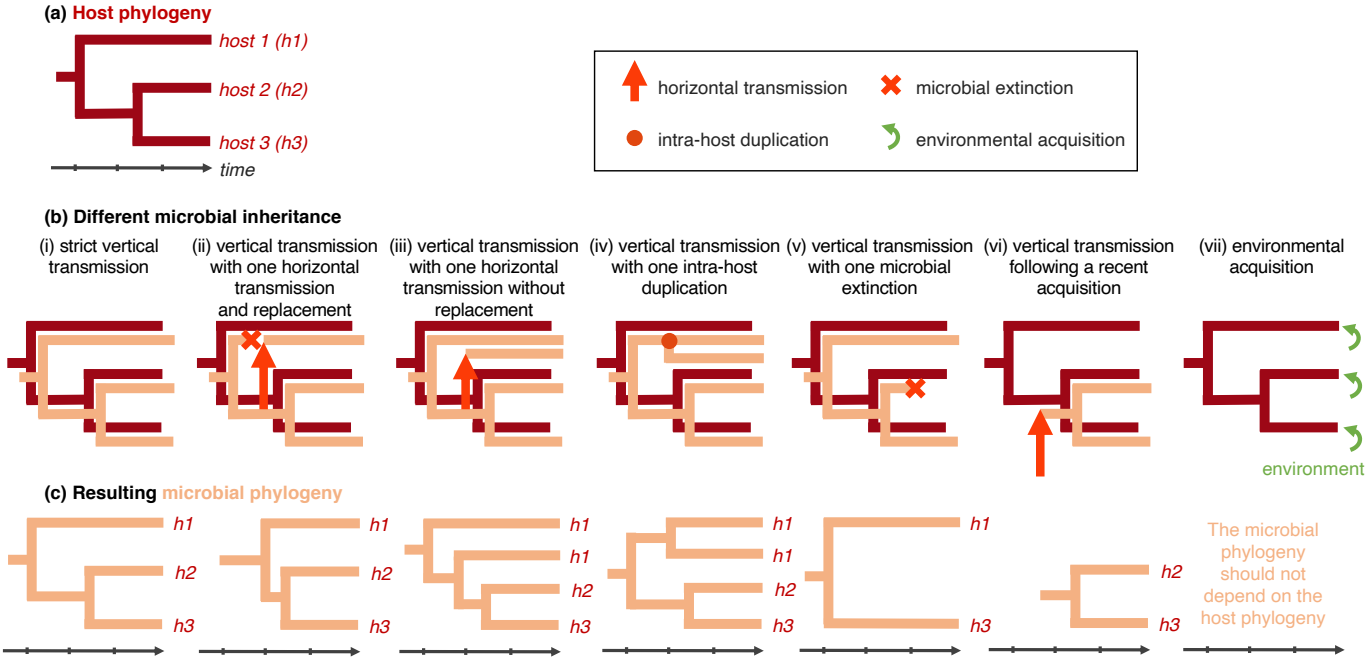


Figure 0.2.7: Different modes of inheritance of a given host-associated microbial lineage and their consequences on the microbial phylogenies. On a phylogenetic tree of 3 host species (a), we represent the different modes of microbial inheritance (b) and their resulting microbial phylogeny (c): extreme scenarios correspond to strict vertical transmission (i; perfect cophylogenetic pattern) or environmental acquisition (vii; no cophylogenetic pattern expected). Under vertical transmission, other punctual processes can result in a loss of perfect congruence between the host and the microbial phylogenies: punctual events of horizontal transmissions (the horizontal transfer from one donor host to a receiver host can result in microbial replacement in the receiver host lineage (ii), or to a duplication of the microbial strains in the receiver host lineage (iii)), intra-host duplication (iv), or microbial extinction (v). Additionally, the interaction could have been acquired recently in only a sub-clade of hosts (vi; the microbe was therefore absent in the most recent common ancestor of all hosts). Here, note that we do not call 'cospeciation' the fact that the microbes split into two lineages at host speciation, because these resulting isolated microbial populations can be differentiated without being necessarily "two species".

crobes phylogenies (Moeller *et al.*, 2016; see Figure 0.2.7 for a list of processes that can generate imperfect congruence between the host and microbe phylogenies). In great apes, transmitted bacteria include the host-restricted Bacteroidaceae (Bacteroidetes) and Bifidobacteriaceae (Actinobacteria), whereas the Lachnospiraceae (Firmicutes), which form spores and can survive outside animal guts harbor no cophylogenetic patterns (Moeller *et al.*, 2016): this suggests that low dispersal ability and host-restrictiveness might be prerequisites for a microbe to be vertically transmitted over long-time scales. These prerequisites could in particular explain why evidence of vertical transmission is rather scarce in mycorrhizal fungi, contrary to non-soil dwelling fungal endophytes (Rodriguez *et al.*, 2009).

Although only a subset of the host-associated microbes is vertically transmitted, it can be sufficient to generate phylosymbiosis (Nishida & Ochman, 2019). Importantly, if vertical transmission often tends to generate cophylogeny, the opposite is not true:

a cophylogenetic pattern is not by itself sufficient evidence for demonstrating vertical transmission, as it can also emerge if both hosts and microbes have experienced concomitant speciation events (*e.g.* there are subject to the same vicariance events) or if the microbes have recently colonized the host clades through horizontal transfers between closely related hosts (host-shift speciations; de Vienne *et al.*, 2013).

2.1.2. Conservatism driven by host filtering

Second, phylosymbiosis has also been observed in host-microbiota systems where the evidence of vertical transmission is scarce and microbes are instead acquired from the environment (Benucci *et al.*, 2020): when microbiota are constituted by environmental acquisitions, host microbiomes can be seen as specific niches with particular environmental conditions that microbes may colonize (Moran & Sloan, 2015; Kohl, 2020). Because of its properties, a given host microbiome may not be suitable for all microbes: a **host filtering**, where the host conditions act as ecological conditions, is operating during microbiota assembly and selecting particular microbes. In animal guts, we can think about the gut morphology or physiology (*e.g.* pH) or the expression of particular antimicrobial peptides (Franzenburg *et al.*, 2013; Nishida & Ochman, 2018). Similarly, in plant roots, root morphology and the emission of particular molecules in the soil (*e.g.* salicylic acid) can impact microbiota composition (Lebeis *et al.*, 2015). If these host traits affecting microbiota assembly are relatively conserved and slowly evolving, we would expect closely related host species to harbor similar traits, and therefore, they should host similar microbial communities in their microbiomes (Moran & Sloan, 2015). Simulations have recently demonstrated that such processes can produce phylosymbiosis (Mazel *et al.*, 2018). Importantly, the evolutionary changes in the host traits involved in microbiota assembly can be completely neutral, and do not have to be under selection.

2.1.3. Conservatism driven by inhomogeneous microbial pools in host environments

Third, besides transmission and host filtering, phylosymbiosis can simply appear if closely related host species tend to live in environments with similar pools of available microbes. In particular, in animals, if gut microbes are mainly acquired through the food, host species with similar diets have similar microbiota (Kohl, 2020). Similarly, spatial distances and environmental conditions are major determinants of the plant-associated arbuscular mycorrhizal fungal communities at the global scale (Davison *et al.*, 2015): as closely related plant species tend to occupy close geographic areas with similar environments, this can simply contribute to the significant phylosymbiosis. **Inhomogeneous microbial pools** rely either on the differential ecological filtering of microbes across environments or on the existence of dispersal limitations in microbes. The latter appears to be particularly important in the mammalian microbiota as their similarity of composition decreases with geographic distances (Moeller *et al.*, 2017).

Importantly, these different processes are not exclusive, and in many hosts, several

of these processes likely contributed to phylosymbiosis (Davison *et al.*, 2015; Mazel *et al.*, 2018). Whether initial divergences can be primed by transmitted microbes, host filtering, or inhomogeneous microbial pools, these divergences can later be exacerbated by priority effects between microbial species (*i.e.* the first settled microbes influence the establishment of others, *e.g.* through competition or facilitation) that can influence the assembly of the host microbiota during the rest of the host life (Kohl, 2020).

2.2. Shifts in microbiota composition and the acquisition of new functions

Although host-microbiota tend to be evolutionary conserved, major shifts in the microbiota composition have also occurred punctually during animal and plant evolutions and were often concomitant with ecological innovations, *i.e.* significant changes in the niches or environments of their hosts.

New transient microbes are continuously passing through the gut and the root microbiota, and resident symbiotic microbes also tend to largely vary among individuals of a given host species, but nonetheless, the emergence of new host-microbe symbioses is rather rare. During land plant evolutions, only a few shifts of the main mycorrhizal symbionts occurred (Selosse & Le Tacon, 1998; Brundrett & Tedersoo, 2018), and were mostly concomitant with drastic changes in plant niches (Martos *et al.*, 2012; Werner *et al.*, 2018). Similarly, in mammals or arthropods, major shifts in microbiota composition tend to correspond to changes in diets or niches (*e.g.* the transitions toward aquatic environments in whales; Russell *et al.*, 2009; Sanders *et al.*, 2015; Groussin *et al.*, 2017). In addition, these shifts in microbiota composition are often non-random but rather convergent between host lineages that evolved to living in similar niches (Bittleston *et al.*, 2016). For instance, several plant lineages have convergently developed mycorrhizal interactions with the same fungal lineages, like the Sebacinales or the Helotiales (Weiß *et al.*, 2016; Hoysted *et al.*, 2018). Similarly, transitions from carnivory to herbivory in mammals tend to be associated with an increase of the proportion of fibrolytic Firmicutes (Groussin *et al.*, 2017) and ant-eating lineages also convergently acquired similar gut microbes (Delsuc *et al.*, 2014), although host phylogenetic inertia is also frequent (Ley *et al.*, 2008; Delsuc *et al.*, 2014). Indeed, in some cases, the gut morphology and physiology might have constrained the shifts in microbiota compositions (Nishida & Ochman, 2018), which suggests that such shifts are primarily constrained by the abilities of the animals or plants to host these new microbial communities.

Moreover, these new symbiotic microbes frequently appear to be beneficial for their new hosts by bringing them new functions. We can distinguish three different origins for these new microbial-mediated host functions according to the origin of the microbes (Figure 0.2.8): microbes can either (i) be horizontally transmitted from other host lineages, (ii) be acquired from the new environment of the hosts, or (iii) already be present in the host

environment. First, horizontal transmission of new symbionts may indeed directly provide the host with new functions, as illustrated by the dynamics of horizontal transfers of bacterial endosymbionts in aphids that provide a range of new functions (Henry *et al.*, 2013). Second, if the animal or plant hosts change their environment, symbionts acquired from the new environment may facilitate their establishment: for instance, some insects rely on microbes from their new environment for metabolizing environmental toxins preventing their establishment (Kikuchi *et al.*, 2007). Third, the host and some microbes already present in the host's environment can evolve a new type of association. For instance, the Sebacinales forming mycorrhizas with many plant species were originally saprotrophs that evolved the abilities to associate with plants as endophytes before eventually developing truly functional mycorrhizas (the "waiting room hypothesis"; Selosse *et al.*, 2009, 2018). Whatever their origins, by bringing new functions, these microbial symbionts can rapidly foster host adaptation, which can generate selective pressures for the hosts to associate with these new microbes and promote the host dependence toward them (see Box 2).

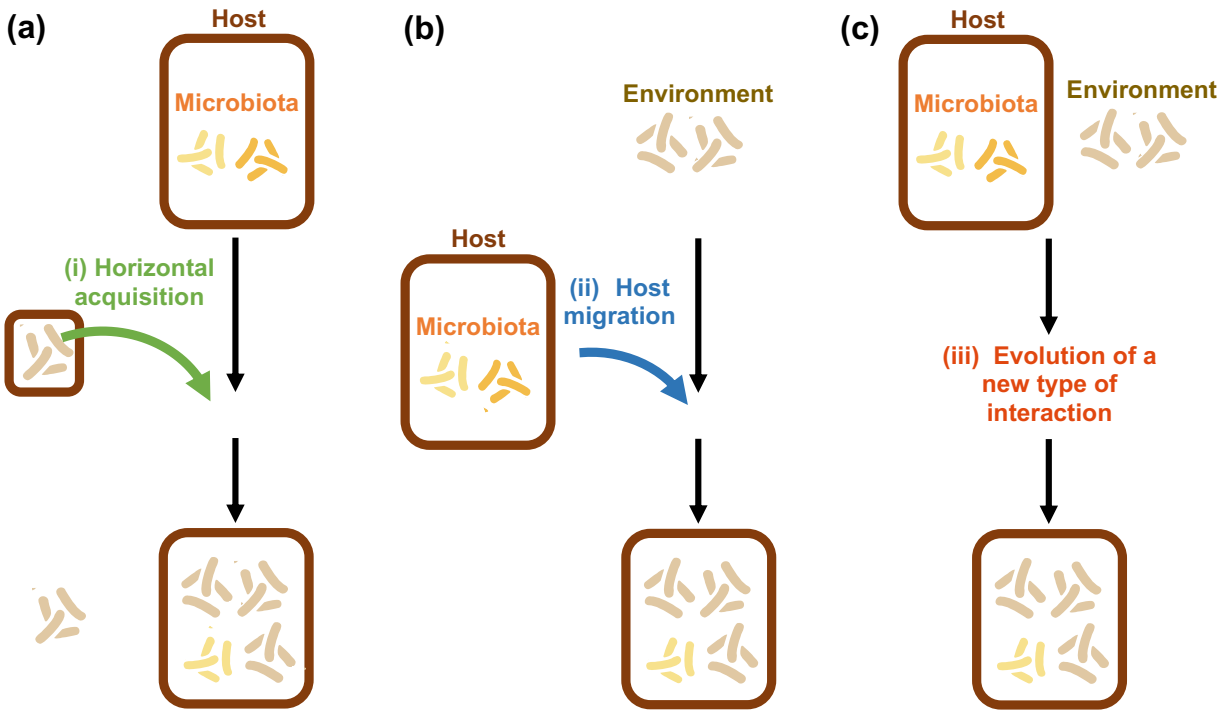


Figure 0.2.8: Three different origins for new microbe-mediated host functions. Hosts (brown rectangles) acquire a new function thanks to their association with a new microbial taxon (the grey microbes). (a) The new microbial symbiont can be horizontally acquired from another host. (b) The host can change its environment/niche and acquire new symbionts from their new environment. (c) A new type of interaction can involve *de novo* if a host and an environmental microbe are in close contact in the environment. Importantly, a host-associated microbe (*e.g.* the yellow microbes) can also acquire new abilities (*e.g.* through horizontal gene transfer) without requiring any taxonomical change (*i.e.* new microbes are not mandatory for new microbial-mediated host functions).

However, systemically linking major shifts in host microbiota composition to the ac-

quisition of new microbes or new functions is over-simplistic. First, major shifts in the microbial relative abundances can also occur when host-microbe interactions are simply lost: for instance, birds and bats seem to have convergently lost their associations with Bacteroidetes but conserved their associations with Proteobacteria during the evolution of flight (Song *et al.*, 2020). Second, like most mutations in a genome are deleterious, most random changes in a microbiota are likely not advantageous for the host (Suzuki & Ley, 2020): many changes, including major ones, may thus not convey any changes in their functional properties to the hosts, given that microbial functional properties are highly redundant (Louca *et al.*, 2016). Finally, the taxonomic composition of a microbial community revealed using marker genes like the 16S SSU rRNA gene might not reflect the functional properties of the microbes, given microbial genomes are very labile thanks to frequent (non-homologous) horizontal gene transfers. The acquisition of new functions could thus appear without any changes in the microbiota composition (Hehemann *et al.*, 2010). Therefore, we should rather track the presence of the microbial genes involved in host functioning rather than looking at microbial taxonomy, as illustrated by the legumes-associated rhizobia that present a large heterogeneity in terms of nitrogen-fixing abilities (Young & Haukka, 1996).

Therefore, it appears that the acquisition of new microbial symbionts may greatly help the hosts to expend their ecological niches (*e.g.* acquisition of new functions or colonization of new niches or environments; Margulis & Fester, 1991; Moran *et al.*, 2019). Such associations can therefore significantly impact the evolutionary success of the hosts over long-timescales.

Box 2: How did host-microbe dependences emerge?

Most animals and plants have developed dependences on their microbial symbionts, and *vice versa*, for realizing essential functions (Chomicki *et al.*, 2020a). However, a dependence on microbes for insuring essential host functions can represent a serious risk for the hosts, as they now depend on the presence of these microbes and their ability to interact with them. To minimize the risk of not recovering the specific microbes from the environment, some hosts have evolved strategies to faithfully transmit their microbial symbionts to the next host generation (see section 1.3). Alternatively, when the host functions do not need specific microbes (like the priming of the host immune system), there is often little risk of relying on the presence of microbes as they are everywhere, given that, since their origins, animals and plants dwell in a microbial world (McFall-Ngai *et al.*, 2013).

Box 2: How did host-microbe dependences emerge? (end)

Then, if the risk of not recovering the beneficial microbes is minor for the host, is evolving a dependence always beneficial? Sometimes these dependences are related to new functions that the host cannot realize alone, like new nutritive strategies or detoxifying abilities. The symbiotic microbes therefore allow the hosts to expend their ecological niches (Margulis & Fester, 1991; Moran *et al.*, 2019). However, in some cases, the benefits of such dependence are unclear: for instance, mammals rely on microbes for priming the development of their immune systems (see section 1.2), so if microbes are absent, they lack an efficient immune system (Chung *et al.*, 2012). Is it advantageous to have developed such a dependence toward microbes while many organisms can prime their immune systems of their own? If priming their own immune system requires costly mechanisms for the hosts, it can be advantageous to rely instead on microbes: dependence can be selected (Black Queen hypothesis; Morris *et al.*, 2012). However, such dependence can also appear without any selective pressures: neutral evolution toward interdependence can emerge by random drift if two redundant mechanisms exist (Selosse *et al.*, 2014). The term "evolutionary addiction" has been proposed for the emergence of such dependences (Moran, 2002). In other words, animals and plants have frequently developed a dependence on their associated microbes for realizing some functions without any apparent additional benefits (Moran *et al.*, 2019).

Reciprocally, such dependence has also emerged in microbes: if it exists many pathways of vertical transmission (parental-to-offspring direct transmissions, proximity between hosts, ...), microbes may no longer need a free-living stage and a dependence toward the hosts can appear, like observed in mammal guts where many transmitted bacteria are restricted to the gut niches (Moran *et al.*, 2019) or in mycorrhizal fungi that often became obligate plant symbionts (*e.g.* Glomeromycotina or some Basidiomycota lineages; Miyauchi *et al.*, 2020). In the same way as their hosts, dependence can be particularly reinforced through adaptive or neutral gene losses (*i.e.* Black Queen hypothesis *versus* evolutionary addiction; Hosokawa *et al.*, 2006).

2.3. The macroevolutionary impact of host-microbiota interactions

Why some clades of species are more diverse than others is a central question in biology. The Red Queen hypothesis proposes biotic interactions as a central force promoting endless adaptive changes in interacting species that result in diversification (Van Valen, 1973). Indeed, selective pressures mediated by host-microbiota interactions are frequent. For instance, gut microbes in *Drosophila melanogaster* can promote the adaptive genomic changes in their host over very short time scales (Rudman *et al.*, 2019) and reciprocally, nitrogen-fixing rhizobia associating with legumes have been found to rapidly adapt to their hosts to increase cooperation (Batstone *et al.*, 2020). Hembry *et al.* (2014) distinguished two ways by which biotic interactions may increase species diversification: either by directly spurring their speciation or by indirectly increasing their diversification.

2.3.1. Biotic interactions can directly promote speciation

First, biotic interactions can directly promote speciation, in particular host speciations (Hembry *et al.*, 2014). Also referred to as "speciation by symbiosis", host-associated microbes can promote pre-mating isolations or post-mating isolations, and such isolations between host populations would spur speciation (Brucker & Bordenstein, 2012). Particularly well documented in arthropods (Vavre & Kremer, 2014), pre-mating isolations include evidence of isolations that are behavioral (Sharon *et al.*, 2010; but see Leftwich *et al.*, 2017) or ecological (Hosokawa *et al.*, 2007), while post-mating isolations often consist in microbe-mediated incompatibilities (Duron *et al.*, 2008). However, if such mechanisms can easily spur population differentiation, especially in a geographic metapopulation context (Thompson, 2005), there is currently only very little evidences that they can successfully result in complete speciation, or directly increase the diversification of the clade (Althoff *et al.*, 2014; Hembry *et al.*, 2014).

2.3.2. Biotic interactions can indirectly promote diversification

Second, biotic interactions can create an ecological opportunity for the interacting clades of species that may promote the ecological and evolutionary success of one or both of them (Hembry *et al.*, 2014). Technically, biotic interactions may increase the net diversification rates of the clades (*i.e.* increasing the speciation rates or decreasing the extinction rates; see section 3.4) without needing to actively promote the speciation. Many animal or plant radiations have been only possible because of their symbioses with microbes that have offered them the ability to expand their niches or colonize new environments. For instance, the colonization of land by plants and their latter radiations would have not been possible without their nutritive symbiosis with mycorrhizal fungi, including the Glomeromycotina (Pirozynski & Malloch, 1975; Selosse & Le Tacon, 1998) or the Mucoromycotina (Strullu-Derrien *et al.*, 2018; Chang *et al.*, 2019). Similarly, the ancestors of all mammals were likely carnivorous, but nowadays, 80% of the extant mammals are herbivorous despite the lack of enzymes degrading complex plant fibers in mammalian

genomes (Flint *et al.*, 2012): the radiation of herbivorous clades has thus only been possible thanks to fibrolytic microbial symbionts (Hacquard *et al.*, 2015). In addition, the numerous beneficial functions ensured by microbial symbionts might improve host survival and increase their population sizes, resulting in lower extinction rates (Chomicki *et al.*, 2019). Besides spurring the diversification of their associated animals and plants, colonizing host microbiomes can also promote the radiation of the corresponding microbial lineages. For instance, more than 80 fungal lineages (representing more than 40,000 extant fungal species) have evolved mycorrhizas (van der Heijden *et al.*, 2015; Brundrett & Tedersoo, 2018), suggesting that developing symbiosis with plants was a significant evolutionary success for fungi (Wilson *et al.*, 2017).

However, if host-microbe interactions are often thought to increase diversification, such associations can also decrease it (Chomicki *et al.*, 2019). For instance, if the hosts depend on their microbes (or reciprocally) and if the interactions are difficult to establish (*e.g.* because of the rarity of the partners), this can increase the risk of extinction of the obligately dependent organisms (Kiers *et al.*, 2010; Chomicki *et al.*, 2020a).

Therefore, there is plenty of evidence that hosts and microbes have greatly influenced their respective evolutions through their interactions in many ways. If this evolutionary interplay between hosts and microbes is reciprocal and due to selective pressures, we may refer to it as coevolution (Janzen, 1980; Box 3).

Box 3: Coevolution in host-microbiota interactions

In its strict definition, coevolution happens when reciprocal selective pressures induce evolutionary changes in two interacting lineages (Janzen, 1980). Importantly, although durable associations are often necessary for coevolution, the intimacy of an interaction and its conservatism over long time-scales does not mean coevolution (Moran & Sloan, 2015): in that sense, coevolution is not a prerequisite nor a consequence of patterns of phylosymbiosis or cophylogeny (Poisot, 2015).
Some examples of host-microbe pairwise coevolutions have been demonstrated, especially in insect-endosymbiont interactions, but proofs of strict pairwise coevolution are scarcer in species-rich microbiota (like in the animal guts) or in microbiota constituted of microbes mainly acquired from the environment (like in plant roots). The detoxifying bacteria acquired from the host environment (see section 2.2) is a good example of microbes that promote host adaptation without requiring any coevolutionary processes (Suzuki & Ley, 2020).
Instead, coevolution is more likely to be diffuse in such systems, *i.e.* the interacting clades of hosts and microbes selectively influence each other as a group (Janzen, 1980), like in mycorrhizal symbioses (Brundrett, 2002) or in the mammalian guts where milk oligosaccharide productions have coevolved with the digestive abilities of several bifidobacteria (Asakuma *et al.*, 2011).

2.4. The stability of host-microbe mutualistic interactions

2.4.1. The breakdown of host-microbiota mutualism

Mutualistic host-microbiota interactions often come at a cost for one or both cooperators (Douglas, 2008): for instance, plant hosts have to provide organic carbon to their mycorrhizal partners in return for mineral matter. Such costs are not problematic if the benefits are larger, which guarantees a positive net benefit for the interacting organisms. However, such a balance between costs and benefits can vary (Figure 0.1.5).

First, the benefit-cost ratio of an interaction depends on the environmental conditions or the niche occupied by the host (Figure 0.1.5d). For instance, under high phosphorous concentration in the soil, mycorrhizal fungi are often no longer beneficial for the plants (Thomson *et al.*, 1986). In such conditions, the hosts can be under selective pressure to get rid of their useless microbial symbionts. When environmental conditions strongly impact the benefit-cost ratio of the interaction, it likely favors the strategy of facultative symbioses: for instance, 7% of the extant plant species only interact with mycorrhizal fungi when needed, and alternatively abandon the mycorrhizal mutualism (Smith & Read, 2008; Brundrett & Tedersoo, 2018). Definitive mutualism breakdowns have also repeatedly occurred during land plant evolution when plant lineages have evolved alternative nutritive strategies, like carnivory or cluster roots, and mycorrhizal symbionts were therefore no longer needed (Werner *et al.*, 2018).

Second, the benefit-cost ratio can be modified by the interacting partners themselves (Figure 0.1.5c). Indeed, some hosts or microbes frequently evolved adaptive uncooperative strategies for retrieving higher benefits from the interaction at the expense of their partners (Sachs *et al.*, 2004; Douglas, 2008). Such strategies, referred to as **cheating**, are parasitism evolved in the framework of a mutualism and can emerge for individuals within a species, or for a whole species in multiple-partners interactions. For instance, several plant lineages have evolved a cheating strategy toward their associated mycorrhizal fungi in terms of carbon supply: these lineages, referred to as **mycoheterotrophic plants**, such as some gentians or orchids, stopped providing organic carbon to their fungal partners and instead rely on them for both mineral and organic matter (Merckx, 2013; Figure 0.2.9). While many achlorophyllous lineages are full mycoheterotrophs, others have conserved their abilities to perform photosynthesis and only partially rely on their fungi for organic matter (mixotrophy or partial mycoheterotrophy) or only for one part of their development (initial mycoheterotrophy; see Box 4 and Figure 0.2.9; Selosse & Roy, 2009; Jacquemyn & Merckx, 2019). Similarly, some arbuscular mycorrhizal fungi have evolved antagonistic strategies as they reduce the fitness of their plant hosts (Johnson *et al.*, 1997; Selosse *et al.*, 2006). However, mycorrhizal plants and fungi provide a range of functions to their partners, from nutrition to protection, which can depend on environmental conditions, such that assessing the limits between a “mutualistic” and a “cheating” strategy can be difficult (Frederickson, 2017); indeed, even if one considers

the effect on fitness as an inclusive estimator, it remains difficult to assess especially for the microbial partners. For instance, in terms of carbon, mycoheterotrophic plants are clearly cheaters that are costly for their mycorrhizal partners; however, it is still debated whether or not the net benefit of the interactions is negative for fungi as mycoheterotrophic plants might be beneficial in other ways (*e.g.* vitamin synthesis; Merckx, 2013). Gut microbiota are also subject to cheating strategies: for instance, gut-associated *Bacteroides* or *Escherichia coli*, which are beneficial symbionts under many conditions, can eventually turn into opportunistic pathogens (Leung *et al.*, 2018), suggesting that cheating is latent in most host-microbiota interactions.

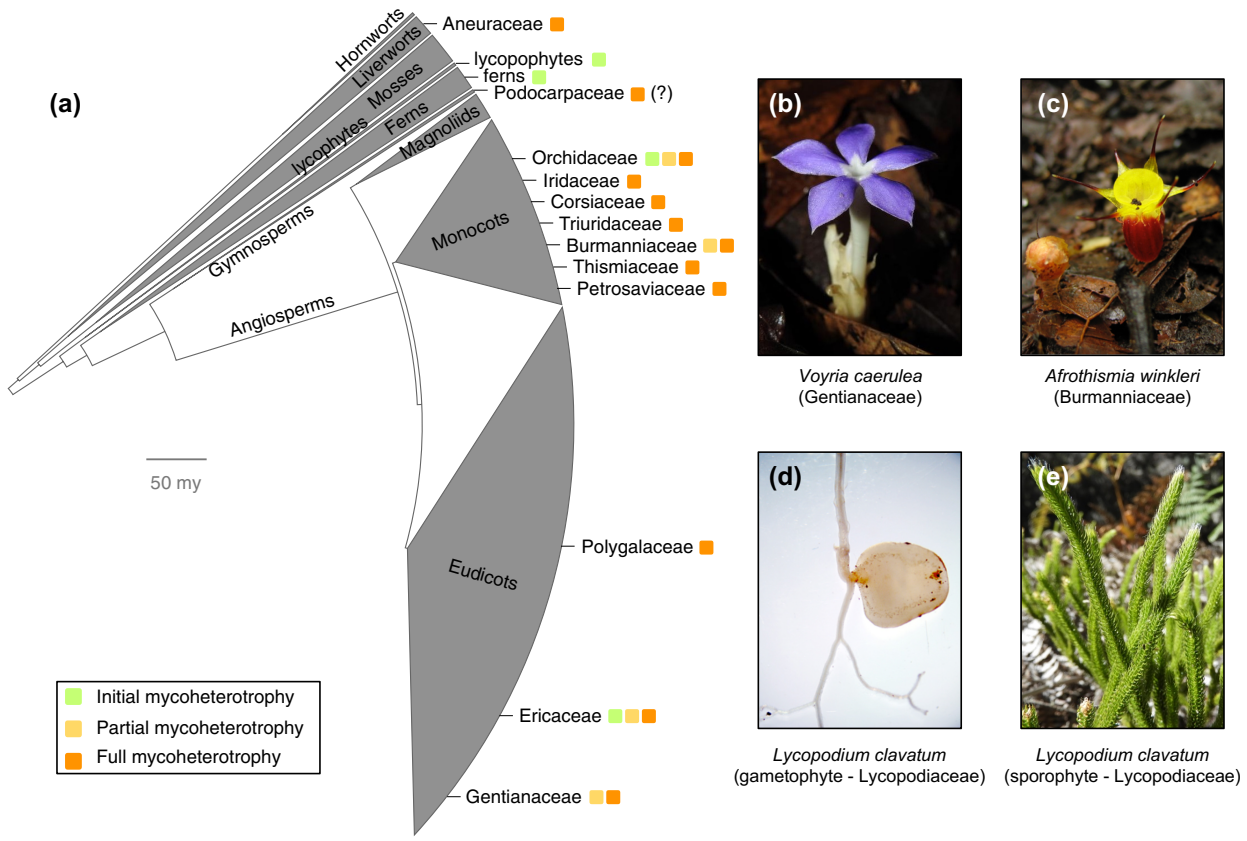


Figure 0.2.9: Multiple emergences of mycoheterotrophic cheating in land plants. (a) Phylogenetic tree of the land plants indicated the lineages having evolved mycoheterotrophic strategies (Figure from Jacquemyn & Merckx (2019)). Colors indicate full mycoheterotrophs, partial mycoheterotrophs, and initial mycoheterotrophs. Note that most cheating plants associate with Glomeromycotina symbionts, but others, like orchids, associate with different mycorrhizal lineages (*e.g.* Sebacinales or Russulales). (b-c) Photos of two full mycoheterotrophic plants: *Voyria caerulea* (b, eudicot, photo by Sébastien Sant) and *Afrothismia winkleri* (c, monocot, photo by Vincent Merckx). (c-d) Photos of the two life stages of *Lycopodium clavatum*, an initial mycoheterotrophic species: achlorophyllous gametophyte (d, photo by Curtis Clark under CC BY-SA 3.0) and autotrophic sporophyte (e, photo by Benoît Perez-Lamarque, Box 4).

Box 4: Initial mycoheterotrophy

Several lineages among ferns (Ophioglossaceae and Psilotaceae) and lycopods (Lycopodiaceae) and orchids have developed initially mycoheterotrophic strategies (Merckx, 2013).
Ferns and lycopods are characterized by a strict alternation of generations between haploid gametophytes and diploid sporophytes. While sporophytes are photosynthetic organisms mostly hosting arbuscular mycorrhizal fungi in their 'roots' (Lehnert *et al.*, 2017), gametophytes can be underground and achlorophyllous, and rely upon mycorrhizal fungi for their mineral and organic nutrition (Boullard, 1979). At first sight, these gametophytes are likely parasitic toward their mycorrhizal fungi. However, contrary to full mycoheterotrophic species that cheat during their entire development, adult sporophytes can repay the carbon invested by their mycorrhizal partners during their gametophytic stage. Therefore, rather than representing a parasitic cost to their associated fungi, they would be mutualistic over their entire development by differing in time their reward toward fungi: a model referred to as "take now, pay later" (Field *et al.*, 2015). Moreover, recent work has demonstrated that gametophytes and sporophytes likely shared the same fungi (Winther & Friedman, 2008), and suggested that the carbon could be directly transferred from the autotrophic sporophytes to the mycoheterotrophic gametophytes (a 'parental nurture') thanks to a shared 'wood-wide-web' (Leake *et al.*, 2008).
Similarly, orchids are mycoheterotrophic at germination since their seeds lack nutritional reserve. Then, most species turn to be full autotrophs or partial mycoheterotrophs as adults (Merckx, 2013).

2.4.2. Constraints upon cheating strategies

The origination and persistence of cheating among host-microbe mutualistic interactions could compromise their evolutionary stability (Ferriere *et al.*, 2002). However, it exists several mechanisms prevent or limit cheating emergence among mutualisms (Foster & Wenseleers, 2006). First, **partner-fidelity feedback** can prevent the emergence of cheating (Sachs *et al.*, 2004). Indeed, if host-microbe interactions are stable over-time, benefits provided by one species to its partner would inevitably feedback to it (Weyl *et al.*, 2010). In other words, under partner-fidelity feedbacks, both fitnesses are aligned, and there is no advantage to harm its partner (Sachs *et al.*, 2004; Selosse & Rousset, 2011). Partner-fidelity feedbacks are trivial in vertically transmitted symbioses and in obligate symbioses where there is no alternative choice among the partners, resulting in a durable and exclusive association. Second, **partner selection** can limit the interactions with cheaters. Indeed, hosts or symbionts can favor and invest more in the interactions with most cooperative partners through conditional investment (Roberts & Sherratt, 1998), stop interacting with the cheaters (Pellmyr & Huth, 1994), or even sanction them (Kiers *et al.*, 2003). Such mechanisms therefore counterselect the propensity

of cheating (Noë & Hammerstein, 1994). Partner selection is particularly important in multiple-partners interactions with partners acquired from the environment (Sachs *et al.*, 2004). There is evidence that both partner-fidelity feedbacks and partner selections stabilize the host-microbiota interactions: for instance, root-associated symbionts can align their fitness with that of their plants over a few generations (Batstone *et al.*, 2020) and both plants and mycorrhizal fungi can avoid cheaters and actively select their partners by conditional investments (Kiers *et al.*, 2011) or by controlling the rhizosphere composition (Lebeis *et al.*, 2015). In gut microbiota, there is also some proofs of partner selection from the hosts: for instance, by secreting different antimicrobial peptides, the host can have control over its symbiotic microbes (Foster *et al.*, 2017).

In addition, the recurrent compartmentalization of the microbial symbionts within their host also enables better control over the interaction (Chomicki *et al.*, 2020b). Fine-scale compartmentalization allows to specifically operate fine-scale partner selection, like demonstrated in the root mycorrhiza where both plants and mycorrhizal fungi actively control the outcome of the interaction (Kiers *et al.*, 2011) or in the gut microbiota where microbes tend to be contained in microenvironments with conditions controlled by the host (pH, nutrients, antimicrobial secretions, etc.; Foster *et al.*, 2017; Chomicki *et al.*, 2020b). In addition, strict compartmentalization of the microbial symbionts is also a way to isolate symbionts and control their reproduction (Chomicki *et al.*, 2020b).

Importantly, partner fidelity feedbacks can also directly promote mutualism by rapidly selecting for transitions from antagonism to mutualism. Such alignments of the host and microbe fitnesses have been demonstrated in parasitic endosymbiotic bacteria that became mutualistic symbionts of *Drosophila* over a few decades (Weeks *et al.*, 2007). Similarly, in mammal guts, some mucin-consuming bacteria (*e.g. Akkermansia*) likely derived from opportunistic saprotrophs (Moran *et al.*, 2019), but their presences are now inversely correlated with obesity in humans (Everard *et al.*, 2013) suggesting that these microbes now play essential roles for their hosts (but see Box 2 on evolutionary addiction).

There is an accumulation of evidence that biological interactions can be seen using a 'market framework' (Selosse & Rousset, 2011; Werner *et al.*, 2014). In such biological marketplaces, multiple-partners mutualistic interactions are maintained stable because there is limited advantages to cheat (partner fidelity feedbacks) and because emerging cheaters are rapidly identified and avoided (partner selection, which is continuously evolving). In addition, market theory (together with greater ecological adaptability) predicts that specialization may not be particularly advantageous in such systems, which can explain why multiple-partner interactions are maintained in animal and plant microbiota (Werner *et al.*, 2014).

3. Analytical tools to study the evolution of host-microbiota interactions

3.1. Characterizing host-associated microbiota

For more than a century, microscopic observations and *in vitro* isolation have been the only way to characterize host-associated microbes. However, such taxonomic characterizations of bacteria and fungi were somewhat limited. Bacteria were classified based on their shapes or the properties of their cell wall (Gram-positive or Gram-negative) and their metabolism *in vitro*, while microscopic mycorrhizal fungi were classified according to their cellular organization (septate or aseptate hyphae) and the structure they form within plant roots (arbuscules, vesicles, coils, mantels…). However, such morphological traits remained very limited to delineate and identify microbial species. In-depth characterization of the microbial communities associated with animals and plants have made substantial progress thanks to advances in molecular biology and DNA sequencing (Hugenholtz, 2002; Heather & Chain, 2016). However, microscopic observations still allow quick and reliable reports of actual host-microbes interactions, like mycorrhizal structures. In addition, combined with recent advances in molecular biology, such as fluorescent in situ hybridization (FISH), microscopic observations allow us to specifically and precisely visualize host-associated microbes, like the hyphae of some specific fungi in unexpected plant species (Schneider-Maunoury *et al.*, 2020).

3.1.1. From microbiota samples to microbial DNA sequences

Following the discovery of the structure of deoxyribose-nucleic acids (DNA) in 1953, several technics to 'read' the DNA sequences have been developed (Heather & Chain, 2016) and can be used to characterize the composition of a microbial community. The idea is to sequence a given core gene, called barcode, that is present in all the microbes, but with a polymorphic DNA sequence, as this DNA region has accumulated mutations during the evolution of these different microbial lineages. Then, by looking at the different DNA barcoding sequences and comparing them to databases of known microbes, we can deduce what are the microbial 'species' present in a given sample: this is called metabarcoding. Several DNA regions can be selected as a barcode, especially among the ribosomal ribonucleic acid (rRNA) operon, as the small subunit (16S rRNA gene) for bacteria and the internal transcribed spacer (ITS) for fungi (Figure 0.3.10).

Metabarcoding requires several steps (Figure 0.3.11): Given the total extracted DNA of a microbial community, the first step is to specifically amplify the barcoding sequences thanks to polymerase chain reactions (PCR). The exact targeted barcode region depends on a match with the primer pair that is used during the PCR and determines the extremities of the amplified region (called amplicon). Second, we have to 'read' the barcoding sequences (amplicon sequencing). One of the first sequencing technologies was Sanger sequencing that allows the 'reading' of a DNA fragment of one thousand base pairs at

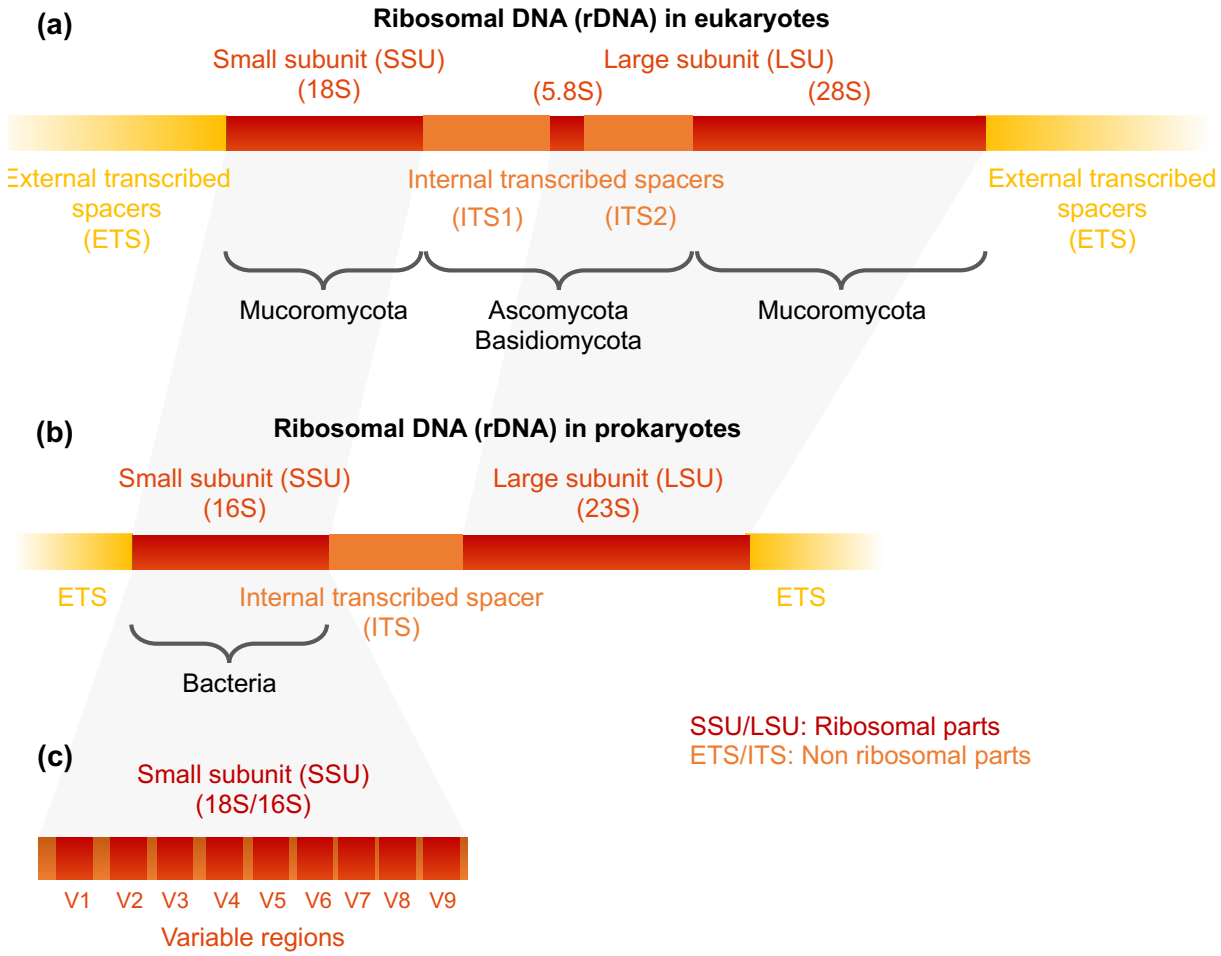


Figure 0.3.10: The ribosomal RNA operon contains traditional barcoding regions for microbes Panels (a) and (b) represent the organization of the rDNA operon in eukaryotes and prokaryotes respectively. The ribosomal DNA operon is composed of a small subunit (SSU - the 16S rRNA gene for prokaryotes and the 18S rRNA gene for eukaryotes) and a large subunit, which is split in two parts in eukaryotes. Panel (c) indicates the 9 variable regions of the SSU gene that alternate with conserved regions. Primer pairs used for metabarcoding generally match with conserved regions at the extremities of one or two variables regions that are thus amplified (*e.g.* the V6 region for prokaryotes).

most. However, Sanger sequencing can only read one unique DNA sequence per sample: therefore, diverse microbial communities cannot be sequenced at once. One solution is to isolate the different microbes before sequencing them, either by separately culturing them (but in that case only culturable microbes are characterized) or by cloning PCR products (*e.g.* Martos et al 2012). Since the beginning of the 21st century, the development of high-throughput sequencing technologies, like pyrosequencing or sequencing by synthesis based on fluorescence (Heather & Chain, 2016), allows in-depth characterizations of the different microbes present in a sample, without cultivation or cloning anymore. However, such methods, like the Illumina MiSeq technology, can only amplify barcodes shorter than 500 bp.

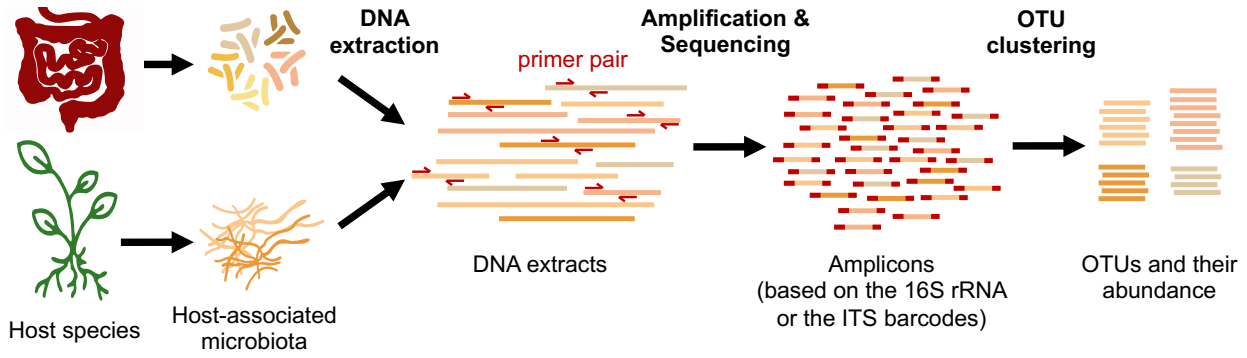


Figure 0.3.11: The different steps of metabarcoding to characterize the composition of a host-associated microbiota. The plant- or animal-associated microbiota have first to be isolated and DNA extraction has to be performed. Next, a PCR has to be performed to specifically amplify the chosen barcode using a specific primer pair and the resulting amplicons can then be sequenced (*e.g.* using the Illumina MiSeq technology). A specific tag is generally inserted into the primers of each sample to be able to recover from which sample each amplicon comes from. Finally, a bioinformatic pipeline is used to process the amplicon reads and cluster them (OTUs).

3.1.2. From DNA sequences to 'microbial species'

An Illumina run generally generates up to 25 million amplicon reads and bioinformatic pipelines need to be used to convert these sequencing data into information about the microbial species present and their abundances (Figure 0.3.11). In short, after filtering the amplicon reads based on their quality to remove PCR and sequencing errors, reads with similar sequences are merged together as they likely come from the same microbial species: the resulting clusters are called operational taxonomic units (OTUs; see Box 5). It exists several OTU clustering methods (Figure 0.3.12): a classical one is to choose a global similarity threshold (*e.g.* 97%) and cluster together all the reads with a sequence similarly larger than 97%. Intra-OTU nucleotide variation is thus assumed to be mainly due to intra-specific variations (microbial differentiation), intra-genomic variations (*e.g.* the rRNA operon is often contained in several copies per genome), or PCR/sequencing errors to a lesser extent. Alternatively, methods like Swarm cluster reads based on the abundances and local thresholds of similarity (Mahé *et al.*, 2014), whereas amplicon sequence variant (ASV) clustering keeps all the unique sequences after stringent quality filtering (Callahan *et al.*, 2017). Besides these phenomenological OTU clustering, one can also consider phylogenetic-based clustering (Box 5 and Figure 0.3.12). Whatever the clustering method used, one sequence is selected to represent each OTU and a taxonomic assignation is performed thanks to global databases censusing all known microbes. For each sample, we can therefore obtain the list of the OTUs that it contained, their read counts (used as a proxy for their abundances in the sample), and their taxonomy (altogether contained into an OTU table).

Box 5: What is a microbial species?

What is a species has been intensively debated for more than a century by biologists and a strong controversy arose because of a mix between the concept of species (what is a species?) and the methods to delineate species (how to put a boundary between species?). De Queiroz (2007) proposed a unified species concept as "separately evolving metapopulation lineages". Reproductive isolation, morphological or ecological dissimilarities, or monophyly are thus only "lines of evidence" (referred to as operational criteria) that can be used to evaluate species delineation and propose species hypotheses (De Queiroz, 2007).

The operational criteria are limited for delineating microbial species. Indeed, as they are mainly asexual organisms (*e.g.* bacteria or Glomeromycotina), the criterium of reproductive isolation (Mayr, 1942) is difficult to apply, and because their morphological variation is often reduced, one cannot reliably use phenotyping to delineate species (Giraud *et al.*, 2008). Thus, microbial species delineation often only relies on DNA sequences from metabarcoding datasets and can be classified in two methods: the phenomenological and the phylogenetic approaches.
First, the phenomenological approaches rely only on current phenotypes (*i.e.* the variation in the DNA barcoding sequence) to cluster the individuals into species hypotheses (referred to as **operational taxonomic units**, OTUs). This includes clustering at a given global threshold (*e.g.* 97% or 99%), a popular approach, or clustering into amplicon sequence variants (ASV). While using a global threshold is arbitrary (and therefore results in arbitrary species hypotheses), ASV might result in over-splitting where single variants likely represent intra-species differentiation rather than different species. Finally, approaches using local thresholds based on abundance profiles (*e.g.* Swarm) have the advantages to be less arbitrary and could thus propose more realistic species hypotheses (Mahé *et al.*, 2014).
Second, the phylogenetic approaches rely on the phylogenetic reconstruction of the evolutionary relationships between all individuals (all unique sequence reads) to propose species hypotheses. One of these approaches, the Generalized Mixed Yule Coalescent (GMYC) model assumes that species diversifications and intra-specific differentiations leave different signals in the time-calibrated phylogenetic tree of all individuals (Pons *et al.*, 2006). Looking at the waiting times between splits in the tree, the GMYC estimates the time t that separates species diversification (birth-death process – before t) and intraspecific differentiation (coalescent process – after t). Therefore, the species hypotheses proposed by the GMYC model do not include any arbitrary threshold. However, it requires a robust time-calibrated phylogenetic tree, which can be difficult or computationally intensive to get. Similarly, the Glomeromycotina are delimited into *virtual taxa* (VT) based on a phylogenetic approach (Öpik *et al.*, 2010): a phylogenetic tree of all the 18S SSU rRNA sequences of Glomeromycotina is built and sequences (called *virtual taxa*) are clustered into OTUs based on a criterium of monophyly and a minimal similarity of 97%.

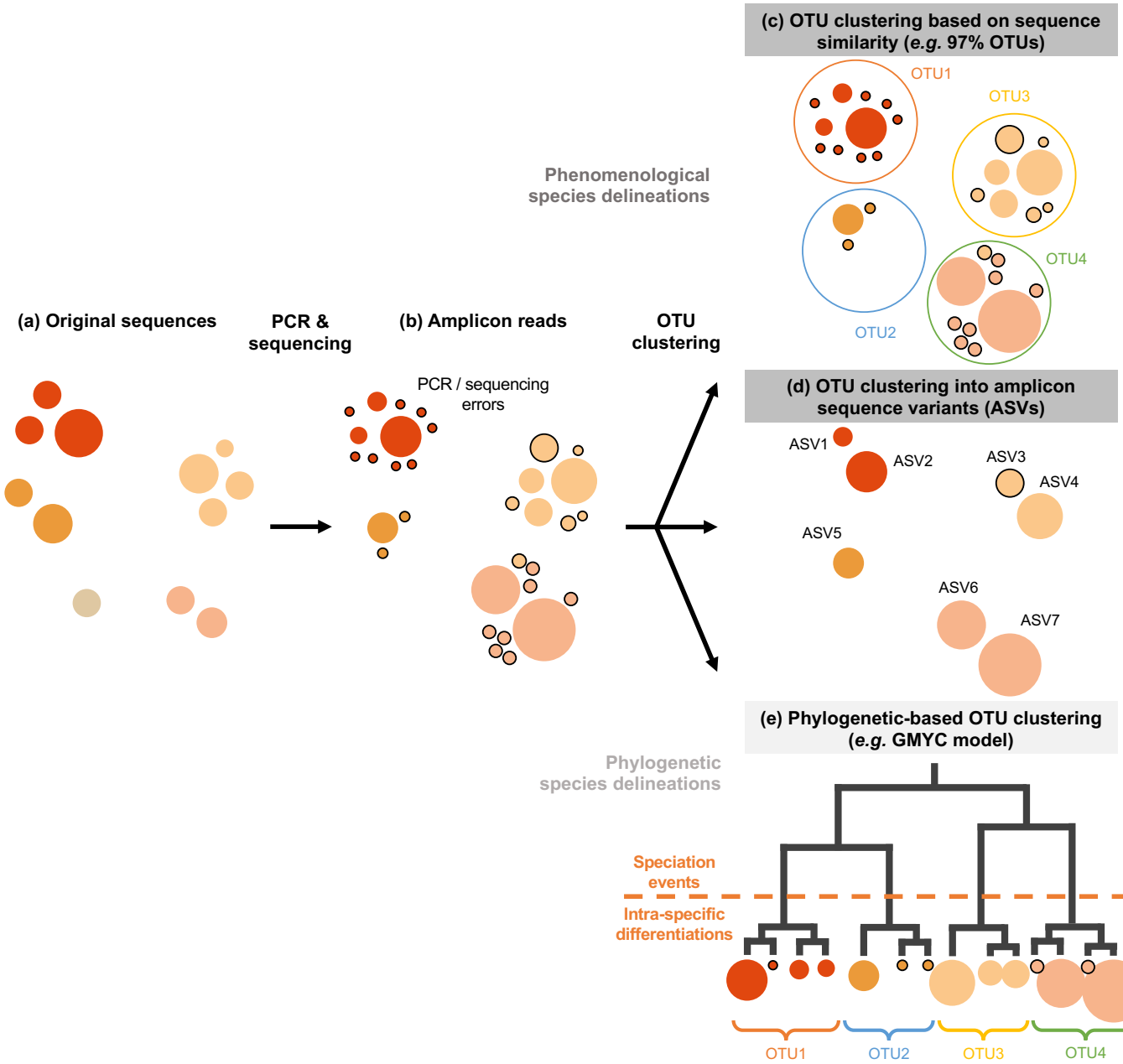


Figure 0.3.12: Clustering the metabarcoding sequences into operational taxonomic units (OTUs). Each colored dot represents a unique barcode sequence. The color of the dot indicates to what microbial species it belongs and its size represents the sequence abundance. Dots that are close to each other indicate that the two corresponding sequences are similar. (a) Original sequences present in the sampled microbial communities. Each dot corresponds to a true biological sequence. (b) After PCR amplification and sequencing, most of the original sequences are still present, but their relative abundances have mostly changed, because of primer biases or noise. However, some of the original sequences have not been amplified (*e.g.* the grey species) and new sequences (dots surrounding by a black circle) have appeared because of PCR and sequencing errors (although they can be numerous, their total abundance generally remains relatively low). All these sequences can be clustered into operational taxonomic units (OTUs) using phenomenological (c-d) or phylogenetic approaches (e). OTU clustering can be performed based on sequence similarity (c), or based on the unique variants (amplicon sequence variants – ASV) after stringent quality filtering to remove most of the errors (d). Alternatively, a phylogenetic tree of all the sequences can be reconstructed (after removing the likely errors) and a phylogenetically-based clustering can be performed, *e.g.* using the Generalized Mixed Yule Coalescent model (GMYC).

Box 5: What is a microbial species? (end)

Importantly, if the DNA barcode is evolving too slowly compared to the speciation dynamics of the microbes (*i.e.* that the barcoding sequences are accumulating substitution at a slower pace than the speciation rate), closely related species might have the same DNA barcoding sequence. In this case, the DNA barcode is improper to delineate species, which highlights the importance of barcode choice to investigate microbial evolution. However, it exists technics to investigate whether a given species delineation is adequate or not: for instance, when using the GMYC model, we can evaluate the support of the threshold model (*i.e.* including intraspecific differentiation) compared to a null model in which all the tips of the tree are assumed to be different species (*i.e.* no intraspecific differentiation).

3.1.3. The limits of metabarcoding

However, metabarcoding has several pitfalls. First, we can only obtain information on the relative abundance of the microbes colonizing a host. Ideally, host-associated microbiota should also be studied in complementary ways, with microscopic observations or using quantitative real-time PCRs to obtain information on the absolute microbial abundances (Hammer *et al.*, 2019). Second, the characterization of the community composition is often biased by the primers used: a primer pair can preferentially amplify some clades of microbes and completely miss other groups due to mismatch in the priming sequences, resulting in an incomplete identification (Mao *et al.*, 2012). Third, the amplification of a short DNA region gives little information on the functional abilities of the microbial community (Louca *et al.*, 2016), especially in prokaryotes. To get a better insight into the functional abilities, metagenomics or metatranscriptomics (*i.e.* direct sequencing of all the DNA or RNA of the whole microbial community) can be used. Fourth, having a short DNA region (<500 bp) that is slowly evolving (to be sufficiently conserved in all bacteria or fungi) is often limited to robustly investigate the evolutionary history of the different microbes (see Box 5). The recent development of new sequencing technologies, like Nanopore, that sequence longer DNA fragments up to several thousand of base pairs will likely improve our ability to reconstruct the evolution of these host-associated microbiota.

Metabarcoding characterization of the host-associated microbiota can be performed for multiple host species in a given clade of animals or plants (Figure 0.3.13) and several analyses can be performed to investigate the evolution of these host-microbiota interactions. First, one can be interested in looking at the evolution of individual microbial lineages (OTU per OTU; Figure 0.3.13a): cophylogenetic analyses can be used in this case (see section 3.2). Second, one can be interested in studying the evolution of the microbiota as a whole (Figure 0.3.13b) thanks to network analyses (see section 3.3). A network approach can for instance enable to investigate patterns of phylogenetic signals in species

interactions (phylosymbiosis; see section 3.3) or patterns and drivers of species diversification (see section 3.4). The next sections present the available tools for studying these different aspects of host-microbiota evolutions.

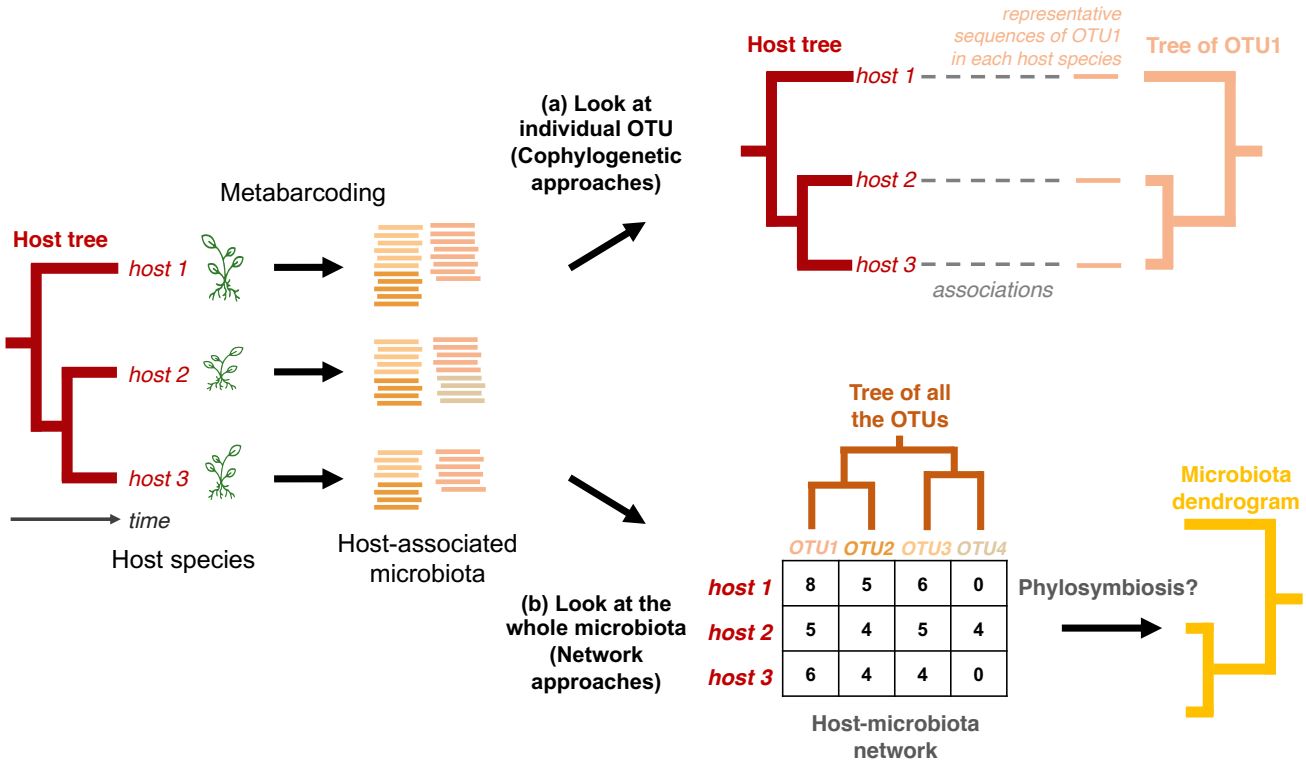


Figure 0.3.13: Investigating the evolution of host-microbiota interactions. The evolution of microbiota associated with a given clade of animal or plant hosts can be analyzed by looking independently at the different OTUs shared across host species using cophylogenetic approaches (a) or at the microbiota as a whole using a bipartite network approach (b). Importantly, we represent here three "microbial trees" that are fundamentally different: (i) for a given OTU (*e.g.* OTU1), one can get the representative sequence of the OTU in each host species and reconstruct the tree of the sequences belonging to OTU1 (this tree, therefore, represents 'intra-specific' differentiation within the OTU), (ii) the tree of all the OTUs (a 'species-level' phylogenetic tree with one representative sequence per OTU), and (iii) the 'microbiota dendrogram', that is obtained by performing a hierarchical clustering of the host-associated microbiota based on their dissimilarity of composition (this dendrogram is generally used to quantify phylosymbiosis).

3.2. Detecting microbial transmissions over macro-evolutionary time scales

An important aspect when studying host-microbe evolutions is to determine whether the associated microbes are transmitted over long-time scales or not (see section 2.1; Figure 0.3.13a). **Cophylogenetic methods** were originally developed to compare the evolutionary histories of both hosts and symbionts, with the underlying idea that vertical transmissions along host lineages lead to concomitant divergences, and thus congruent phylogenies with similar topologies and divergence times (de Vienne *et al.*, 2013), while processes such as microbial horizontal transmissions (host-switches), extinctions, or du-

plications disrupt this congruence (Figure 0.2.7). These cophylogenetic approaches are particularly well-adapted to infer the mode of inheritance of the symbionts when two robust phylogenetic trees of the hosts and the symbionts are available (see Box 6). However, when studying host-microbiota interactions, obtaining robust phylogenetic trees of the microbial symbionts is often challenging: the DNA sequences from barcodes, like the bacterial 16S rRNA gene, are generally too short and conserved to allow a robust reconstruction of the microbial phylogenetic trees when transmitted microbes diverged for less than a few million years. Indeed, the 16S rRNA gene for instance is expected to accumulate 1% of nucleotide divergence per 50 million years (Ochman *et al.*, 1999).

Existing cophylogenetic methods can be roughly classified into two categories (de Vienne *et al.*, 2013): the global-fit methods (that test whether there is an overall significant congruence between the host and the symbiont evolutionary histories) and the event-based methods (that fit evolutionary events to reconciliate the host and the symbiont phylogenies). The significance of the cophylogenetic congruence (correlative methods) or the reconciliated scenario (event-based methods) is then generally evaluated by comparing it to null expectations obtained by randomizing the host-symbiont associations. Here, we mainly present the methods that consider uncertainty in symbiont phylogenies and that are thus well-adapted for investigating the transmission of host-associated microbes.

3.2.1. Global-fit methods

First, global-fit methods can be separated between topology- and distance-based methods (de Vienne *et al.*, 2013). While the topology-based methods require to have a robust phylogenetic tree for the symbionts, distance-based methods are more flexible. For instance, Mantel tests, which were originally developed to test if there is a correlation between two dissimilarities matrices (Mantel, 1967), have been applied to test the correlation between host and microbial genetic or patristic distances. Given two dissimilarity matrices, their elements are first standardized (by subtracting their mean and dividing them by their standard deviation) and the Mantel statistics (R) corresponds to the mean of the products of the corresponding elements in the two matrices (Pearson correlation). Alternatively, if the correlation is not suspected to be linear, the dissimilarities can be first transformed into ranks (Spearman correlation). The null hypothesis states that the two dissimilarity matrices are not expected to be correlated: the significance of the observed correlation (R) is therefore tested using randomizations that permute the rows and the corresponding columns of one of the matrices. However, Mantel tests can only be used if there are one-to-one host-symbiont associations, although a more recent extension allows considering multiple microbes per host (Hommola *et al.*, 2009).

Box 6: Reconstructing phylogenetic trees using DNA sequences

Molecular phylogenetics deal with the reconstruction of phylogenetic trees using DNA sequences (Felsenstein, 2004). In short, to reconstruct the phylogenetic tree of different organisms, their corresponding DNA sequences have first to be aligned, such that each nucleotide site in the alignment corresponds to a homology. Second, one has to assume a model of DNA evolution and reconstruct the phylogenetic relationships between the species that best fit this model given the aligned sequences. Besides parsimony and distance-based methods that enable quick phylogenetic reconstruction, probabilistic methods have been developed and are mainly used to get more robust phylogenetic trees (Felsenstein, 2004). In short, assuming a probabilistic model of DNA substitutions occurring along the branches of the tree, Felsenstein pruning algorithm allows to easily compute the likelihood of a tree given the observed nucleotide alignment at present (Felsenstein, 1981). Phylogenetic algorithms are then designed to explore the tree space and output the most likely phylogenetic tree (maximum likelihood estimation) or an *a posteriori* distribution of likely phylogenetic trees (Bayesian inference). Importantly, probabilistic methods infer both the tree topology and the branch lengths. By default, the branch length unit is in a number of substitutions per nucleotide site. Time-calibrated trees (or ultrametric trees) can be obtained by converting the number of substitutions into a relative time given by a molecular clock. In addition, an extragroup of sequences is often used to root the tree, *i.e.* to determine the most recent common ancestor of the organisms of interests.
Finally, fossils can be used to calibrate the trees in an absolute time or to constraint the monophyly or the age of certain clades during the phylogenetic reconstruction. However, fossils are generally not abundant for bacteria or fungi, and often difficult to robustly identify based on their morphology.

Reconstructing the species evolutionary history with a single gene can be problematic. Indeed, species trees and gene trees can be incongruent because of gene evolutionary events (*e.g.* horizontal gene transfers, gene duplications, or losses) or incomplete lineage sorting, resulting in different topologies (Szöllősi *et al.*, 2015). In addition, a gene contains only a limited number of segregating sites, meaning that they contain a limited amount of information to reconstruct the tree: identical likelihoods can be obtained for different tree topologies if the number of segregating sites is too low, resulting in large uncertainty in the reconstructed tree. This is particularly problematic if one wants to reconstruct the recent evolution of a slowly evolving barcode gene like the SSU rRNA gene.

Next, like Mantel tests, the ParaFit test has also been developed to test whether closely related symbiont species tend to interact with closely related host species, but while incorporating the possibility of multiple symbionts per host (Legendre *et al.*, 2002). It first transforms the host and symbiont phylogenetic distance matrices using principal coordinates analyses to obtain the matrices H and S respectively (where each row of a matrix corresponds to the principal coordinate decomposition of a given species), and then computes the matrix D as the matrix product $S'MH$, where M is the matrix of interactions (with hosts on columns and symbionts on rows) and S' denotes the transpose of S. The ParaFit statistics is then the trace of $D'D$ and its significance is calculated by randomly permuting the values within each row of M. Contrarily to Mantel tests, ParaFit is not symmetrical: the null hypothesis states that each symbiont species interacts with any host species independently from the host evolutionary history (Legendre & Legendre, 2012).

Finally, a procrustean approach to cophylogeny (PACo) has been proposed (Balbuena *et al.*, 2013) to test the dependence of the symbiont phylogeny on the host phylogeny. Like ParaFit, PACo first transforms the host and symbiont phylogenetic distance matrices using principal coordinates analyses to obtain the matrices H and S respectively. Second, given M, the symbiont matrix S is rotated and scaled using a Procrustes analysis to fit the host matrix H (*i.e.* to minimize the squared differences of the superimposition). The null hypothesis states that the host phylogeny does not predict the parasite ordination and null expectations are obtained by permuting the columns of M (Balbuena *et al.*, 2013). Like ParaFit, PACo also allows multiple symbionts per host and *vice versa*. Built upon these global-fit methods, Random TaPas (Balbuena *et al.*, 2020) has been recently developed to additionally identify the individual host-symbiont interactions, the species tips, and the nodes that mostly contribute to the overall congruence.

Importantly, all these global-fit methods do not account for phylogenetic non-independence (de Vienne *et al.*, 2013), as closely related pairs of species have the same weight as distantly related ones. MRCALink has been developed to tackle this problem (Schardl *et al.*, 2008), however, it requires binary ultrametric trees, which limits its uses for investigating microbial transmissions based on metabarcoding datasets.

These correlative methods have been used several times for investigating microbial transmission in host clades having their microbiota characterized with metabarcoding datasets (*e.g.* Youngblut *et al.*, 2019; Amato *et al.*, 2019). However, we currently lack a comparative analysis of the relative strengths and weaknesses of these different global-fit methods when applied on host-microbe datasets with very little resolution in the microbial tree.

3.2.2. Event-based methods

Second, many event-based methods have been developed to infer the evolutionary events (*e.g.* horizontal transmissions, duplications, or extinctions) that can explain the loss of congruency between the host and symbiont phylogenies (de Vienne *et al.*, 2013). These methods differ according to the types of evolutionary events they consider, their ability to consider multiple associations per species, and their inference technics (parsimony, cost-based methods, Bayesian inference…). Many of these approaches, like TREEMAP (Pagel, 1994) require robust phylogenies, but a few of them also consider phylogenetic uncertainty. For instance, Huelsenbeck *et al.* (2000) developed a Bayesian approach that jointly reconstructs the phylogenetic trees of the hosts and symbionts while simultaneously fitting host-switching events. In addition, methods originally developed to reconciliate gene trees and species trees can also be used in the context of symbiotic transmissions (Bailly-Bechet *et al.*, 2017). For instance, the amalgamated likelihood estimation (ALE) approach models lateral gene transfers, gene duplication, and gene losses that can happen during the species evolution (Szöllősi *et al.*, 2013a,b). By extension, ALE can be used to model microbial horizontal transmissions, duplications, or extinctions on the host phylogeny (Bailly-Bechet *et al.*, 2017). In short, ALE takes as input a robust phylogenetic tree of the hosts and a distribution of microbial phylogenies (obtained using Bayesian phylogenetic reconstruction) and estimates by maximum likelihood (or using Bayesian inference) the rates of horizontal transmissions, duplications, and extinctions. ALE then generates reconciled scenarios of host-microbe evolutions, and for each scenario, outputs the number of co-divergences, horizontal transmissions, duplications, and extinctions. In a recent application investigating transmitted microbes in the mammalian gut microbiota, Groussin *et al.* (2017) considered that a microbe has been vertically transmitted if the difference between the estimated number of co-divergences and the estimated number of horizontal transmissions is positive and larger than the differences obtained with ALE when randomizing host-microbe associations. ALE does not consider branch lengths (it only accounts for tree topologies), and it has not been tested when the number of segregating sites in the microbial sequences is very low (*i.e.* when there is a lot of phylogenetic uncertainty in the microbial trees).

Therefore, a range of methods to look at symbiont transmissions in a host clade is already used to investigate microbial transmissions among host-associated microbiota. While event-based methods offer a complete understanding of the host-microbe evolutionary histories, global-fit approaches indicate only whether there is a significant congruence and therefore do not provide mechanistic details about the evolution of host-microbe interactions. Moreover, it is unclear how these methods deal with the low number of segregating sites in the microbial barcoding sequences (*i.e.* limited amount of information on the microbial evolutionary history).

3.3. Representing and analyzing host-microbe interactions using bipartite networks

Host-microbiota interactions are often multiple-partner interactions: host species regularly interact with a large number of microbes, and the microbial 'species' (OTUs) can be shared between host species (generalist microbes) or only present in a few hosts (specialist ones). These interactions can be represented using a **bipartite network**, where hosts and microbes are nodes and interactions are links between nodes. A bipartite network can be simply visualized using a matrix, with host species on columns and microbial OTUs on rows (Figure 0.3.14). Binary (or unweighted) networks only carry the information of the presence (1) or absence (0) on an interaction, whereas quantified (or weighted) networks inform each link with the abundance of the corresponding host-microbe interaction (*e.g.* the number or the proportion of reads of a given OTU in a host species).

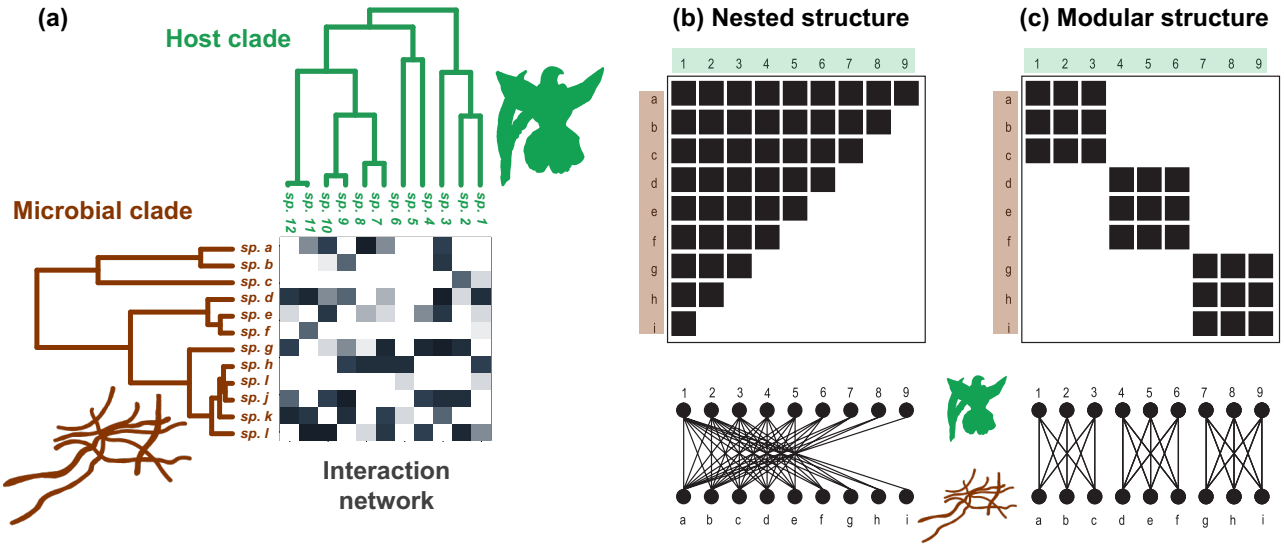


Figure 0.3.14: Examples of network representations of host-microbiota interactions. (a) Interaction network between orchids (in green) and mycorrhizal fungi (in brown) with associated phylogenetic trees. The bipartite interaction network is represented by a matrix with orchids species on columns and fungal species on rows and which elements indicate the frequency of the interaction using shades of grey from white (no interaction) to dark grey (many interactions). Panels (b) and (c) illustrate examples of perfectly nested (b) and modular (c) networks, represented either as a matrix (top) or as a graph with species being the nodes and interactions the links between nodes (bottom; Figures modified from Fontaine *et al.* (2011)). For each species in the network, its degree corresponds to its number of partners, and species with a high (resp. low) degree are generalist (resp. specialist): nested networks are characterized by asymmetrical specializations, whereas modular networks present reciprocal specializations. The connectance of the networks, defined as the ratio between the number of observed interactions and the total number of possible interactions, is higher in the nested network (0.5) than in the modular one (0.33).

Importantly, converting host-microbiota interactions into a species-level bipartite network is an overall simplification as we merge altogether interactions with different physiological and ecological importance, but this reductionist approach has allowed us to

better understand the patterns and processes behind multiple-partners interactions (Jordano, 2010; Guimarães, 2020). Epistemologically, it is an important step toward considering the interactions rather than the individuals and species. In addition, bipartite networks can be used at different temporal (punctual or not) and spatial scales, from the scale of a community in a given plot (local scale) to the regional (ecosystem scale) or the global scale. According to its scales, an interaction network does not convey the same type of information. The absence of interactions between two species can be due either to a lack of detection (metabarcoding technics can miss some OTUs in some samples) or the fact that both species are actually never interacting because of constraints, *e.g.* temporal or spatial mismatches, physiological or morphological incompatibilities, or evolutionary inertia (Bascompte & Jordano, 2013).

3.3.1. Analyzing the structure of interaction networks

Well before being applied to host-microbiota interactions, bipartite networks have been mostly developed to study interspecific interactions between macroorganisms, such as plant-pollinator interactions, seed dispersals, herbivory, or host-parasite interactions (Bascompte & Jordano, 2013). It rapidly appeared that the structures of these empirical networks could be roughly divided into two categories of networks: the nested networks (where specialist species tend to interact with generalist partners, and generalist species form a core on interactions; Bascompte *et al.*, 2003) and the modular networks (where several subsets of species, called modules, tend to preferentially interact with each other rather than with species from other modules; Olesen *et al.*, 2007). In terms of specialization, asymmetrical specialization (resp. reciprocal specialization) is frequent in nested (resp. modular) networks.

One way to compute the nestedness of a network is to use the Nestedness metric based on Overlap and Decreasing Fill (NODF; Almeida-Neto *et al.*, 2008) and modularity can be obtained using one of the many available algorithms to maximize the modular structure of a network (Beckett, 2016). The obtained values for nestedness and modularity are generally difficult to compare in an absolute way, thus to determine whether a network is significantly nested or modular, one has to compare these values to null expectations (null models) obtained by randomizing the interactions in the network (Gotelli, 2000). Null models are often chosen to test whether a particular process of interest can generate similar structures (*e.g.* if interactions are shuffled based on species abundances, does it generate networks as nested as the original one?). Besides global metrics, such as connectance, nestedness, and modularity, simple patterns of interactions between 2 to 6 species (referred to as bipartite motifs) offer also more insights into the direct and indirect interactions between species (Simmons *et al.*, 2019). Indeed, these "building blocks" of the network can be used either (i) to study the position of a given species in the interaction network or (ii) to compute the frequencies of the different motifs and thus compare networks (Simmons *et al.*, 2019).

By comparing mutualistic and antagonistic networks, it appeared that overall, mutualistic networks tend to be nested whereas antagonistic networks are rather modular (Thébault & Fontaine, 2010). Such a dichotomy is rather widespread in empirical networks, although the intimacy of the interactions or the proportion of realized links in the networks (connectance) can complicate the picture (Fortuna *et al.*, 2010; Fontaine *et al.*, 2011). In addition, it has been shown that a pollination network dominated by cheating insects was modular, while a similar network dominated by mutualistic pollinators was nested (Genini *et al.*, 2010), suggesting that the emergence of cheaters might disturb the nested structure of mutualistic networks. Structural analyses of host-microbiota interactions have also been performed, in particular in mycorrhizal networks. It has been found that networks between plants and arbuscular mycorrhizal fungi (Glomeromycotina) are rather nested (Montesinos-Navarro *et al.*, 2012; Sepp *et al.*, 2019), but can also contain a certain level of modularity (Chagnon *et al.*, 2012). Network structures seem more variable in other mycorrhizal systems like the ectomycorrhiza or the orchid mycorrhiza, ranging between strong nestedness and clear modularity (Jacquemyn *et al.*, 2011; Martos *et al.*, 2012; Toju *et al.*, 2014; Põlme *et al.*, 2018). Importantly, network structure is influenced by the scale of the studied community: networks at the local-scale, regional-scale, or global scale, or network focusing only on a subset of interactions in a community will likely have different structures.

Several ecological and evolutionary processes have been proposed to explain the differences in structural patterns between mutualistic and antagonistic networks (Fontaine *et al.*, 2011). For instance, one explanation relates to community stability (Thébault & Fontaine, 2010): in mutualistic communities, nested structures may minimize the competition between species and facilitate the integration of new species, while in antagonisms, interspecific competitions may limit the sharing of partners. Another explanation states that nested patterns in mutualistic networks may be due to convergence and complementarity of traits between interacting species (Thébault & Fontaine, 2008; Maliet *et al.*, 2020), whereas strong selective pressures and coevolutionary arms race in antagonisms (or in mutualism with high intimacy) may lead to network compartmentalization by reciprocal adaptation of the partners (Thompson, 2005; Guimarães *et al.*, 2007). However, it remains difficult to directly link structural patterns in interaction networks to clear processes explaining their assembly or their evolution, especially in the case of host-microbiota networks (Chagnon, 2016).

3.3.2. Investigating how evolution has influenced interaction networks

A bipartite network can also be informed with the host and symbiont phylogenetic trees to investigate the evolution of host-microbiota interactions. This can be done by looking at how current patterns of interactions can be explained by the host and symbiont evolutionary histories.

First, one can use correlative approaches, like Mantel tests (see section 3.2), to mea-

sure the phylogenetic signal in the interactions, *i.e.* do closely related host species tend to interact with similar microbial symbionts (phylosymbiosis) and *vice versa*? Mantel tests can be computed between one phylogenetic distance matrix (*e.g.* for the hosts) and a matrix comparing the dissimilarity of the sets of microbial partners interacting with pairs of host species. Such dissimilarities generally correspond to beta diversity metrics, like Jaccard or UniFrac distances; the former being based on only the sharing of partners between species, while the later also considers the phylogenetic distances between partners (Lozupone & Knight, 2005). Phylogenetic signals in host-microbiota interactions are frequently measured using Mantel tests (Jacquemyn *et al.*, 2011; Groussin *et al.*, 2017; Song *et al.*, 2020; Armstrong *et al.*, 2020) and they directly allow to test for phylosymbiosis. An alternative method to measure phylosymbiosis is to perform a hierarchical clustering of the microbiota composition (based on their beta diversities) to obtain a dendrogram, and then to assess whether this dendrogram recapitulating microbiota differentiation tends to (qualitatively or quantitatively - using the Robinson-Foulds metric or the matching cluster metric) mirror the host phylogeny (Lim & Bordenstein, 2020).

Second, a few model-based approaches of network evolution have been developed by assuming that species interactions are influenced by species traits. For instance, the phylogenetic bipartite linear model (PBLM; Ives & Godfray, 2006) assumes that host-symbiont interactions are mediated by a host trait and a symbiont trait evolving according to an Ornstein-Uhlenbeck process of trait evolution and that the probability of interaction between a host and a symbiont is proportional to the product of their trait values. By fitting this model to an empirical network (with both the host and microbe phylogenies), one can get a measure of phylogenetic signal indicating the extent to which host-microbe interactions are conserved (Jacquemyn *et al.*, 2011; Martos *et al.*, 2012). Several extensions of this model have been proposed (Rafferty & Ives, 2013; Hadfield *et al.*, 2014), but we currently lack a comparative analysis of their advantages and weaknesses.

Third, besides measures of phylogenetic signals, one can be interested in how host-microbe interactions are acquired or lost. For instance, a few approaches have investigated how the modes of symbiont diversification (*e.g.* radiation after host switches) can affect extant patterns of network structure and interaction specificity (Braga *et al.*, 2018; Jousselin & Elias, 2019). More recently, a model aiming at reconstructing the ancestral host repertoire of the symbionts has been developed (Braga *et al.*, 2020) and it represents the first step to infer ancestral host-microbiota networks.

3.4. Quantifying the effect of biotic interactions on species diversification

Quantifying the effect of the associated microbes on host diversification (and *vice versa*) can be done using models of species diversification. Usually, species diversification is modeled using **birth-death processes** (Nee, 2006; Figure 0.3.15). We first present the

birth-death models, then examine how they can be used to test the effect of biotic interactions, and finally discuss the challenges when applying such models to host-microbiota datasets.

3.4.1. Birth-death models of species diversification

Under a homogenous constant-rate birth-death model, a lineage is assumed to have a constant probability to speciate and a constant probability to go extinct, with λ and μ being the per-lineage speciation and extinction rates respectively. Given a reconstructed time-calibrated phylogenetic tree of the extant species, one can fit this model and estimate (by maximum likelihood or using Bayesian inference) the parameters λ and μ (Nee, 2006). Intuitively, these parameters are estimated by using the information contained in the lineages-through-time (LTT) plot representing the logarithm of the number of lineages in the reconstructed tree as a function of time. Under a homogenous constant-rate birth-death model, the slope of the LTT close to the present equals the speciation rate (λ), whereas in the past, it equals the net diversification rate ($\lambda - \mu$) (Figure 0.3.15). However, a reconstructed tree often does not include all the existing species of a clade, as only a fraction of them (ρ) has been sampled, but the three parameters λ, μ, and ρ are unidentifiable (*i.e.* several combinations of parameter values have the same likelihood). Thus, one has to first estimate ρ (using different approaches) to infer λ and μ (Morlon *et al.*, 2010).

The assumption of constant rates can then be relaxed and one can assume that both speciation and extinction rates are piece-wise constant (Stadler, 2011) or vary as a continuous function of time, $\lambda(t)$ and $\mu(t)$ (Morlon *et al.*, 2011). For instance, $\lambda(t)$ can be a linear or an exponential function of time (which can successfully model the slowdown of diversification rates frequently observed close to the present (Moen & Morlon, 2014)). Alternatively, we can assume that speciation and extinction rates are functions of an environmental variable that varies through time (*e.g.* global temperature; Condamine *et al.*, 2013) or depend on the species diversity of the clade (Rabosky & Lovette, 2008). However, under a homogenous time-varying birth-death model and in the absence of any hypothesis on the functional form of $\lambda(t)$ and $\mu(t)$, one wishes to estimate two variables $\lambda(t)$ and $\mu(t)$ from the LTT plot that is a single function of time, which is therefore not asymptotically identifiable (Lambert & Stadler, 2013, Louca & Pennell, 2020). In other words, there is not enough information in a reconstructed phylogenetic tree of the extant species to recover the rates $\lambda(t)$ and $\mu(t)$ without further hypotheses on the functional form of $\lambda(t)$ and $\mu(t)$. Thus, rather than estimating the 'true' $\lambda(t)$ and $\mu(t)$, one can only test whether there is or not some support for a given scenario of diversification, represented by hypotheses on $\lambda(t)$ and $\mu(t)$ (*e.g.* is there support for a diversification slowdown represented by an exponential decline? is their support for a linear association between past temperature and the diversification rates of the clade?; Morlon, 2014). The support of the different models can then be compared using model selection (*e.g.* based on corrected Akaike Information Criterion (AICc)) in a hypothesis-driven frame-

work (Morlon *et al.*, 2020). More recently, it has been proposed to rather estimate "pulled rates" of speciation and diversification, which are identifiable from the reconstructed phylogenies of extant species (Louca & Pennell, 2020), but are more difficult to interpret.

Besides homogenous birth-death models, one can consider models of diversification where there are potential shifts in the rates of diversification occurring at some speciation events (Morlon *et al.*, 2011; Rabosky, 2014), or models where shifts occur at each speciation event (Maliet *et al.*, 2019). The latter model, ClaDS, assumes that each lineage has its own speciation rate, that the parental speciation rates are transmitted at speciation events to the two daughter lineages with a small stochastic variation (σ) around the parental speciation rates multiplied by a general trend (α). In addition, the model ClaDS2 assumes that the turnover (ϵ), the ratio between extinction and speciation rates, is constant. Importantly, contrary to the homogenous birth-death models under which all tree topologies are equally likely (and therefore only the waiting times between speciation events in the reconstructed tree, *i.e.* the LTT plot, are informative), models with heterogenous rates across lineages predict different tree topologies, and therefore tree topology is useful for distinguishing different models.

Finally, there is a range of models (referred to as State Speciation Extinction, SSE models) that specifically test the effect of particular traits on species diversification. For instance, the binary-state speciation and extinction (BiSSE) model assumes that speciation and extinction rates depend on the state of a given trait of the lineages that can be in two states (0 and 1) (Maddison *et al.*, 2007). Trait states are supposed to evolve according to a continuous-time Markov process, and transition rates and state-specific diversification rates can be jointly estimated. Extensions include testing for the effect of multiple discrete traits, continuous traits, or geographical distributions, while adding hidden states to avoid biased model selections (Caetano *et al.*, 2018).

3.4.2. Investigating the effect of biotic interactions on species diversification

One can envisage several ways to test for the effect of biotic interactions on species diversification. For instance, let's assume that one wants to test fungal lineages that evolved mycorrhizal symbiosis with plants diversified faster or slower than the lineages that remained saprotrophs. First, one can consider separately the mycorrhizal fungal clades or the saprotrophs clades and fit homogenous constant rates birth-death model to see if mycorrhizal fungal clades have significantly higher diversification rates than saprotroph ones. In addition, one can fit environment-dependent birth-death models to test whether past land plant diversity could have influenced the tempo of diversification of mycorrhizal fungi, whereas saprotroph fungi could have experienced other drivers. Second, one can consider all the fungi together, estimate their lineage-specific speciation rates (using ClaDS), and test whether mycorrhizal fungal species have higher speciation rates (at present) than non-mycorrhizal ones. Alternatively, one can use SSE models, to directly test whether diversification rates are actually different according to these traits

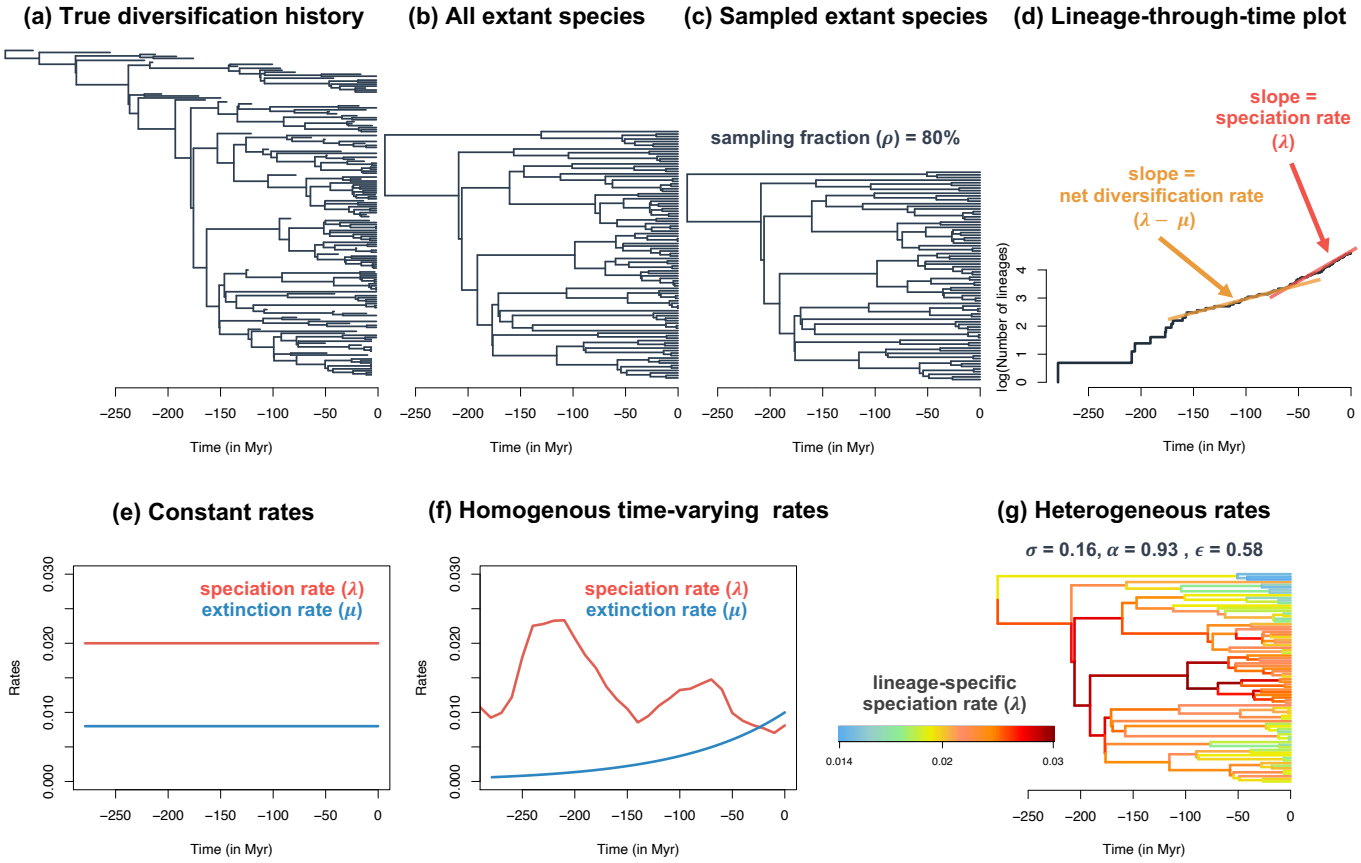


Figure 0.3.15: Modeling diversification rates using time-calibrated reconstructed phylogenies of extant species. (a) Phylogenetic tree simulated under a homogenous birth-death model with constant rates (speciation rate $\lambda = 0.02$ and extinction rate $\mu = 0.008$). (b) Phylogenetic tree of only extant species (lineages that did not leave any extant descendants are absent from this phylogeny). (c) Phylogenetic tree of only sampled extant species (80% of the extant species are sampled here); this phylogeny corresponds to the reconstructed phylogenetic tree obtained from the sampled species at present. (d) Lineage-through-time (LTT) plot of the complete reconstructed phylogeny (b): under a homogenous birth-death model, the slope of the LTT equals the speciation rate close to the present and the net diversification rate far in the past. (e) Constant rates used to simulate the phylogenetic tree in (a). (f) Example of homogenous time-varying rates that can be fitted to the reconstructed phylogenetic trees. Here, for illustration, $\lambda(t)$ is assumed to be an exponential function of the average global temperature and $\mu(t)$ an exponential function of time. (g) Example of lineage-specific speciation rates as assumed by the model ClaDS (Maliet *et al.*, 2019). The colors of the tree branches indicate the speciation rates. Simulations and plots have been performed in R (R Core Team, 2020) using the packages phytools (Revell, 2012) and RPANDA (Morlon *et al.*, 2016).

(mycorrhizal or saprotroph).

3.4.3. Challenges when applying birth-death models to microbial clades

If applying birth-death models to animals or plants can be 'relatively' straightforward, using them on microbes to investigate their diversification history is generally more challenging, especially if these microbes have been characterized using metabarcoding technics. Indeed, birth-death models require correct species delineations, robust phylogenetic trees, and good estimations of the sampling fraction. First, species delineation in many microbial groups is often done using operational taxonomic units (OTUs) without convincingly showing that these units correspond to 'biologically-relevant species' (see Box 5). Over-splitting the species would result in an artefactual increase of the speciation rates toward the present, whereas over-merging them would produce an artefactual slowdown (Moen & Morlon, 2014). Given that estimates of diversification rates strongly rely on species delineation, one should test several species delineations (*e.g.* with different similarity thresholds in the OTU delineations). Second, phylogenetic trees reconstructed using a single DNA region that is slowly evolving (*e.g.* the SSU rRNA gene) generate a lot of phylogenetic uncertainty, and conversely, a fast evolving region like the ITS generate sequence homoplasy. Running the diversification models on several of the alternative evolutionary histories is therefore essential to consider this uncertainty (Lewitus *et al.*, 2018). Finally, the total diversity of the group should be known, which is generally not the case in microbial clades. This total diversity can either be estimated using various methods (like rarefactions or model-based methods; Quince *et al.*, 2008) or diversification models can be replicated for a range of sampling fractions (Morlon *et al.*, 2012). Therefore, applying diversification models on microbial clades is possible but requires additional validations to guarantee the robustness of the results.

4. Goals of the PhD

A plethora of recent studies have characterized the compositions and the functions of microbiota across various clades of host plants and animals. Enough data have thus been collected across a broad range of different systems to investigate the question of the dynamics and conservatism of the host-microbiota association over evolutionary timescales. The overall goal of my PhD is to advance our understanding of how microbiota evolve with their host species. The data we have in hands or have generated comprise robust phylogenetic trees of the animal or plant host species for which the associated microbiota have been characterized using metabarcoding technics targeting the microbial SSU rRNA gene or the ITS region. For each host, its associated microbiota is described as a list of DNA sequences clustered into operational taxonomic units (OTU; Figure 0.3.13). We developed new quantitative tools, collected data, and performed a series of analyses, all directed to the common overarching goal of better characterizing the evolution of host-microbiota interactions. We considered both microbiota-animal and microbiota-plant systems, with a specific focus on mycorrhizal interactions.

In **Chapter I**, we focused on quantifying the prevalence of vertically transmitted microbes among the microbiota of a clade of host (Figure 0.3.13a). We developed HOME, a quantitative approach for inferring the modes of microbial inheritance as host clades diversify. Given a host phylogeny and the microbiota of present-day species, our approach uses nucleotidic variability within OTUs to detect the OTUs that are vertically transmitted. We applied this model to two systems, the gut microbiota of primates (**Article 1**) and a clade of Hawaiian spiders (**Article 2**), in order to evaluate the prevalence of vertical transmissions in microbiota evolution across the animal kingdom. Finally, we compared the performances of HOME to other available approaches (**Article 3**).

In **Chapter II**, we examined the interplay between the evolutionary histories of host and host-associated microbial clades (Figure 0.3.13b). We focus on two specific questions: "To what extent does evolutionary history influence which microbial species interact with which host species?" and "How does the evolutionary history of hosts influence the diversification of host-associated microbial clades?". The first question leads us to compare different methods for estimating phylogenetic signals in host-microbiota interaction networks, *i.e.* whether closely related species share similar sets of partners, with an application on plant-mycorrhizal interactions (**Article 4**). We explore the second question by studying the diversification of the arbuscular mycorrhizal fungi (Glomeromycotina) in the past 500 million years and evaluating how land plants might have affected the diversification of these obligate mycorrhizal symbionts. (**Article 5**).

In **Chapter III**, we focused on the evolution of cheating in host-microbiota mutualism, by taking the mycorrhizal symbioses as a case study. We first explored the constraints upon the evolutionary emergence of cheating in plants (mycoheterotrophy) by

analyzing patterns of plant-mycorrhizal fungus interactions at the global scale (**Article 6**). Then, we investigated whether similar constraints were found in local mycorrhizal networks including initially mycoheterotrophic plants (Lycopodiaceae) that we sampled in La Réunion island (**Article 7**).

We finally discuss how such computational tools, in combination with metabarcoding sequencing data, allow studying how microbiota evolve with their hosts. We consider in particular the challenges and promises of this comparative approach to apprehend host-microbiota evolution.

Chapter I.

Characterizing the inheritance of microbial units on a host phylogeny:

Whether microbial symbionts are transmitted during host evolution is a central question as it has important consequences on both hosts and microbes, as vertical transmission can guarantee a stable association, allow the evolution of dependences, and limit the propensity to cheat. In this chapter, we focused on the detection of vertically transmitted microbes among the microbiota of a clade of hosts (Figure 0.3.13a). We developed a probabilistic model (HOME) to independently infer via maximum likelihood the evolutionary history of each microbial OTU constituting the microbiota. In short, given a host phylogeny and the microbiota of present-day species, our approach uses nucleotidic variability within OTUs to detect which OTUs are vertically transmitted. The model takes as inputs the host phylogeny and, for each OTU, a nucleotidic alignment constituted by the associated sequence from each host. We considered a model where the sequences (i) evolve by substitution along the branches of the host phylogeny, (ii) are vertically transmitted during host speciation events, and (iii) experience a certain number of horizontal switches between host lineages. We computed the likelihood associated with the nucleotidic alignment under the model of vertical transmission with a given number of host switches, estimated the number of host-switches, and evaluated the model support in comparison with scenarios of environmental acquisition or strict vertical transmission. In Article 1, we tested the performances of the approach using simulations. We first applied HOME to the great apes microbiota (Ochman *et al.*, 2010) and found that some bacterial taxa, representing >5%, have been transmitted during great apes evolution. While a few cases of bacterial transmissions in great apes have already been demonstrated recently by amplifying lineage-specific bacterial genes (Moeller *et al.*, 2016), our approach allows scanning the whole microbiota without formulating any *a priori* on the transmitted microbes.

Second, in Article 2, we applied HOME on the bacterial microbiota of *Ariamnes* spiders that recently diversified along the Hawaiian archipelago and demonstrated that despite a significant pattern of phylosymbiosis, there is a few shreds of evidence of bacterial

transmissions in these spider microbiota. To evaluate the robustness of our findings, we further tested the performance of HOME when the number of segregated sites is very low (*i.e.* when host lineages diverged very recently) and when preferential host switches occurred (between closely related host lineages or between host lineages sharing the same geographic areas).

Finally, in Article 3, we compared HOME to other available approaches. Using simulations, we investigated the performances of event-based approaches (ALE and HOME) and global-fit approaches (ParaFit and PACo). We found that HOME has low statistical power compared to the other approaches, but it has the advantage of presenting a very type-I error rate (*i.e.* low number of false-positives). We applied these approaches to the bacterial gut microbiota of 18 worldwide primate species and found that a significant signal of vertical transmission in up to 10% of the gut bacteria, irrespectively of the geographic distribution of the species.

Contents of Chapter I

Chapitre I : Caractériser la transmission microbienne sur une phylogénie d'hôtes

Déterminer si des symbiontes microbiens sont transmis durant l'évolution de leurs hôtes est une question centrale sachant les conséquences de la transmission pour les hôtes et les microbes. En effet, la transmission verticale garantit une stabilité de l'association, permet l'évolution de dépendances et limite la tentation de tricher. Dans ce chapitre, nous nous sommes intéressés à la détection de microbes transmis verticalement au sein des microbiotes d'un clade d'hôtes (Figure 0.3.13a). Nous avons développé un modèle probabiliste (HOME) afin d'inférer, par maximum de vraisemblance, l'histoire évolutive de chaque OTU microbien constituant le microbiote. Brièvement, sachant une phylogénie d'hôtes et les microbiotes des espèces hôtes au présent, notre approche utilise la variabilité nucléotidique au sein des OTUs pour détecter ceux qui sont transmis verticalement. Le modèle prend en entrées la phylogénie des hôtes et pour chaque OTU considéré indépendamment, un alignement nucléotidique constitué par les séquences d'ADN associées à chaque espèce d'hôtes. Nous avons considéré un modèle où les séquences d'ADN microbiens (i) évoluent par substitution le long des branches de la phylogénie des hôtes, (ii) sont transmises verticalement au moment des évènements de spéciations des hôtes et (iii) peuvent subir un certain nombre de transferts horizontaux entre lignées d'hôtes. Nous avons calculé la vraisemblance du modèle, correspondant à la probabilité d'observer l'alignement nucléotidique sous le modèle de transmission verticale avec un certain nombre de transferts. Nous pouvons ainsi estimer le nombre de transferts et évaluer le support du modèle en comparaison avec un scénario d'acquisitions environnementales ou de stricte transmission verticale. Dans l'Article 1, nous avons testé les performances de notre approche à l'aide de simulations. Nous avons appliqué HOME sur le microbiote intestinal des grands singes (Ochman *et al.*, 2010) et trouvé que certaines espèces bactériennes, représentant plus de 8% de leur microbiote, ont été transmises durant l'évolution des grands singes. Alors que certaines transmissions bactériennes ont déjà été récemment démontrées, grâce à l'amplification de gènes bactériens spécifiques à certains groupes (Moeller *et al.*, 2016), notre approche permet de scanner le microbiote entier sans besoin de formuler des hypothèses *a priori* sur les microbes transmis.

Deuxièmement, dans l'Article 2, nous avons appliqué HOME sur les données de microbiotes bactériens d'araignées *Ariamnes* qui se sont récemment diversifiées le long de l'archipel Hawaïen. Malgré un patron significatif de phylosymbiose, nous n'avons pas (ou peu) trouvé de transmissions verticales dans ces microbiotes d'araignées. Afin d'évaluer la robustesse de nos résultats, nous avons exploré plus en détail les performances de HOME lorsque le nombre de sites qui ségrégent est très bas (*i.e.* lorsque les hôtes ont divergé très récemment) et lorsque des transferts horizontaux préférentiels ont lieu (entre des lignées d'hôtes phylogénétiquement proches ou entre des lignées d'hôtes qui partagent la même aire géographique).

Enfin, dans l'Article 3, nous avons comparé HOME aux autres approches disponibles. Grâce à des simulations, nous avons étudié les performances d'approches réconciliant des évènements (ALE et HOME) et d'approches mesurant uniquement un signal global (ParaFit et PACo). Nous avons trouvé que HOME a un plus faible pouvoir statistique comparé aux autres approches, mais qu'il a l'avantage de présenter un taux d'erreur de type-I très faible (*i.e.* un faible nombre de faux positifs). Nous avons appliqué ces approches sur les microbiotes intestinaux bactériens de 18 espèces de primates et trouvé un signal significatif de transmission verticale concernant jusqu'à 10% des bactéries, indépendamment de la distribution géographique des espèces de primates.

Article 1: Characterizing symbiont inheritance during host-microbiota evolution: application to the great apes gut microbiota:

Authors: Benoît Perez-Lamarque[1,2] & Hélène Morlon[1]

[1] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

Abstract

Microbiota play a central role in the functioning of multicellular life, yet understanding their inheritance during host evolutionary history remains an important challenge. Symbiotic microorganisms are either acquired from the environment during the life of the host (*i.e.* environmental acquisition), transmitted across generations with a faithful association with their hosts (*i.e.* strict vertical transmission), or transmitted with occasional host-switches (*i.e.* vertical transmission with horizontal switches). These different modes of inheritance affect microbes' diversification, which at the two extremes can be independent from that of their associated host or follow host diversification. The few existing quantitative tools for investigating the inheritance of symbiotic organisms rely on cophylogenetic approaches, which require knowledge of both host and symbiont phylogenies, and are therefore often not well adapted to DNA metabarcoding microbial data.

Here, we develop a model-based framework for identifying vertically transmitted microbial taxa. We consider a model for the evolution of microbial sequences on a fixed host phylogeny that includes vertical transmission and horizontal host-switches. This model allows estimating the number of host-switches and testing for strict vertical transmission and independent evolution. We test our approach using simulations. Finally, we illustrate our framework on gut microbiota high-throughput sequencing data of the family Hominidae and identify several microbial taxonomic units, including fibrolytic bacteria involved in carbohydrate digestion, that tend to be vertically transmitted.

Keywords: symbiont transmission, microbiota, molecular evolution, likelihood-based framework, holobiont, great apes.

Author contributions: BPL and HM designed research, BPL performed research, BPL and HM analyzed data and wrote the paper.

Acknowledgments: The authors thank Ana Alfonso Silva, Leandro Aristide, Julien Clavel, Carmelo Fruciano, Eric Lewitus, Sophia Lambert, Odile Maliet, Marc Manceau, Olivier Missa, and Guilhem Sommeria-Klein for helpful comments on the article. They also thank Florian Hartig, Marc-André Selosse and Florent Martos for helpful discussions, as well as the Associate Editor and the anonymous reviewers for their constructive comments on the previous version of the manuscript. This work was supported by the Centre national de la recherche scientifique (CNRS), the grant PANDA from the European Research Council (ERC-CoG) attributed to HM, the Ecole Doctorale FIRE - Programme Bettencourt, and a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL.

Data availability: The implementation of HOME is available on GitHub (`https://github.com/BPerezLamarque/HOME`) where we also provide a tutorial and scripts to prepare the data. The sequences used in our empirical applications are available in `https://doi.org/10.5061/dryad.023s6/3`.

Citation: Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host-microbiota evolution: Application to the great apes gut microbiota. Molecular Ecology Resources 19: 1659–1671.

Introduction:

Microbiota – host-associated microbial communities – play a major role in the functioning of multicellular organisms (Hacquard *et al.*, 2015). For example, the gut microbiota plays a significant nutritional role for animals by synthesizing essential nutrients and by helping digestion and detoxification (McFall-Ngai *et al.*, 2013). It is also involved in a broad range of other mutualistic functions important for host protection, development, behavior, and reproduction (Zilber-Rosenberg & Rosenberg, 2008). Other less-studied microbiota, such as those found on animal skins or plant roots also play major ecological roles (Philippot *et al.*, 2013).

Host-microbiota associations have evolved for thousand million years with three major modes of inheritance across phylogenetic host lineages: (i) strict vertical transmission within a host lineage (Rosenberg & Zilber-Rosenberg, 2016), which can happen either by transmission from mother to child (*e.g.* directly through ovaries during reproduction or at birth), or by social contact while sharing life with related individuals (Bright & Bulgheresi, 2010), (ii) vertical transmission with occasional horizontal switches between host lineages (Henry *et al.*, 2013), which can for example happen through direct interactions, via vectors or via shared habitats (Engel & Moran, 2013), and (iii) environmental acquisition, with microbes coming from the environment independently from other related hosts (Bright & Bulgheresi, 2010). The vertical transmission of a given microbial lineage within host lineages can lead to cophylogenetic patterns, with the microbial

phylogeny mirroring the host phylogeny (*e.g. Helicobacter pylori* in humans; Linz *et al.* (2007)). Horizontal switches and environmental acquisitions can play key roles in adaptation, for example by allowing host lineages to adapt to new feeding regimes (Muegge *et al.*, 2011; McKenney *et al.*, 2018). They will tend to erase cophylogenetic patterns linked to vertical transmission. The relative importance of each of the three modes of inheritance depends on the type of host and the type of microbes. For example, vertical transmission is thought to be far more preponderant in the 'core' microbial species, which are shared across hosts regardless of environmental conditions, than in the 'flexible' microbial species, facultative and dependent on internal and external conditions (Shapira, 2016).

Quantifying the relative importance of different modes of inheritance during host-microbiota coevolution remains a major challenge. Patterns of 'phylosymbiosis', *i.e.* a pattern of concordance between a given host phylogeny and the dendrogram reflecting the similarity of microbial communities across these hosts, is frequently observed (Bordenstein & Theis, 2015), for example for great apes gut microbiota (Ochman *et al.*, 2010). Although these phylosymbiotic patterns may suggest that some microbial species within the microbiota are vertically transmitted, such community-wide comparisons of microbiota across hosts do not allow identifying which microbial species are vertically transmitted, nor quantifying the relative importance of the different modes of inheritance across distinct microbial species. More recently, approaches have been developed to apply cophylogenetic concepts to microbial taxa (Groussin *et al.*, 2017; Bailly-Bechet *et al.*, 2017). Cophylogenetic methods were originally developed to study the coevolution between hosts and their symbionts, with the underlying idea that close and long-term associations lead to congruent phylogenies with similar topologies and divergence times (Page & Charleston, 1998; de Vienne *et al.*, 2013), while processes such as host-switches disrupt this congruence. Cophylogenetic tools either quantify the congruence between symbiont and host trees using distance-based methods – *e.g.* ParaFit (Legendre *et al.*, 2002), generalizations of the Mantel test (Hommola *et al.*, 2009), or PACo (Balbuena *et al.*, 2013) – or try to find the most parsimonious sets of events (*e.g.* host-switches) that allow reconciling both trees (*e.g.* TreeMap or Jane; Conow *et al.* (2010)). In the context of microbiota, Groussin *et al.* (2017) and Bailly-Bechet *et al.* (2017) have used the ALE program (Szöllősi *et al.*, 2013b,a), which was initially designed to solve the gene tree - species tree reconciliation problem. Importantly, these event-based methods require first a reconstruction of the microbial tree for each individual microbial taxa. However, microbiota data are typically generated using Next Generation Sequencing (NGS) metabarcoding techniques, providing short DNA reads of a targeted slow-evolving universal gene (*e.g.* the 16S rRNA gene). Such data often contain limited nucleotide variability within each microbial taxa, which can be problematic for reconstructing their tree.

Here, we develop a probabilistic model of host-symbiont evolution, which aims at studying modes of inheritance in the microbiota without building first microbial phy-

logenies. The main idea is to use the host phylogenetic tree to inform the microbial trees, which reduces the problem of low phylogenetic resolution of metabarcoding microbial markers. Huelsenbeck *et al.* (2000) developed a model of cospeciation and host-switches similar to ours, focused on host-parasite associations. However, the authors developed an inference framework reconstructing host and parasite phylogenetic trees jointly, which is not well adapted to the case when the host phylogenetic tree is robust and the symbionts are represented by a sequence alignment with limited phylogenetic information. Here, we fix the host phylogeny and follow the evolution of individual microbial taxa on the host tree. We compute likelihoods associated with microbial sequence alignments under a model including vertical inheritance and host-switches. We find estimates of the number of host-switches and develop tests for evaluating model support in comparison with scenarios of independent evolution and strict vertical transmission. We test our approach using simulations and apply it to gut microbiota high-throughput sequencing data of the family Hominidae.

Methods:

HOME: A general framework for studying Host-Microbiota Evolution:

From metabarcoding microbiota data to separate alignments:

Given a host species tree and metabarcoding microbiota data sampled from each host species (*e.g.* sequences from the 16S rRNA gene, ITS, or any other DNA metabarcoding marker), our framework begins by clustering sequences into Operational Taxonomic Units (OTUs) using bioinformatics pipelines. Each OTU is made of distinct microbial populations, each corresponding to a specific host species (Figure I.1.1a). We assume as a starting point that there is no within-host genetic variability (we discuss later how we relaxed this assumption), such that each microbial population is represented by a unique sequence. In our analysis of these data, for each OTU and each host, we use the most abundant microbial sequence as the representative sequence. The data we consider thus consists in a series of microbial alignments A, each corresponding to a sequence alignment for a specific OTU; a given alignment is composed of N-nucleotidic sites long sequences (with potential gaps considered as missing data), each corresponding to a specific host. In each alignment, we distinguish the segregating sites (*i.e.* those that vary in at least one sequence) to those that do not vary across sequences. Some microbial OTUs may not be represented in all host species (*i.e.* there might be missing sequences in the alignment), which can either be true absences (*i.e.* the corresponding host species do not host the OTU), or a lack of detection (*i.e.* the OTU is present but has not been sampled in these host species). Because we cannot distinguish these two possibilities, we simply treat missing sequences as missing data; we do not explicitly model the extinction of symbiotic populations in certain host species, nor the microbial sampling process. We

apply our model separately to each alignment.

Modeling the evolution of an OTU on a host phylogeny:

We consider the evolution of a given microbial OTU on a host phylogeny T (Figure I.1.1); T is assumed to be a known, ultrametric, rooted and binary n-tips tree. The model is defined as follows:

(i) Vertical transmission: From an ancestral microbial population at the root of the host phylogeny represented by a N-nucleotidic sites long sequence with N_v 'variable' sites (*i.e.* those that can experience substitutions), substitutions occur along host branches. Following classical models of molecular evolution (Strimmer & von Haeseler, 2009), we assume that each variable site evolves independently from the others according to a substitution model with a rate μ that is supposed to be the same for all variable sites and constant along the evolutionary branches (strict-clock model). The substitution model is represented by a continuous-time reversible Markov process, characterized by an invariant measure π (*i.e.* the vector of base frequencies at equilibrium) and an instantaneous transition rate matrix Q between different states (Strimmer & von Haeseler, 2009). At a host speciation event, the two daughter host lineages inherit the microbial sequence from the ancestral host, after which microbial populations on distinct host lineages evolve independently.

(ii) Host-switches: A discrete number (ξ) of host-switches happens during the evolution of the OTU on the host tree. The switches occur from a 'donor' branch, with a probability proportional to its branch length, and at a time uniformly distributed on the branch, to a 'receiving' branch, with equiprobability among the co-existing branches (we do not consider the phylogenetic proximity from the donor branch). When a host-switch happens, for convenience we assume that the microbial sequence from the donor host replaces that of the receiving host and the microbial sequence from the donor host remains unchanged.

Each series of host-switches on T defines a tree of microbial populations T_B that summarizes which populations descended from which ones and when their divergences occurred (Figure I.1.1). In the absence of host-switches ($\xi = 0$), T_B and T are identical. When host-switches occur, they break the congruence between T_B and T (*e.g.* Figure I.1.1c). Hence, the model can be decomposed in two steps: first, host-switches generate T_B from T; second, a sequence (representing a microbial population) evolves on T_B with a constant substitution rate.

Likelihood computation and inference:

We develop a likelihood-based framework in order to fit the above model to data comprising a given (fixed) tree T of hosts and an alignment A_s of microbial sequences characterizing populations of a given microbial OTU for these hosts (here the alignment

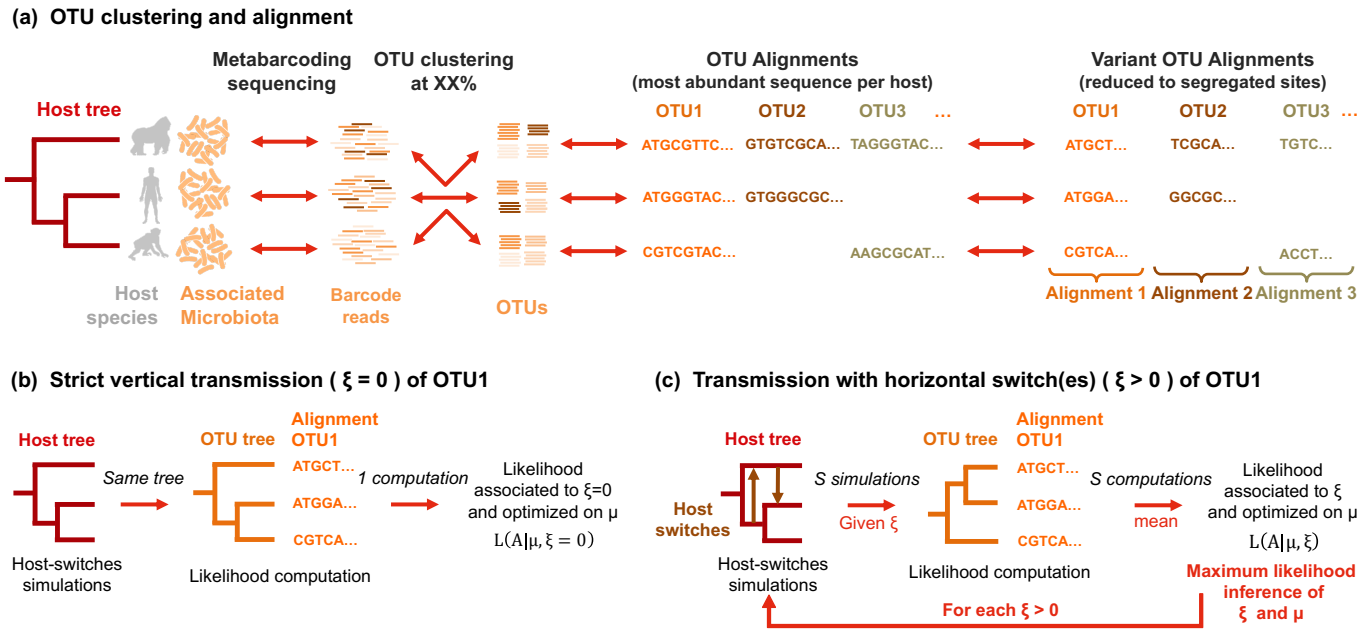


Figure I.1.1: Illustration of the various steps for assessing microbial modes of inheritance in host-microbiota evolution from metabarcoding data. (a) The first step consists in clustering the microbial sequences into OTUs and building for each OTU the corresponding alignment of segregating sites (A_s). (b, c) The second step consists in fitting different models of inheritance to each microbial alignment. We compute the probability of the microbial alignment on hypothetical microbial trees. Under a model with strict vertical transmission (ξ=0, b), the microbial is the same as the host tree; under a model with vertical transmission and host-switches ($\xi > 0$, c), microbial trees are simulated from the host tree with various numbers of switches ξ. We find the substitution rate $\hat{\mu}$ and the number of switches $\hat{\xi}$ that maximize the probability of the alignment.

A_s is reduced to the segregating sites). This will allow estimating the number of switches $\hat{\xi}$ on the host tree. The probability of the alignment assuming that the substitution rate is μ and that there are ξ switches is given by:

$$L(A_s|\mu,\xi) = \int_{T_B} L(A_s|\mu,T_B)dT_B$$

where $L(A_s|\mu,T_B)$ is the probability of the alignment assuming that the substitution rate is μ and the (dated) microbial tree is T_B, and the integral is taken over the space of dated trees obtained with ξ switches on T. In practice, we compute this integral using Monte Carlo simulations: we simulate a large number (S) of dated microbial trees obtained with ξ switches on T (see next section), compute for each T_B the probability of the alignment assuming that the substitution rate is μ, and sum these probabilities:

$$L(A_s|\mu,\xi) = \frac{1}{S}\sum_{T_B} L(A_s|\mu,T_B)dT_B$$

This approximate expression converges to the exact integral form when S is large.

We compute the probability $L(A_s|\mu,T_B)$ of the sequence alignment A_s on a given

dated microbial tree T_B using the Felsenstein pruning algorithm (Felsenstein, 1981). We take into account the possibility of gaps in the microbial alignment, considering them as 'missing values' by pruning off the tips of the tree with a gap (Truszkowski & Goldman, 2016). First, we choose the model of DNA substitution between the K80, F81, and HKY matrices from the alignment reduced to segregating site (A_s) using the function *modelTest* (R-package phangorn) and based on a BIC selection criterion: this function estimates Q and π directly from A_s, where Q, the reversible transition rate matrix, depends on the invariant measure π. We also obtain estimates of the transition/transversion rate ratio κ (K80 and HKY) and of the base frequencies at equilibrium π (F81 and HKY) from these models. Second, we compute the probability of the alignment at each nucleotidic site ν using the pruning algorithm. For a given segregating site among A_s, let $P(t)$ be the vector of probabilities of states A, C, G, and T at time t. $P(t)$ is given by $P(t) = M(t) * P(0)$ where $P(0) = (\mathbb{1}_A, \mathbb{1}_C, \mathbb{1}_G, \mathbb{1}_T)$ with $\mathbb{1}_A$ equals 1 if A is the initial nucleotide is and 0 otherwise, and $M(t) = e^{t\mu Q}$. Let $P_\nu(s)$ be the probability of the alignment corresponding to the clade descending from node s in the phylogeny for site ν. We have:

$$P_\nu(\text{leaf}) = (\mathbb{1}_A, \mathbb{1}_C, \mathbb{1}_G, \mathbb{1}_T) \text{ and } P_\nu(s) = (M(t_1)P_v(s_1)).(M(t_2)P_\nu(s_2))$$

Where s_1 and s_2 are the two nodes descending from s and t_1 and t_2 are their respective times of divergence (t_1 and t_2 are fixed, given by the branch lengths of the simulated dated tree T_B). We iterate this pruning calculation from the leaves to the root of the tree, and obtain the probability of the alignment at site ν:

$$L_\nu = \pi P_\nu(\text{root})$$

Because we consider only segregating sites, we condition this probability on the occurrence of at least one substitution. The probability of a substitution happening on a tree T_B of total branch length B is given by $(1 - e^{-\mu B})$. Finally, the probability of the alignment A_s is obtained by multiplying the probabilities corresponding to each site. Hence the probability of the variable alignment A_s is given by:

$$L(A_s|\mu, T_B) = (1 - e^{-\mu B})^{-N_s} \prod_{\nu=1}^{N_s} L_\nu$$

where N_s is the number of segregating sites.

In practice, we used $S = 10^4$ and plotted the resulting value of $L(A_s|\mu, \xi)$ with an increasing number of trees T_B to ensure that S was large enough to obtain a reliable approximation of the likelihood. For each ξ, we find μ that maximizes $L(A_s|\mu, \xi)$. Finally, we repeat these analyses for a range of realistic ξ values (typically $\xi = [0, 1, 2, ..., 2n]$) and deduce the couple of parameters $\hat{\xi}$ and $\hat{\mu}$ that maximizes the probability of the alignment. Likelihood landscapes typically have a well-defined peak (Supplementary Figure 1), suggesting that ξ and μ are identifiable. We also show later that we can properly es-

timate them under a wide set of scenarios. Low $\hat{\xi}$ values are indicative of OTUs that are transmitted mostly vertically, while high $\hat{\xi}$ values are indicative of those that perform frequent host-switches.

Simulations of host-switches: from T to T_B:

By simulating ξ switches on T, we obtain a (dated) bacterial T_B characterized by its topology and its branch lengths. Each switch is characterized by its 'donor' branch, by its position on the branch, and by its 'receiving' branch. The donor branch is chosen with a probability proportional to its branch length, the time of the switch is drawn uniformly on the branch, and the receiving branch is chosen with equiprobability among the lineages alive at time t. A switch replaces the existing microbial sequence in the receiving host, and creates a new branching event in the microbial tree T_B. Four types of switches can occur and each of them results in different rules to obtain T_B from T (Figure I.1.2):

(i) the switch occurs just after the root on the host tree, before any other speciation event: T_B is obtained from T by re-dating the root of the tree to the time of the host-switch. This switch does not change the topology of the tree (*i.e.* it only affects the branch lengths).

(ii) the switch occurs from an internal branch to a branch directly related to the root, *i.e.* one of the sequences originating at the root no longer has descendants in the current sequences: T_B is obtained from T by re-rooting the tree to the most recent common ancestor to all the current microbial sequences. This switch changes both the topology of the tree and the branch lengths.

(iii) the switch occurs between 2 sister lineages: T_B is obtained from T by re-dating the divergence between the two sister lineages to the time of the host-switch. This switch only affects the branch lengths of the tree.

(iv) the switch occurs between 2 distantly related lineages and the receiving branch is not related to the root: T_B is obtained from T by an internal reorganization of the tree. This switch changes both the topology of the tree and the branch lengths.

Technically, in order to reduce computation time, we simulated a 'bank of trees' with ξ switches on the host tree and use these same trees in our different analyses.

Model selection:

In addition to the general model fitting procedure described above, we designed two model selection procedures: the first aims at testing whether the presence of horizontal switches is statistically supported (versus a simpler model with only strict vertical transmission); the second aims at testing support for a model with a limited number of host-switches *versus* environmental acquisition (OTUs that are environmentally acquired will provide high $\hat{\mu}$ and $\hat{\xi}$ estimates and could thus be interpreted as vertical transmission

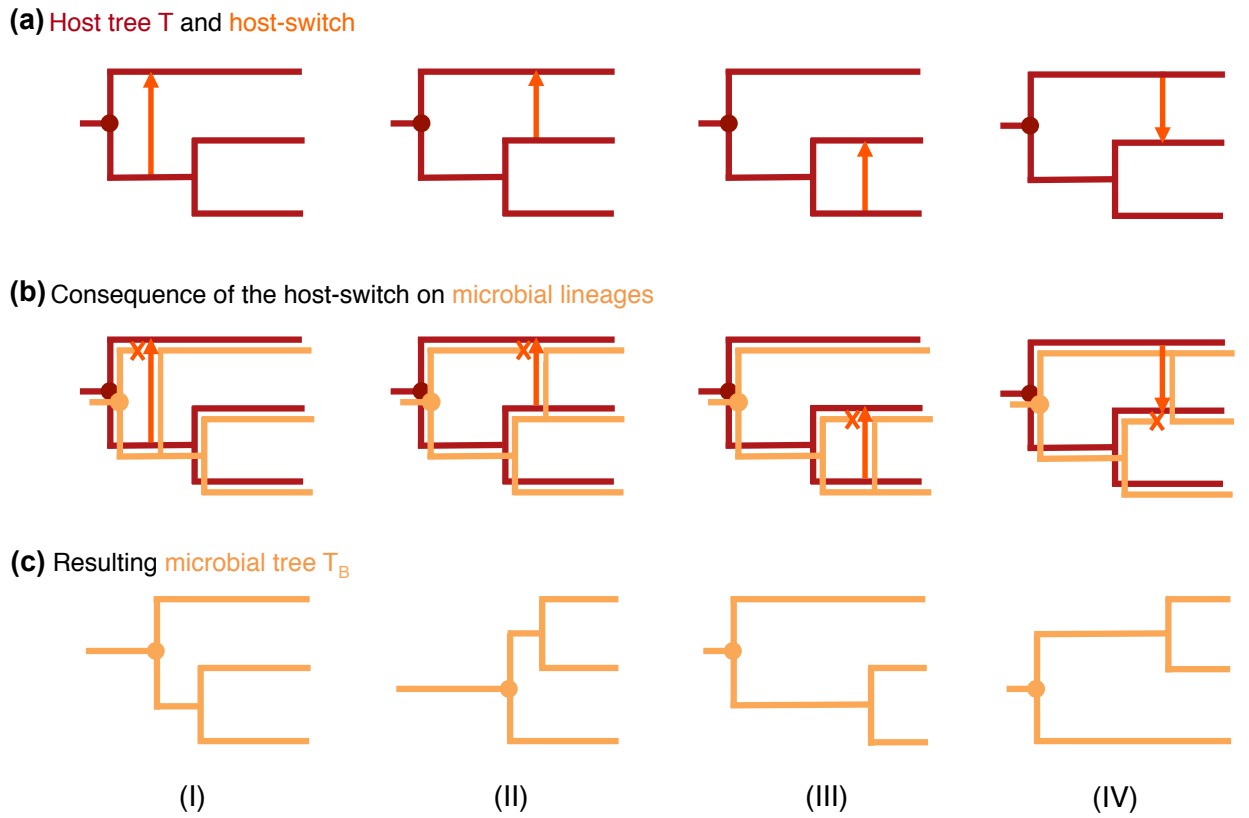


Figure I.1.2: Host-switch simulations. (a) Four types of host-switch can occur on the host tree T (b-c) these host switches generate distinct microbial trees T_B. Orange arrows represent host-switches. Orange crosses represent the extinction of the microbial lineage on the receiving branch.

with frequent horizontal switches and high substitution rates instead of environmental acquisition).

In order to test support for a scenario with horizontal host-switches *versus* strict vertical transmission, we compute $L_0 = L(A_s|\hat{\mu}, T)$, the likelihood corresponding to the best scenario of evolution of the microbial sequences directly on the host tree (*i.e.* no switch) and compare it to the likelihood $L_1 = L(A_s|\hat{\mu}, \hat{\xi})$ corresponding to the best scenario with horizontal switches, using a likelihood ratio test.

In order to test support for a scenario of vertical transmission with horizontal host-switches *versus* environmental acquisition, we test its support when compared to a scenario where microbial populations are acquired at random by host species (thereafter referred to as a scenario of 'independent evolution'): we randomize R times the host-microbe association and run our model on each of these randomized data. Next, we analyze the rank of $\hat{\xi}$ and $\hat{\mu}$ estimated from the original alignment in the distribution of ξ_R and μ_R estimated from the randomized alignments. Ideally, we would perform a large number of randomizations (*e.g.* $R > 100$) and directly compute p-values from the ranks of $\hat{\xi}$ and $\hat{\mu}$. However, for computational reasons we used only $R = 10$ randomized alignments and chose to reject the hypothesis of independent evolution if $\hat{\xi} < \xi_R$ and $\hat{\mu} < \mu_R$ for all R. Conversely, if the estimated number of switches ξ or the substitution rate μ are ranked within the distribution of ξ_R and μ_R, we consider that a scenario of independent

evolution cannot be rejected. There are thus two (indistinguishable) scenarios that will produce microbial alignments that won't reject our test of independent evolution: environmental acquisition and vertical transmission with highly frequent host-switches.

Detecting transmitted OTUs:

Based on the analyses above and our definition of modes of inheritance, we sort the OTUs into two different categories: the transmitted OTUs (those that reject the hypothesis of independent evolution, either because they are strictly vertically transmitted, or because they are vertically transmitted with a few host-switches), and the independent OTUs (those that do not reject the hypothesis of independent evolution, either because they are environmentally acquired, or because they experienced enough host switches to be indistinguishable from a scenario of environmental acquisition). In practice, there is no universal similarity threshold that will provide the 'right' biological unit delineation across all microbial groups (Sanders *et al.*, 2014; Supplementary Figure 2). 'Over-splitting' a biological unit using a similarity threshold that is too high for that biological unit will reduce statistical signal (each sub-unit will be represented in fewer hosts) and will miss host-switches between sub-units (given that sub-units will be analyzed independently). 'Over-merging' OTUs using a similarity threshold that is too low will tend to blur a signal of transmission, and will over-estimate substitution rates, because alignments will mix sequences from distinct biological units. By using several clustering thresholds, we can hope to find one that properly delimitates biological units. Given that vertical transmission tends to be erased by improper delimitation, if it is detected for at least one threshold, then it suggests that it is the 'right' threshold and that vertical transmission does indeed occur.

Implementation:

All the scripts of our model are written in R (R Core Team 2018), using the packages ape, phangorn, and phytools for the manipulations of phylogenetic trees (Paradis *et al.*, 2004; Schliep, 2011; Revell, 2012) and are available on GitHub (`https://github.com/BPerezLamarque/HOME/`). Some internal functions computing the likelihood are coded in C++. We also used the packages parallel, expm, ggplot2, reshape2, Rcpp and R2HTML for the technical aspects of the scripts. All outputs of our model (*e.g.* parameter estimation and model selection) are concatenated in a user-friendly HTML file with different formats (*e.g.* tables, values, pdf plots and diagrams). The computational time depends both on the number of host (n) and on the number of trees (S) used in the likelihood inference; examples of computation time are provided in Supplementary Figure 3.

Testing our approach with simulations:

We performed a series of simulations to test the ability of our approach to recover simulated parameter values and evolutionary scenarios. We calibrated our choices of

tree size, alignment size and parameter values so as to obtain simulated data comparable to those of the great ape-microbiota data (Supplementary Figure 9 and Supplementary Table 2). We considered 3 independent host trees of size $n = 20$ (T1, T2, and T3) simulated under a Yule model (no extinction) using the function *pbtree* from phytools. We scaled these trees to a total branch length of 1. On each of these host trees, we considered a scenario of strict vertical transmission (ξ=0), scenarios of vertical transmission with host-switches $\xi \in [1, 2, 3, 5, 7, 10]$, and a scenario of environmental acquisition; each of these scenarios were obtained by simulating the corresponding microbial trees T_B. For the scenario of strict vertical transmission, $T_B = T$. For scenarios of host-switches, 15 T_B per ξ value were derived from T. For the scenario of environmental acquisition, 20 T_B with n tips were simulated under a Yule model independently from T, using the same procedure as above. Finally, we simulated on each T_B the evolution of microbial sequences of a total length $N = 300$ using our own codes, with a probability 0.1 for each site to be variable. We simulated the K80 stochastic nucleotide substitution process with a ratio of transition/transversion rate $\kappa = 0.66$ and three different values of substitution rate ($\mu = 0.5$, 1, or 1.5). The realized proportion of segregating sites was quite variable and comparable to empirical alignments (Supplementary Figure 9). We simulated 20 alignments A per substitution rate on T for the scenario of strict vertical transmission (180 alignments in total), and 1 alignment per T_B per substitution rate for the scenarios of host-switch (135 alignments per ξ value) and environmental acquisition (180 alignments). Thereafter we call 'ξ-switches alignment' an alignment simulated with ξ switches on T and 'independent alignment' an alignment simulated under the environmental acquisition scenario (*i.e.* independently from T).

We applied our inference approach to each simulated couple of T and A and compared the estimated parameters ($\hat{\xi}$, $\hat{\mu}$, and $\hat{\kappa}$) to the simulated values. We used mixed linear models with the host tree (T1, T2, and T3) as a random effect (R-package nlme). We tested homoscedasticity and normality of the model residuals and considered a p-value of 0.05 as significant. We also evaluated the type-I and type-II errors associated with our tests of strict vertical transmission and independent evolution.

Empirical application: great apes microbiota:

We illustrate our approach using data from Ochman *et al.* (2010); this paper is one of the first paper looking at phylosymbiotic patterns in great apes, and the associated data has been used in other papers aiming at studying transmission (Sanders *et al.*, 2014). The dataset consists of fecal samples collected from 26 wild-living hominids, including eastern and western African gorillas (2 individuals of *G. gorilla* and 2 individuals of *G. beringei*), bonobos (6 individuals of *P. paniscus*), and three subspecies of chimpanzees (5 individuals of *P. t. schweinfurthii*, 7 individuals of *P. t. troglodytes* and 2 individuals of *P. t. ellioti*), as well as two humans from Africa and America (*H. sapiens*).

Ochman *et al.* (2010) extracted DNA from the fecal samples, PCR-amplified the DNA

for the 16S rRNA V6 gene region using universal primers, and finally sequenced the PCR products using 454 (Life Sciences/Roche). They obtained 1,292,542 reads after sequence quality trimming and barcodes removal. Gut microbiota are now sequenced with more coverage than what was possible at the time of the Ochman paper, yet these data represent a good application of our approach. We obtained the reads from Dryad (`http://datadryad.org/resource/doi:10.5061/dryad.023s6`). We used python scripts from the Brazilian Microbiome Project (BMP, available on `http://www.brmicrobiome.org/`; Pylro *et al.*, 2014) which combines scripts from QIIME 1.8.0 (Caporaso *et al.*, 2010) and USEARCH 7 (Edgar, 2013) as well as our own bash codes. We merged raw reads from all the hosts and processed them step by step:

(i) Dereplication: we discarded all the singletons and sorted the sequences by abundance using USEARCH commands *derep_fulllength* and *sortbysize*.

(ii) Chimera filtering and OTU clustering: we removed chimeras and clustered sequences into OTUs using the *cluster_otus* command of the UPARSE pipeline (Edgar, 2013). We chose a 1.0, 3.0, or 5.0 OTU radius (the maximum difference between pairs of OTU member sequences), which corresponds to a minimum identity of 99, 97, and 95%. We performed an additional chimera filtering step using *uchime_ref* with the RDP database as a reference (`http://drive5.com/uchime/rdp_gold.fa`). We obtained 1,074 OTUs at 95%, 1,793 at 97%, and 4,935 at 99% (Supplementary Table 1).

(iii) Taxonomic assignation: we assigned taxonomy using a representative sequence for each OTU generated (with *cluster_otus*), using *assign_taxonomy.py* from QIIME and the latest version of the Greengenes database (`http://greengenes.secondgenome.com`), or using BLAST when Greengenes did not assign taxonomy with enough resolution.

(iv) Mapping reads to OTUs and OTU table construction: we used the *usearch_global* command to map all the reads from the different samples to these taxonomy-assigned OTUs. Then we used *make_otu_table.py* and BMP scripts to build the OTU table (a list of all the OTUs with their abundance by host individual).

(v) Core-OTUs selection: we selected the 'core' OTUs as the ones that occurred in at least 75% of the host individuals, using the *compute_core_microbiome.py* script from QIIME. This resulted in 134 core OTUs at 95%, 120 at 97%, and 71 at 99% (there are more OTUs at 99% than at 97% and 95%, but a much smaller proportion that are core OTUs, Supplementary Table 1).

(vi) Making intra-OTU alignments: discarding the few OTUs that had unvaried alignments, we obtained 130 core OTUs at 95%, 110 core OTUs at 97%, and 66 core OTUs at 99% similarity thresholds (Supplementary Table 1). Microbial genetic variability within each OTU and within each host individual (hereafter referred

to as 'intra-individual variability') was quite high, sometimes higher than inter-individual variability (Supplementary Figure 10a-c), suggesting that it was due to PCR and sequencing artefacts rather than true variability. Therefore, we built the bacterial alignment for a given OTU by selecting for each host individual the most abundant sequence among all the reads mapped to that OTU. This sequence is less likely to be subject to PCR/sequencing errors.

Finally, we applied HOME to each core OTU separately, and to the tree of the 26 host individuals, constructed with mitochondrial markers provided in the supplementary data of the article, scaled to a total branch length of 1. We used this individual-level tree instead of the species- or sub-species level tree in order to increase tree size (there are only 7 subspecies in our great apes tree); this approach also provides a way to account for microbial genetic variability within host subspecies (hereafter referred to as 'intraspecific variability'). We arbitrarily resolved intra subspecies polytomies by assigning quasi-null branch lengths (10^{-4}) to the corresponding branches. We classified the OTUs into 'transmitted' and 'independent' OTUs; among the transmitted OTUs, we distinguished those where the transmission is strictly vertical, and for the others we recorded the estimated number of host switches. In order to get an idea of the proportion of the microbiota that is transmitted we also recorded the number of reads corresponding to the transmitted OTUs.

Accounting for intra-host genetic variability:

Our treatment of the great ape data illustrates an approach to account for intra-host microbial genetic variability: instead of running HOME on a species-level host tree (with a single representative microbial sequence per host species), it can be run on an individual-level host tree, with arbitrarily small intra-specific branch-lengths. Because this usage of HOME is slightly different from the case envisioned in our description of the approach, we tested its behavior. We simulated the evolution of microbial alignments on the great apes sub-species tree with a range of intraspecific variability similar to the range observed in the great apes alignments. For each OTU alignment, we defined intraspecific variability (V) as the mean nucleotidic diversity within host subspecies (computed using Nei's estimator; Ferretti *et al.* (2012)) divided by the total nucleotidic diversity computed on the entire alignment. We simulated a total of 180 alignments according to 3 scenarios: strict vertical transmission ($\xi = 0$), transmission with 5 host-switches ($\xi = 5$), and environmental acquisition. For every scenario, we simulated intraspecific variability by extending the stochastic process generating nucleotidic substitution on every sequence for a time range that allowed to obtain levels of intraspecific variability that corresponded to the empirical level of intraspecific variability (Supplementary Figure 10d-i). We ran HOME on each of these simulated alignments and evaluated its performance, in terms of parameter estimation and model selection, when there was no intraspecific variability (V=0), low and intermediate intraspecific variability (0<V<0.5), and high intraspecific

variability (V>0.5).

Results:

Performance of HOME:

Likelihood landscapes typically display a single peak, illustrating that ξ and μ are in general identifiable (Supplementary Figure 1). Rarefactions curves also indicate that using $S = 10^4$ trees to compute the likelihood provides a good approximation (Supplementary Figure 4). Testing the performance of HOME using intensive simulations, we find a reasonable ability to recover simulated parameter values (Figure I.1.3). Estimates of the number of switches $\hat{\xi}$ are highly correlated with simulated ξ values, although the approach tends to overestimate the true number of switches when there are very few (less than 2) and to underestimate this number when there are many (Figure I.1.3A). The linear regression confirms these results $\hat{\xi}$ = 2.15 ($F_{dl=606}$=1015, p-value<0.0001) + $\xi \times 0.58$ ($F_{dl=606}$=141, p-value<0.0001). The ability to recover the true number of switches does not depend on the simulated substitution rate ($F_{dl=606}$=0.26, p-value=0.61; Supplementary Figure 5). The substitution rate is rather well estimated (Figure I.1.3b), although it tends to be slightly overestimated when the simulated number of switches exceeds 3 (slope 0.04; $F_{dl=606}$=45.9, p-value<0.0001; Figure I.1.3b). The simulated transition/ transversion rate ratio κ is well estimated (median $\pm$ s.d. = 0.68 $\pm$ 0.17), although it is slightly underestimated when the substitution rate is high (slope of -0.015; $F_{dl=606}$=12, p-value=0.0007). For alignments simulated independently from the host tree, the approach estimates a high number of switches (median $\pm$ s.d. = 16 $\pm$ 6.2, Figure I.1.3a), and highly overestimates the substitution rate (Figure I.1.3b). The type of host tree (T1, T2 or T3) has little impact on the estimation of ξ (it explains less than 3% of the total variance, Supplementary Figure 5), μ (around 10%, Supplementary Figure 6) and κ (less than 0.01%).

Our model selection procedure has very low type-I error rates, and type-II error rates that depend on the situation (Figure I.1.4): the hypothesis of strict vertical transmission was nearly never rejected when transmission was indeed strictly vertical (1/180, type-I error rate = 0.0056%) and always rejected under environmental acquisition (Figure I.1.4a); conversely, the hypothesis of independent evolution was almost always rejected when transmission was strictly vertical (1/180) and almost never rejected under environmental acquisition (3/180, type-I error rate = 0.017%, Figure I.1.4b). While the type-I error rates of the two tests are low, their power to detect a scenario of strict vertical transmission with host-switches is variable. In the case of the test of strict vertical transmission, the power ranges from 95% for ξ=10 to 45% when ξ=1 (Figure I.1.4a). In the case of the test of independent evolution, the power ranges from 100% for ξ=1 to 60% for ξ=10, and it would decrease further with more switches (Figure I.1.4b). In both cases, the power increases when the substitution rate μ is larger (Supplementary Figure 7).

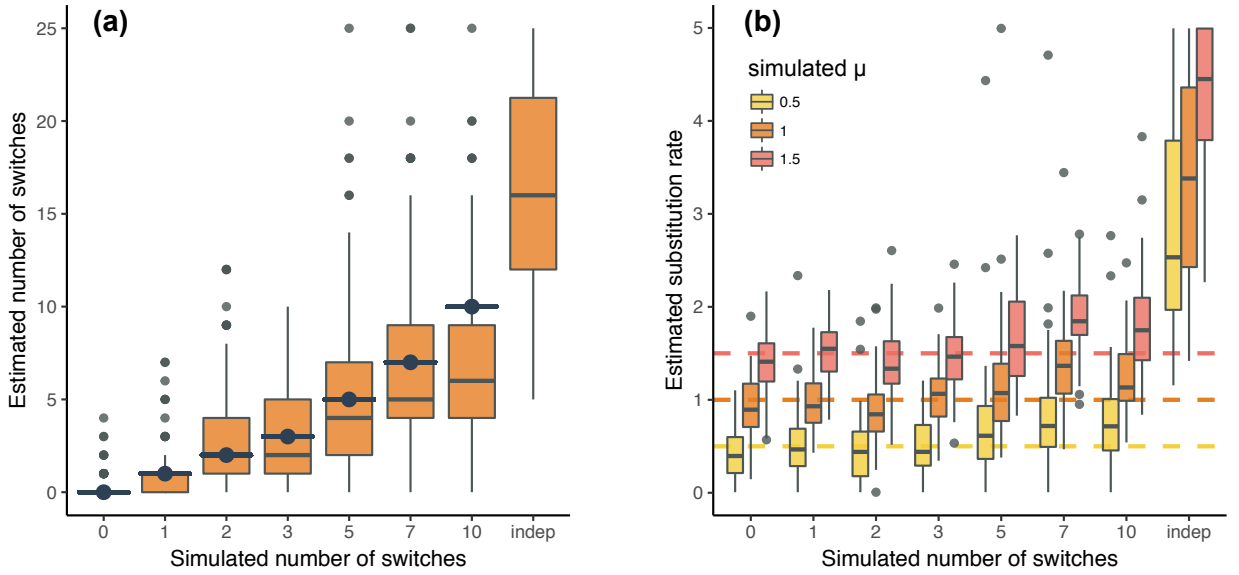


Figure I.1.3: Parameter estimation. Estimated *versus* simulated number of switches ξ (a) and substitution rate μ (b) under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution, referred in the figure as "indep"). Simulated values are represented by blue ticks in (a) and dashed lines in (b). Boxplots present the median surrounded by the first and third quartile, and whiskers extended to the extreme values but no further than 1.5 of the inter-quartile range.

When HOME is applied to an individual-level host tree in order to account for intraspecific microbial genetic variability, type-I error rates associated to the test of independent evolution remain very low regardless of the magnitude of the variability (Supplementary Figure 8). The confidence in the estimation of the parameters (ξ and μ) remains good for low values of intraspecific variability (V<0.5), but decreases with increasing variability (V>0.5). The type-I error rate associated to the test of strict vertical transmission increases with increasing variability, and the power of the two tests decreases with increasing variability.

Modes of inheritance in the great apes microbiota:

Applying HOME to great apes gut microbiota data, we found that among the core OTUs with at least one segregating site, approximately 1 in 10 OTUs is transmitted (*i.e.* rejects the test of independent evolution, Figure I.1.5a); more specifically, the ratios of transmitted OTUs (and strictly vertically transmitted OTUs) were the following: 12(8)/130 at 95%, 12(10)/110 at 97%, and 4(4)/66 at 99%. In terms of relative abundance, 108,206 raw sequences in a total of 1,292,542 (8.4%) belonged to transmitted OTUs (Supplementary Table 3). Almost half of the sequences from transmitted OTUs (49,508) were from an *Acinetobacter* bacterium (family Moraxellaceae; phylum Proteobacteria); another important pool of these sequences was from the family Prevotellaceae (28,843 reads; phylum Bacteroidetes). In total, 12 bacterial families (in 27) contained OTUs that were transmitted, including Veillonellaceae, Lachnospiraceae, Ruminococcaceae , and Paraprevotellaceae (Figure I.1.5b, Supplementary Table 4). Some of these

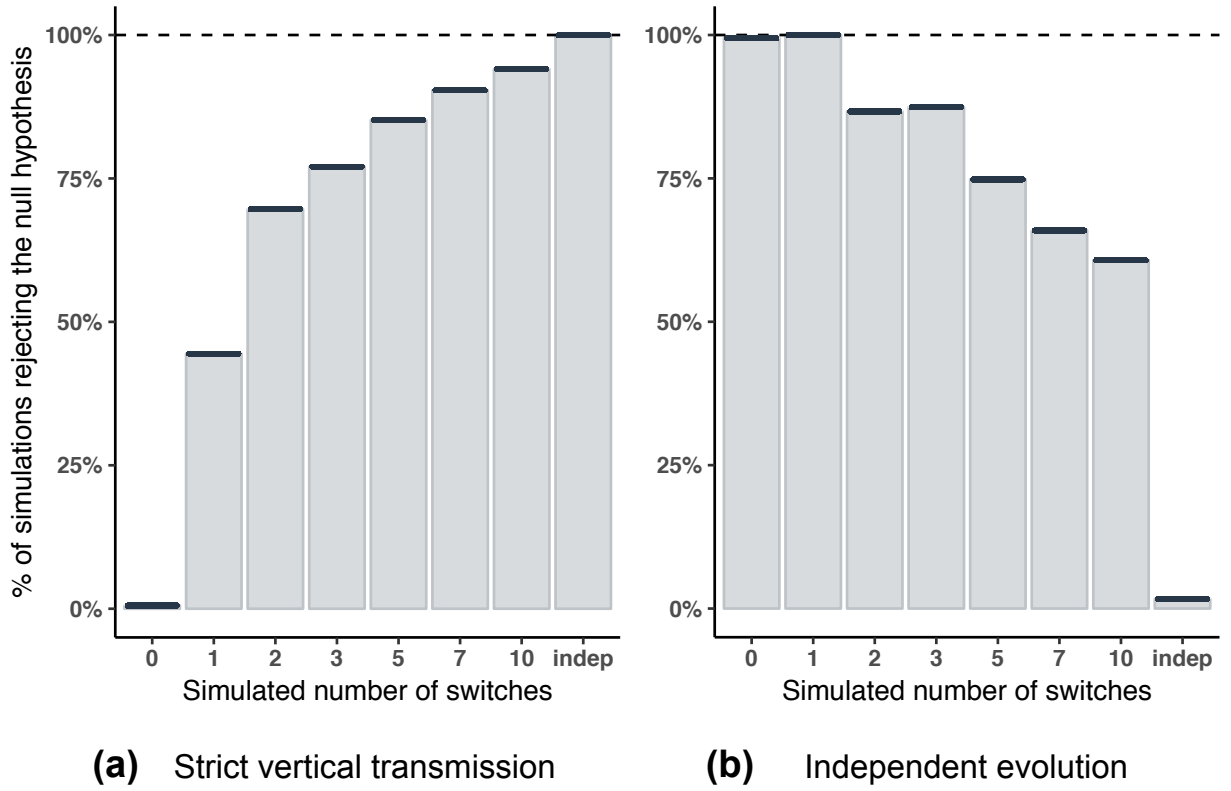


Figure I.1.4: Model selection. Percentage of simulated alignments for which the null hypothesis of strict vertical transmission (a) or independent evolution (b) is rejected under various evolutionary scenarios (strict vertical transmission, vertical transmission with a given number of switches, and independent evolution, referred in the figure as "indep").

families (*e.g.* Desulfurococcaceae, Pelobacteraceae, Rhodocyclaceae, and Eubacteriaceae) were entirely made of a transmitted OTU, while others also had many OTUs and/or sequences that were independent (*e.g.* Ruminococcaceae, Lachnospiraceae, and Coriobacteriaceae). Transmitted OTUs were in general not more abundant in a particular group of host species, except for the Prevotellaceae, that were overall more abundant in bonobos and chimpanzees than in gorillas and humans (Figure I.1.5b).

The sequence length, the proportion of segregating sites, and the intra-individual variability of the OTUs inferred as transmitted were similar to those of other OTUs (Supplementary Figure 9 and Supplementary Table 2), suggesting that HOME is not biased towards detecting vertical transmission in OTUs with specific characteristics. However, the intra-specific variability of OTUs inferred as transmitted tend to be smaller than that of other OTUs (Supplementary Table 5), which is consistent with our simulation results showing that the power to detect vertical transmission decreases with increasing intraspecific variability.

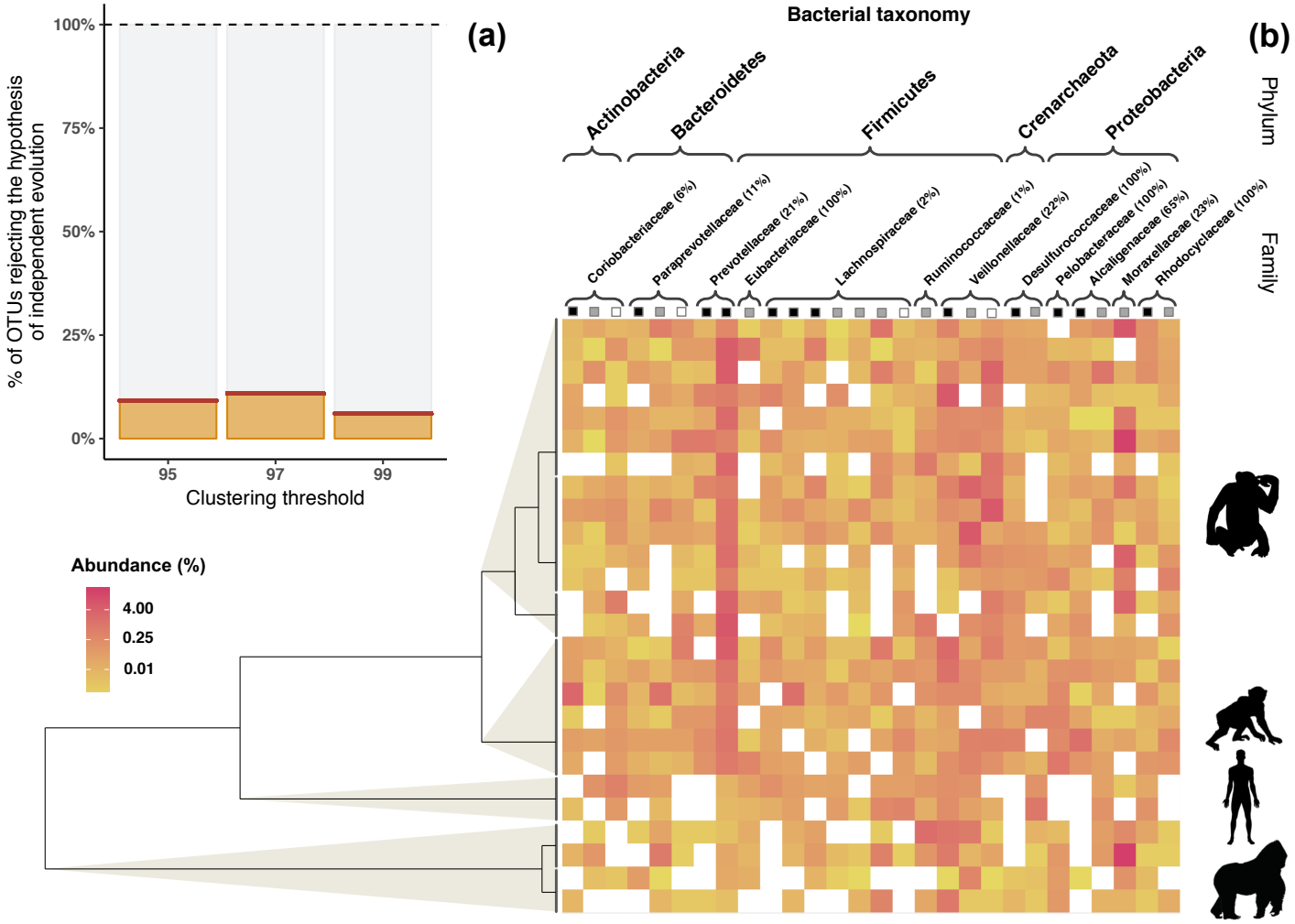


Figure I.1.5: Transmitted OTUs in the great ape microbiota. (a) Percentage of OTUs rejecting the hypothesis of independent evolution at the three clustering thresholds (b) Phylogenetic tree of great apes and their associated transmitted OTUs (black: 95% similarity threshold, grey: 97%, white: 99%). The percentage indicated in parenthesis for each family is the estimated percentage of transmitted raw reads in the family. The color of the heat map represents for OTU each host the percentage of raw reads of the OTU in the entire microbiota of the host. A white square means that the OTU is not found in the host.

Discussion:

We developed HOME, a likelihood-based approach for studying the inheritance of microbiota during the evolution of their hosts from metabarcoding data. We showed using simulations that even relatively short DNA reads can help identify modes of inheritance, without the need to build a microbial phylogenetic tree. Applying HOME to great apes microbiota data, we identified a set of transmitted gut bacteria that account on average for 8.4% of the total reads of the gut microbiota.

Our combination of model fitting and hypothesis testing helps identify modes of inheritance. We see the estimate of the number of switches as a good indicator of modes of inheritance (from strict vertical transmission for low ξ values to transmission with high rates of horizontal switches or environmental acquisition for high ξ values) rather than as an accurate estimation of past host-switches. We have indeed shown that ξ tends to be underestimated when quite many switches are simulated on a fixed host tree. In nature this underestimation may be even more pronounced, as our model ignores host-switches that happened in lineages not represented in the phylogeny, as a result of either extinction or undersampling (Szöllősi *et al.*, 2013b). In line with these results, we find that the hypothesis of strict vertical transmission is often not rejected when there are in fact host-switches. On the other hand, we can also estimate a positive ξ from data simulated under strict vertical transmission; however, in this case, a model with host-switches will in general not be selected when compared to a model of strict vertical transmission. Hence, if the hypothesis of strict vertical transmission is rejected, one can conclude with confidence that host-switches occurred (or that the microbial unit was environmentally acquired). Similarly, the hypothesis of independent evolution is often not rejected when the transmission is actually vertical with rather frequent host-switches, and rarely rejected in scenarios of environmental acquisition, such that when it is rejected, one can conclude with confidence that the microbial unit is transmitted. Said differently, our approach is conservative in its identification of transmitted OTUs; and when an OTU is identified as being transmitted, our approach is conservative in its identification of switches.

We assessed the performance of HOME in a limited set of conditions (*e.g.* host tree size, sequence length, substitution rates) calibrated on the great apes microbiota data. We can expect that the power of the model will increase with host tree size and the number of segregating sites in the microbial alignment. As the latter is a combination of sequence length, substitution rate, and hosts divergence times, there is no universal guidelines on the applicability of the model to a particular marker, sequencing technology, and host clade age. Rather, the marker and sequencing technologies must be adapted to the study system. For example, the 200-300 bp-long 16S rRNA V6 gene region sequenced with 454 sequencing used on great apes in our empirical application was enough to identify some transmitted microbial OTUs, but it probably missed others that had too low substitution rates to leave a detectable signal. Similarly, it might have a low resolution to detect

variability between host species that diverged more recently than the great apes. In such cases, using longer sequences and/or markers that evolve more quickly can be necessary. Finally, we can expect that PCR and sequencing errors will blur the signal and reduce the power to detect transmitted OTUs, although this should be limited by selecting the most abundant sequence representative of each OTU for each host.

HOME is currently best suited to the study of microbiota transmission in recent, well-sampled host clades in which no or few extinctions occurred, since it does not account for unsampled host lineages, nor for host extinctions. For example, HOME would be well adapted to the study of microbiota transmission in some vertebrates and invertebrate clades, for which microbiota sequencing data are already available (*e.g.* Amato *et al.*, 2019; Brooks *et al.*, 2016; Ren *et al.*, 2016). Ignoring extinction is reasonable at the small evolutionary scales of such groups or the great apes (Ochman *et al.*, 2010), but it would not be at larger evolutionary timescales such as across invertebrate or vertebrate species; in this case accounting for host switches from now-extinct lineages is necessary (Szöllősi *et al.*, 2013b). Another reason why HOME is currently better adapted to studying recent rather than ancient host clades is that it does not account for extinction of symbiont lineages, and therefore can only model the inheritance of OTUs shared across most species (*i.e.* core OTUs); the more divergent the host species, the less core OTUs there will be. Further developments of the model that would allow extending its relevance to a broader range of data include accounting for extinction and incomplete sampling in the host clade, as well as incorporating symbiont extinctions.

When it occurs, the support for vertical transmission of a given microbial unit arises from a phylogenetic signal in microbial sequences (*i.e.* a congruence between the phylogenetic similarity of host species and the molecular similarity of the microbes they host). However, such congruence can also arise from processes not accounted for in our model, such as geographic or environmental effects; for example, if there is a phylogenetic/molecular signal in the geographic or habitat distribution of hosts/microbes, or if the host environment creates microbial selective filters, this could result in a phylogenetic signal in microbial sequences that could be misleadingly interpreted as vertical transmission. We have not evaluated the robustness of our approach to such effects. Future developments could involve reconstructing ancestral areas/habitats or host environments on the host phylogeny in order to distinguish a phylogenetic signal truly driven by vertical transmission *versus* other effects.

In the construction of the model, we have made the important assumption that there is no microbial genetic variability within host species, such that each microbial OTU is represented by at most one sequence in each host. This is quite unlikely in natural microbial populations where multiple microbial strains can colonize a host species (Ellegaard & Engel, 2016). In our empirical application, we tackled this limitation by representing each host species by several individuals, using approximately zero-length branches to

split conspecifics in the host phylogeny. Although our simulations show that the statistical power of our tests decreases strongly when intraspecific variability is high, they also show that the hypothesis of environmental acquisition is rarely rejected when the acquisition is indeed environmental. Hence, HOME is unlikely to misleadingly identify transmitted OTUs, especially in the presence of intraspecific variability. Another (more satisfying) approach would be to directly account for intraspecific variability in microbial sequences in the likelihood computation; this could for example be done by representing the data by – at each tip of the host phylogeny and for each nucleotidic site – a vector of probabilities of states A, C, G, and T representing the intra-host relative abundance of the four bases at the given nucleotidic position. In this case, we would directly use the variation given at the level of amplicon sequence variants (ASVs; Callahan *et al.*, 2016). Alternatively, further developments of HOME incorporating horizontal host-switches without replacement (*i.e.* the persistence of both ancestral and newly-acquired symbionts in a lineage), as well as dynamics of duplication and recolonization, would allow better accounting for intra-host genetic variability. In addition, rather than considering each OTU as a separately evolving unit, it would be interesting to account for interactions between these units, that can for example lead to competitive exclusion (Koeppel & Wu, 2014) or interdependency (*e.g.* adaptive gene loss; Morris *et al.* (2012)), and are crucial aspects of microbial community assembly.

In the great apes gut microbiota, we found that the major part of the microbiota (91.6%) is constituted of bacteria which acquisition scenario is not distinguishable from one that is independent from the great apes phylogeny (Moeller *et al.*, 2013; Amato *et al.*, 2019). Still, we identified OTUs representing 8.4% of the total number of reads that are transmitted across generations during millions of years of evolution. Given the low phylogenetic signal in the geographic distribution of the hosts (see Ochman *et al.*, 2010), these OTUs are likely truly transmitted vertically. And given that HOME is conservative in its identification of transmitted OTUs, 8.4% is a lower bound estimate of the relative abundance of the microbiota that is vertically transmitted. Thus, our results suggest that the phylosymbiosis pattern observed by Ochman *et al.* (2010) is partially driven by vertically transmitted bacteria, as suggested by Sanders *et al.* (2014). Our approach offers the advantage of investigating the whole microbiota without an *a priori* on which families might be transmitted; it identified 12 microbial families with transmitted OTUs. This complements approaches that focus on few candidate families (*e.g.* Moeller *et al.*, 2016). Indeed, Moeller *et al.* (2016) used 3 specific primer pairs to focus on 3 bacterial families (Bacteroidaceae, Bifidobacteriaceae, and Lachnospiracea) and showed that phylogenies representing the Bifidobacteriaceae and Bacteroidaceae were congruent with the apes phylogeny, suggesting that co-diversification occurred in these two families. Unfortunately, neither Bifidobacteriaceae nor Bacteroidaceae were represented in the core OTUs in Ochman *et al.*'s data, even with a 95% similarity threshold: those bacteria were either not sampled, badly processed during DNA extraction and PCR, poorly taxonomically annotated, or too divergent to be merged into a single core OTU defined at 95%. Conversely,

while Moeller *et al.* (2016) did not find any signal of co-phylogeny in the Lachnospiraceae family, we found 3 transmitted OTUs belonging to this family. The authors investigated the phylogenetic relationships between all the amplified strains of Lachnospiraceae and whether they match the phylogenetic tree of great apes. This illustrates the utility of our approach, which investigates transmission modes of separate OTUs within bacterial families, rather than considering in a single evolutionary framework all the sequences from the same family.

Among the families in which we found transmitted OTUs, some are well known for having mutualistic properties. For example, the Lachnospiraceae, Paraprevotellaceae, and Rhodocyclales families are involved in breaking down complex carbohydrates in the gut; they have even evolved fibrolytic specialization in gut communities (Biddle *et al.*, 2013). These vertically transmitted fibrolytic bacteria, which have been co-diversifying for millions of years with the great apes, would thus constitute for the great apes a conserved reservoir of gut symbionts able to digest carbohydrates, and might have facilitated frequent and rapid dietary shifts during the evolutionary history of hominids (Head *et al.*, 2011; Muegge *et al.*, 2011; Hardy *et al.*, 2015). However, why these particular bacteria are faithfully vertically transmitted while other digesting gut bacteria seem largely environmentally acquired (or vertically transmitted with frequent host-switches) remains unclear.

DNA metabarcoding data for microbiota is being collected across multiple hosts at an unprecedented scale. Our approach allows identifying, among numerous microbial units, those that are vertically transmitted and potentially coevolving with their hosts. The current implementation of our model is entirely adapted to applications to other datasets using different sequencing techniques, clustering methods, and de-noising algorithms. Being able to identify vertically transmitted microbial units is an important step towards a better understanding of the role of microbial communities on the long-term evolution of their hosts.

References:

Amato KR, G. Sanders J, Song SJ, Nute M, Metcalf JL, Thompson LR, Morton JT, Amir A, J. McKenzie V, Humphrey G, *et al.* 2019. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. The ISME Journal 13: 576–587.

Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot M-FF, Charlat S. 2017. How long does *Wolbachia* remain on board? Molecular Biology and Evolution 34: 1183–1193.

Balbuena JA, Míguez-Lozano R, Blasco-Costa I. 2013. PACo: A novel procrustes application to cophylogenetic analysis (CS Moreau, Ed.). PLoS ONE 8: e61048.

Biddle A, Stewart L, Blanchard J, Leschine S. 2013. Untangling the genetic basis of fibrolytic specialization by Lachnospiraceae and Ruminococcaceae in diverse gut communities. Diversity 5: 627–640.

Bordenstein SR, Theis KR. 2015. Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. PLoS Biology 13: 1–23.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 8: 218–230.

Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. 2016. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history (D Relman, Ed.). PLOS Biology 14: e2000225.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, *et al.* 2010. QIIME allows analysis of high- throughput community sequencing data. Nature Methods 7: 335–336.

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. 2010. Jane: A new tool for the cophylogeny reconstruction problem. Algorithms for Molecular Biology 5: 1–10.

Edgar RC. 2013. UPARSE: Highly accurate OTU sequences from microbial amplicon reads. Nature Methods 10: 996–998.

Ellegaard KM, Engel P. 2016. Beyond 16S rRNA community profiling: Intra-species diversity in the gut microbiota. Frontiers in Microbiology 7: 1475.

Engel P, Moran NA. 2013. The gut microbiota of insects - diversity in structure and function. FEMS Microbiology Reviews 37: 699–735.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17: 368–376.

Ferretti L, Raineri E, Ramos-Onsins S. 2012. Neutrality tests for sequences with missing data. Genetics 191: 1397–1401.

Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nature Communications 8: 14319.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, *et al.* 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host and Microbe 17: 603–616.

Hardy K, Brand-Miller J, Brown KD, Thomas MG, Copeland L. 2015. The Importance of Dietary Carbohydrate in Human Evolution. The Quarterly Review of Biology 90: 251–268.

Head JS, Boesch C, Makaga L, Robbins MM. 2011. Sympatric chimpanzees (*Pan troglodytes troglodytes*) and gorillas (*Gorilla gorilla gorilla*) in Loango National Park, Gabon: dietary composition, seasonality, and intersite comparisons. International Journal of Primatology 32: 755–775.

Henry LM, Peccoud J, Simon JC, Hadfield JD, Maiden MJC, Ferrari J, Godfray HCJ. 2013. Horizontally transmitted symbionts and host colonization of ecological niches. Current Biology 23: 1713–1717.

Hommola K, Smith JE, Qiu Y, Gilks WR. 2009. A permutation test of host-parasite cospeciation. Molecular Biology and Evolution 26: 1457–1468.

Huelsenbeck JP, Rannala B, Larget B. 2000. A Bayesian framework for the analysis of cospeciation. Evolution 54: 352–364.

Koeppel AF, Wu M. 2014. Species matter: The role of competition in the assembly of congeneric bacteria. ISME Journal 8: 531–540.

Legendre P, Desdevises Y, Bazin E. 2002. A statistical test for host-parasite coevolution (RDM Page, Ed.). Systematic Biology 51: 217–234.

Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, Falush D, Stamer C, Prugnolle F, Van Der Merwe SW, *et al.* 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. Nature 445: 915–918.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey H V., Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, *et al.* 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences 110: 3229–3236.

McKenney EA, Maslanka M, Rodrigo A, Yoder AD. 2018. Bamboo specialists from two mammalian orders (Primates, Carnivora) share a high number of low-abundance gut microbes. Microbial Ecology 76: 272–284.

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

Moeller AH, Peeters M, Ndjango J-BJB, Li Y, Hahn BH, Ochman H. 2013. Sympatric chimpanzees and gorillas harbor convergent gut microbial communities. Genome Research 23: 1715–1720.

Morlon H, Lewitus E, Condamine FL, Manceau M, Clavel J, Drury J. 2016. RPANDA: An R-package for macroevolutionary analyses on phylogenetic trees. Methods in Ecology and Evolution 7: 589–597.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. mBio 3: e00036-12.

Muegge BD, Kuczynski J, Knights D, Clemente JC, Fontana L, Henrissat B, Knight R, Gordon JI, González A, Fontana L, *et al.* 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332: 970–974.

Ochman H, Worobey M, Kuo CH, Ndjango JBN, Peeters M, Hahn BH, Hugenholtz P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biology 8: 3–10.

Page RDM, Charleston MA. 1998. Trees within trees: Phylogeny and historical associations. Trends in Ecology and Evolution 13: 356–359.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Philippot L, Raaijmakers JM, Lemanceau P, Van Der Putten WH. 2013. Going back to the roots: The microbial ecology of the rhizosphere. Nature Reviews Microbiology 11: 789–799.

Pylro VS, Roesch LFW, Ortega JM, do Amaral AM, Tótola MR, Hirsch PR, Rosado AS, Góes-Neto A, da Costa da Silva AL, Rosa CA, *et al.* 2014. Brazilian Microbiome Project: Revealing the Unexplored Microbial Diversity-Challenges and Prospects. Microbial Ecology 67: 237–241.

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

Revell LJ. 2012. phytools: An R-package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rosenberg E, Zilber-Rosenberg I. 2016. Microbes drive evolution of animals and plants: The hologenome concept. mBio 7: 1–8.

Sanders JG, Powell S, Kronauer DJC, Vasconcelos HL, Frederickson ME, Pierce NE. 2014. Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Molecular Ecology 23: 1268–1283.

Schliep KP. 2011. phangorn: Phylogenetic analysis in R. Bioinformatics 27: 592–593.

Shapira M. 2016. Gut microbiotas and host evolution: scaling up symbiosis. Trends in Ecology and Evolution 31: 539–549.

Strimmer K, von Haeseler A. 2009. Genetic distances and nucleotide substitution models. In: The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing. Cambridge University Press, 111–125.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013a. Efficient exploration of the space of reconciled gene trees. Systematic Biology 62: 901–912.

Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013b. Lateral gene transfer from the dead. Systematic Biology 62: 386–397.

Truszkowski J, Goldman N. 2016. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. Systematic Biology 65: 328–333.

de Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: Methods for testing, evidence from natural associations

and relation to coevolution. New Phytologist 198: 347–385.

Zilber-Rosenberg I, Rosenberg E. 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiology Reviews 32: 723–735.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3dZWWjs` or by scanning:

Article 2: Limited evidence for microbial transmission in the phylosymbiosis between Hawaiian spiders and their microbiota

Authors: Benoît Perez-Lamarque[1,2], Henrik Krehenwinkel[3], Rosemary Gillespie[4], Hélène Morlon[1]

[1] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

[3] Department of Biogeography, Trier University, Trier, Germany

[4] Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA

Abstract

The degree of similarity between the microbiota of host species often mirrors the phylogenetic proximity of the hosts. This pattern, referred to as phylosymbiosis, is widespread in animal and plant kingdoms. While phylosymbiosis was initially interpreted as the signal of symbiotic transmission and coevolution between microbes and their hosts, it is now recognized that similar patterns can emerge even if the microbes are environmentally acquired. Distinguishing between these two scenarios, however, remains challenging. We recently developed HOME (HOst-Microbiota Evolution), a cophylogenetic model designed to detect transmitted microbes and host-switches from amplicon sequencing data. Here, we apply HOME to the microbiota of Hawaiian spiders of the genus *Ariamnes*, which experienced a recent radiation on the archipelago. We demonstrate that although Hawaiian *Ariamnes* spiders display a significant phylosymbiosis, there is little evidence of microbial vertical transmission. Next, we perform simulations to validate the absence of transmitted microbes in *Ariamnes* spiders. We show that this is not due to a lack of detection power because of the low number of segregating sites nor an effect of phylogenetically-driven or geographically-driven host-switches. *Ariamnes* spiders and their associated microbes therefore provide an example of a pattern of phylosymbiosis likely emerging from processes other than vertical transmission.

Keywords: microbiota, phylosymbiosis, vertical transmission, host-filtering, Hawaiian arthropods.

Author contributions: All the authors designed the study. BPL, HK, and RG gathered the data and BPL performed the analyses. BPL and HM wrote the first version of the manuscript and all authors contributed to the revisions.

Acknowledgments: The authors acknowledge E. Armstrong, J. Lim, A. Rueda, T. Schol, S. Kennedy, and S. Prost for helpful discussions. BPL and HM also thank D. de Vienne, M. Elias, M.-A. Selosse, F. Martos, L. Aristide, C. Fruciano, I. Quintero, S. Lambert, I. Overcast, O. Maliet, A. Silva, and G. Sommeria for comments on an early version of the manuscript. This work was supported by a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL and the École Doctorale FIRE – Programme Bettencourt. RG acknowledges support from the National Science Foundation, DEB 1241253. HM acknowledges support from the European Research Council (grant CoG-PANDA). BPL and HM also acknowledge support from the iBioGen Twinning project (H2020 research and innovation program, grant agreement No 810729).

Data availability: The raw data can be found on dryad (`https://datadryad.org`) DOI: https://doi.org/10.5061/dryad.nzs7h44qj. The implementation of HOME used in this study and a tutorial are available on GitHub (`https://github.com/BPerezLamarque/HOME`).

Citation: Perez-Lamarque B, Krehenwinkel H M, Gillespie R, Morlon H. Limited evidence for microbial transmission in the phylosymbiosis between Hawaiian spiders and their microbiota, ***under review***.

Introduction:

Most multicellular organisms such as plants and animals host complex microbial communities, referred to as microbiota, which provide important functions to their hosts (Selosse *et al.*, 2004; McFall-Ngai *et al.*, 2013). Although these microbial communities can fluctuate over short time scales and according to external variables such as animal diet or soil composition (Parfrey & Knight, 2012; Tkacz *et al.*, 2015; Kennedy *et al.*, 2020), microbiota of host individuals from the same species often tend to be more similar than microbiota from different host species (Hacquard *et al.*, 2015). Over long-time scales, the extent to which these microbiota, and the functions they provide to their hosts, are conserved will depend on the relative tendency of microbes to colonize host individuals at each generation (Nyholm & McFall-Ngai, 2004), which is influenced by their modes of inheritance. At the two extremes, microbes can be either transmitted from generation to generation (vertical transmission, *e.g.* directly through the maternal germline or by social contacts with host relatives; Moran *et al.*, 2008; Funkhouser & Bordenstein, 2013), or acquired from the environment during the lifetime of each host individual independently of the previous host generation (environmental acquisition; Bright & Bulgheresi, 2010;

Funkhouser & Bordenstein, 2013). In the latter case, the maintenance of the microbes from host generation to host generation depends on their presence in the host environment and their availability to colonize the host niche, which can be seen as an ecological filter (Moran & Sloan, 2015).

Microbiota of closely related host species are often more similar than those of distantly related species, such that host dendrograms constructed from the similarities of whole microbiota communities tend to mirror the host phylogeny (Brooks *et al.*, 2016; Lim & Bordenstein, 2020). This pattern, referred to as phylosymbiosis, has for example been documented for the gut microbiota of primates (Ochman *et al.*, 2010; Amato *et al.*, 2019) and arthropods (Armstrong *et al.*, 2020), or in the roots of plants (Yeoh *et al.*, 2017). However, how and why this pattern emerges has been intensively debated (Moran & Sloan, 2015; Kohl, 2020). In particular, phylosymbiosis is expected to emerge if some specific microbial lineages are vertically transmitted during host evolution (Ochman *et al.*, 2010; Sanders *et al.*, 2014; Moeller *et al.*, 2016, 2018). In this case, these vertically transmitted microbial lineages follow host diversification, resulting in co-phylogenetic patterns between host species and each transmitted microbial lineage (Moeller *et al.*, 2016). Alternatively, phylosymbiosis can emerge in the absence of vertical transmission, for instance, if the community assembly of microbes acquired from the environment is dominated by mechanisms of ecological filtering by the hosts, and if the host traits involved in this filtering are phylogenetically conserved (Moran & Sloan, 2015; Mazel *et al.*, 2018). In the latter case, no specific congruence between the host phylogeny and the individual microbial phylogenies is expected (Moran & Sloan, 2015). Some correlative approaches are available to investigate whether phylosymbiosis is supported by recent or ancient microbial divergences, such as the beta diversity sensitivity analysis (Sanders *et al.*, 2014; Amato *et al.*, 2019), however they cannot directly assess whether a pattern of phylosymbiosis is linked to the vertical transmission of individual microbial lineages or to alternative processes such as ecological host-filtering.

An alternative approach to inferring if and which microbial lineages are transmitted among members of whole microbial communities would be to use cophylogenetic methods, which quantify the congruence between trees (de Vienne *et al.*, 2013; Kohl, 2020), however this is challenging for several reasons. First, transmitted microbes can experience not only vertical transmissions, but also events of horizontal switching between host lineages, which result in a loss of phylogenetic congruence between host and microbial lineages (Charleston & Perkins, 2006). Horizontal switches between particular host lineages are also expected to be more likely, for instance between closely related host species (that likely represent similar niches for the microbes), or between host species sharing the same geographic area (Charleston & Robertson, 2002). Such preferential host-switches can strongly influence the observed cophylogenetic patterns (de Vienne *et al.*, 2007). Second, because of the short length of the amplicons used to characterize microbiota in high throughput sequencing studies (*e.g.* the 16S rRNA gene; Taberlet *et al.*, 2012), the micro-

bial DNA sequences available have accumulated only a few mutations since their host diverged, providing limited information on the evolutionary history of the transmitted microbes (Ochman *et al.*, 1999). The difficulty in reconstructing a robust phylogeny for each microbial lineage is one of the most problematic challenges, especially as phylosymbiosis is often observed in recent host radiations, whereas it is "erased" by various factors such as diet shifts over longer timescale (Groussin *et al.*, 2017; Baldo *et al.*, 2017). To address these challenges, we recently developed a likelihood-based model, called HOME (HOst-Microbiota Evolution; Article 1), designed to infer modes of microbial inheritance in the presence of host-switches without reconstructing the phylogenetic tree of the microbial lineages. Instead, HOME directly models the evolution of the microbial DNA sequences on the host phylogeny, with potential host-switches, and tests whether this model of host-dependent evolution is supported in comparison with scenarios where microbial sequences evolve independently.

Here, we investigated the modes of inheritance of the bacterial microbiota of a lineage of Hawaiian spiders *Ariamnes* that exhibit a significant pattern of phylosymbiosis (Armstrong *et al.*, 2020) to examine whether this empirical pattern of phylosymbiosis (at the whole microbiota community level) is at least partially explained by transmitted microbes (at the level of individual microbial lineages). The Hawaiian *Ariamnes* spiders are non-model organisms and whether or not their microbes are transmitted is unknow. They are predators of other spiders (Kennedy *et al.*, 2018) and are mostly restricted to the wet forest habitats (Gillespie *et al.*, 2018). Their radiation into 15 species within the last 2 million years across the Hawaiian archipelago shows a classic pattern of colonization from older to younger islands, which results in strong geographical clustering (Gillespie & Rivera, 2007; Gillespie *et al.*, 2018). Because these species are highly specialized (similar habitat and very narrow and conserved diet relative to other spiders) and closely related within an island, we would expect their important niche conservatism to favor microbial transmission and that their geographical clustering would influence the host-switch dynamics. We applied HOME and, contrary to these expectations, we did not find any transmitted microbial symbiont, suggesting that *Ariamnes*' microbes are not transmitted over long timescales. To confirm this finding, we performed further validations of the method to show that (i) HOME performs well to infer the modes of inheritance of microbial lineages even with few informative segregating sites in the DNA sequences and that (ii) preferential host-switches are unlikely to bias our conclusions.

Methods:

Study system: microbiota of Hawaiian spiders:

Hawaiian *Ariamnes* spiders had been sampled under the leaves in understory vegetation in 8 sites with similar abiotic conditions (temperature and precipitation) to control for environmental variables and to guarantee that all individuals had similar micro-

niches (Armstrong *et al.*, 2020). We selected the individuals for which both genome-wide sequencing of the host and 16S rRNA metabarcoding of their associated microbiota have been performed: We obtained 63 "gold ecomorph" individuals (Gillespie *et al.*, 2018) from one species on Molokai (*A. n. sp.*), one on West Maui (*A. melekalikimaka*), and one on Hawaii Island (*A. waikula*).

We used a robust phylogenetic tree of the 63 individuals reconstructed using genomic ddRAD markers and 16S rRNA metabarcoding data characterizing their microbiota from (Armstrong *et al.*, 2020). In short, the calibrated host phylogenetic tree was obtained using the Stacks pipelines (Catchen *et al.*, 2013), IQ-TREE (Nguyen *et al.*, 2015) with 100 bootstraps, and r8s (Sanderson, 2002) (see Supplementary Methods 1 and Armstrong *et al.* (2020) for details). Spider-associated bacterial communities were studied using short DNA metabarcoding sequences from the 16S rRNA gene. Microbiota DNA extractions were performed using the Gentra Puregene Tissue Kit (Qiagen, Hilden, Germany) on the spiders' abdomens, which contain the gut as well as other organs such as the gonads (Kennedy *et al.*, 2020). Bacterial 16S rRNA genes were targeted using a primer pair amplifying approximately 310 base pairs (bp) of the V1-V3 variable regions (Gibson *et al.*, 2014). The amplicon library was sequenced using Illumina MiSeq technology generating 2×300 bp paired reads. Negative controls (extraction blanks and no template controls) were carefully performed.

The corresponding microbiota raw data (obtained from Armstrong *et al.*, 2020) encompassed 4,932,236 microbial reads, that were assembled, demultiplexed, and quality checked using VSEARCH v2 (Rognes *et al.*, 2016). We clustered reads according to their sequence similarities into Operational Taxonomic Units (OTU) using two different algorithms. We first used the Swarm v2 (Mahé *et al.*, 2015), using the fastidious option, that groups reads into OTUs without specifying a global similarity threshold, and thus can accurately identify clustered structure at a finer scale. We performed a second clustering using a classical OTU clustering at 97% similarity using VSEARCH. We assigned a taxonomy to each OTU using the SILVA database (Quast *et al.*, 2013), filtered-out chimera, and built an OTU table indicating for each OTU its abundances in the different spiders' microbiota. Finally, non-bacterial OTUs and contaminant OTUs present in high abundances in the negative controls were filtered out of the OTU table: We obtained a total of 413 Swarm OTUs and 414 97% OTUs, which respectively correspond to a total of 1,297,307 and 1,178,325 reads.

Assessing phylosymbiosis:

Following Brooks *et al.* (2016) and Lim & Bordenstein (2020), we assessed phylosymbiosis by (i) performing a Mantel test between the host phylogenetic distances and the microbiota beta diversities and (ii) evaluating using matching cluster analyses (Bogdanowicz & Giaro, 2013) the topological congruence between the *Ariamnes* phylogenetic tree and the microbiota dendrogram reconstructed using a hierarchical clustering from

the beta diversities (Supplementary Methods 2). Weighted or unweighted beta diversities were computed using OTU tables rarefied at 5,000 reads per sample.

Inferring transmitted symbionts:

The inheritance modes among the microbial OTUs were inferred independently for each OTU using HOME (Article 1). Each OTU was characterized by a nucleotide alignment made of the microbial sequences obtained across the different host lineages, with at most one single representative DNA sequence per host individual. In short, HOME uses the intra-OTU variation (segregating sites) contained in the nucleotide alignment to test whether each microbial OTU has likely evolved on the host phylogeny by vertical transmission or alternatively has been acquired from the environment. It assumes that for a given OTU, microbial populations, represented by a DNA sequence in each host lineage, (i) are vertically transmitted along branches on the host phylogeny, (ii) can experience DNA substitutions with a constant rate μ, which generate segregating sites in the OTU alignment, (iii) are inherited at host splitting events by the two daughter host lineages, and (iv) can experience a certain number of host-switches ξ, where one microbial sequence from a donor host branch is horizontally transmitted at a given time to another receiving branch where it replaces the previous sequence. By default, host-switches are assumed to be uniformly distributed on the host branches. For each OTU independently, HOME uses a combination of likelihood-based and simulation-based approaches to estimate ξ and μ, and to test whether a scenario of transmission is more likely than a scenario of host-independent evolution where the links between microbial sequences and host lineages are randomized (see Article 1 for more details).

To run HOME, we selected the OTUs shared by at least five individual spiders, as HOME cannot perform well with lower occurrence. In addition, based on the content of negative PCR controls and on previous estimates (Minich *et al.*, 2019), we assumed that if an OTU occurs with less than 5 reads in a spider, it is likely the result of cross-contaminations during the library preparation, and consider the OTU as absent in this spider's microbiota. For a given OTU, we selected one representative sequence per host spider by taking the most abundant read confidently assigned to this OTU present in the spider microbiota. Neglecting the microbial intra-individual variation is equivalent to considering that host individuals are colonized by only one unique microbial strain per OTU, and that the intra-individual variability observed in the data is caused by PCR and sequencing errors. To relax this hypothesis, we repeated the analyses by instead picking, when available, the second most abundant read as the representative sequence of one OTU in one host spider, although these sequences were likely artifacts (Schloss *et al.*, 2011). We then assembled the sequence alignment for each OTU and ran HOME separately. As HOME does not model microbial extinction, in the case of incomplete representation of an OTU (either because the OTU is truly absent or undetected), we pruned out of the phylogenetic tree the host spiders where the OTU was absent (Article 1).

Finally, to confirm that these few OTUs shared by multiple host individuals contributed for the pattern of phylosymbiosis at the whole community level (encompassing all "shared" and "unshared" OTUs), we randomized the "unshared" OTUs among *Ariamnes* samples while keeping the "shared" OTUs untouched (and *vice versa*) and we re-assessed phylosymbiosis using Mantel tests.

Simulating the performance of HOME under low intra-OTU variation and preferential host-switches :

To confirm that HOME would detect transmitted OTUs in *Ariamnes* microbiota despite the recent host divergences and the likely occurrence of preferential host-switches, we tested the performance of HOME using simulations by (i) checking the statistical power of HOME when there are only a few segregating sites in the microbial OTU alignments and (ii) testing the effect of phylogenetically-driven and geographically-driven host-switches.

First, we simulated, on the *Ariamnes* host tree, microbial phylogenies of OTUs evolving under vertical transmissions with 0, 5, 15, 25, or 35 host-switch(es), or OTU phylogenies that evolved independently from the host phylogeny. For each scenario, we simulated 15 independent OTUs. Given the phylogenetic tree of each OTU, we simulated the corresponding evolution of a nucleotide alignment of 300 bp, with a probability 0.1 for each site to be variable under a stochastic K80 (Kimura, 1980) nucleotide substitution process (with a ratio of transition/transversion rate $\kappa = 0.66$) and we tested the effect of very low relative substitution rate ($\mu = 0.1$), compared to intermediate values ($\mu = 1.5$). We then ran HOME on each alignment independently. Furthermore, as some OTUs occurred in only a few host individuals (either because they were absent or undetected), we tested the effect of low OTU occurrence on the statistical power of HOME by simulating the alignment corresponding to each OTU on a host phylogeny randomly pruned to 5 or 20 tips.

Next, we simulated vertically transmitted microbial symbionts which experience events of host-switches driven by host relatedness (de Vienne *et al.*, 2007). We considered that the probability that there is a host-switch between times t and $t + dt$ depends on the phylogenetic relatedness between pairs of host lineages among the $N(t)$ other coexisting hosts at this time such that:

$$P(\text{host switch at time } t) \sim \frac{1}{N(t) - 1} \sum_{i \in [1, N(t)]} \sum_{j \in [1, N(t)]; j \neq i} e^{-h d_{i,j}(t)}$$

where $d_{i,j}(t)$ represents the phylogenetic distance, measured as branch length, between the coexisting hosts i and j at time t, and h is a parameter tuning the effect of host-phylogenetic relatedness (if $h = 0$, there is no effect of host relatedness, and the higher h is, the more likely host-switches occur between closely related hosts). If a host-

switch occurs at time t, a pair (i,j) of host lineages involved in the host-switch is chosen proportionally to $e^{-hd_{i,j}(t)}$ (if $h > 0$, pairs of hosts that are phylogenetically distant - large $d_{i,j}(t)$ - are unlikely to be chosen). For $h = 0$ (uniform distribution of host-switches), the probability of a switch is proportional to the number of host lineages at time t and pairs of hosts involved in the switch are chosen uniformly. We performed the same simulations as detailed above (no host-switch) using 4 h values: $h \in (0, 1, 10, 100)$.
Similarly, we simulated vertically transmitted microbial symbionts which experience events of host-switches that are more likely between hosts sharing the same geographical area (*i.e.* same sampling site). We assumed that hosts occupy a unique discrete area, which can change through time. We first reconstructed the ancestral biogeography of the host phylogeny using stochastic mapping (*make.simmap* function, phytools R-package; Revell, 2012) considering that migrations between areas are punctuated events occurring with different estimated probabilities, represented by a symmetrical matrix of transition Q. We considered g, the probability for a host-switch to occur between hosts from different areas divided by the probability for a host-switch to occur between hosts of the same area: if $g = 1$, host-switches between hosts from different areas are as likely as host-switches within the same area, whereas $g = 0$ corresponds to a scenario where host-switches only occur between hosts sharing the same area. We considered that the probability that a host-switch occurs at time t depends on the total number of hosts sharing the same area at time t and the number of hosts alone on their own area at time t multiplied by g, such that:

$$P(\text{host switch at time } t) \sim \sum_{A \in \text{areas}, N_A(t) > 1} N_A(t) + \sum_{A \in \text{areas}, N_A(t) = 1} g N_A(t)$$

with $N_A(t)$ being the number of hosts in area A at time t.
If a host-switch occurs at time t, pairs of hosts are then chosen with a relative weight of 1 if they are in the same area, or g if not. These simulations could be further improved by considering geographic distances between the different areas. We performed the same simulations as detailed above using 4 g values: $g \in (1, 0.5, 0.25, 0)$. For each simulated OTU, we used a different stochastic mapping of the ancestral biogeography to consider the uncertainty in the reconstruction.

In addition to these tests, we investigated the ability of HOME to detect preferential host-switches on the *Ariamnes tree* (Supplementary Methods 3).

Results:

Phylosymbiosis and inheritance of the microbiota of Hawaiian spiders:

As previously shown in Armstrong *et al.* (2020), the evolutionary history of the 63 sampled *Ariamnes* spiders reconstructing using ddRAD markers presented a significant

clustering by geographic areas as indicated by the ancestral state (Figure I.2.1a). Looking at their associated microbiota as a whole using 16S rRNA amplicon sequencing, we found significantly congruent topologies between the host spider phylogeny and their microbiota dendrogram reconstructed from the weighted or unweighted beta diversities (Figures I.2.1b & Supplementary Figure 1). Phylosymbiosis was also confirmed by Mantel tests indicating a significant correlation between the host patristic distances and the microbiota dissimilarities when using weighted beta diversity indices (Supplementary Figure 1). Conversely, Mantel tests were no longer significant when using unweighted indices, suggesting that shifts in the presence/absence of abundant OTUs participated to phylosymbiosis.

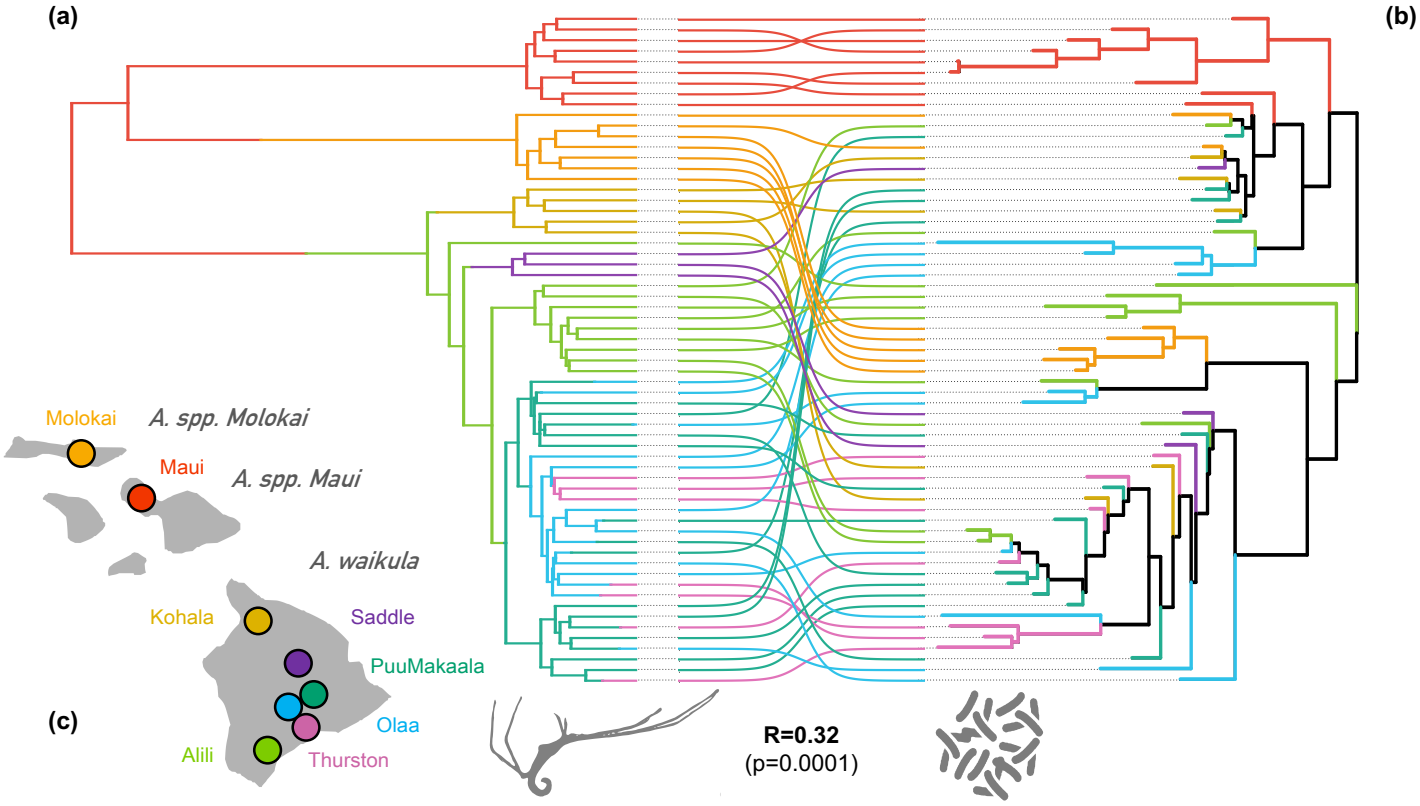


Figure I.2.1: Phylosymbiosis in the microbiota of Hawaiian spiders. (a) Phylogenetic tree of the host *Ariamnes* spiders obtained using ddRAD markers. Branches are colored according to the geographic area occupied by the spiders as inferred using a Bayesian ancestral state reconstruction. Interspecific nodes are supported at 100%, whereas most intraspecific nodes have bootstrap supports greater than 70%. (b) Microbiota dendrogram obtained from the Bray-Curtis dissimilarities as a measure of beta diversities comparing the Swarm composition between spiders' microbiota (similar patterns are obtained with 97% OTUs - not shown). Tips are colored according to the geographic area of the host spiders, and internal branches are colored accordingly if all their descendant tips come from the same area. Colored links represent the interaction between spiders and its microbiota. The Mantel test between the host phylogenetic distances and the microbiota Bray-Curtis dissimilarities indicated a significant correlation. (c) Map of the Hawaiian archipelago where the *Ariamnes* spiders were sampled.

Next, we used HOME to infer the inheritance modes for each of the 96 Swarm OTUs and 103 97% OTUs that were shared by at least 5 spider individuals. When selecting the most abundant sequence per host individual, only 51 Swarm OTUs (resp. 66 97% OTUs) had at least one segregating site, while we had 81 Swarm OTUs (resp. 90 97% OTUs) when selecting the second most abundant sequence. These "shared" OTUs presented a relatively low number of segregating sites (Supplementary Figure 2) and occurred in 5 to 55 host individuals (Supplementary Figure 3), but they represented 88% (Swarm) and 89% (97% OTU) of the total bacterial reads. When applying HOME, no OTU rejected the null hypothesis of host-independent evolution (Figure I.2.2), except 2 Swarm OTUs out of 132, but this ratio falls into the global type-I error of HOME (Article 1; Supplementary Figure 4b). In addition, we confirmed that these small fraction of "shared" OTUs used to run HOME (only ∼25% of the OTUs occurred across multiple host individuals) were mainly responsible for the global pattern of phylosymbiosis: we still found a significant phylosymbiosis when randomizing the "unshared" OTUs while keeping the "shared" OTUs untouched, but conversely, the Mantel correlations were no longer sig-

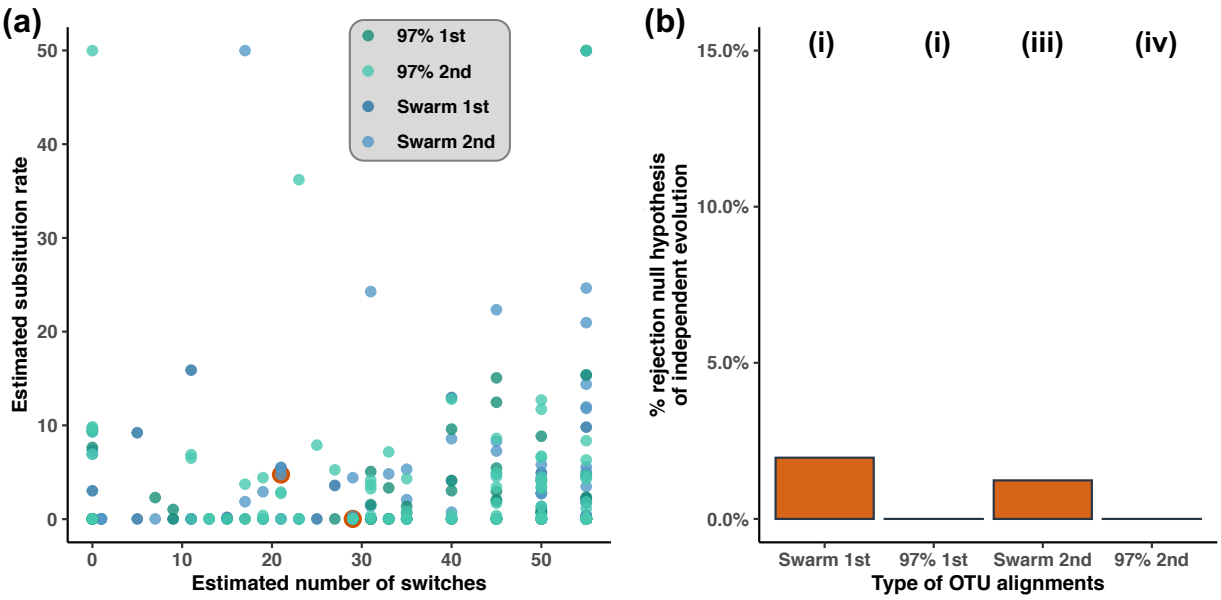


Figure I.2.2: HOME results on OTU alignments from the *Ariamnes* microbiota. (a) Estimated substitution rate as a function of the estimated number of switches for the different OTUs. Each dot corresponds to an empirical OTU and is colored according to the type of OTU (Swarm or 97% OTUs) and the representative sequence (the most or the second most abundant sequence per sample). (b) Percentage of OTU rejecting the null hypothesis of independent host-microbial evolution. Columns (i) and (iii) correspond to Swarm OTUs, whereas (ii) and (iv) are 97% OTUs. Representative OTU sequences in columns (i) and (ii) are obtained by taking the most abundant sequence per sample, whereas in (iii) and (iv) they correspond to the second most abundant sequences per sample. The two OTUs that rejected the null hypothesis of independent evolutions (columns (i) and (iii)) belong to the genera *Bacillus* and *Erythrobacter* respectively, occurred in 12 and 51 host individuals respectively, have 1 and 13 segregating sites respectively, and have an estimated number of host-switches equaled to 12 and 21 respectively (very high compared to the number of hosts where they occurred): they are likely false-positives. They are highlighted by two surrounding orange circles on panel (a).

nificant when randomizing the "shared" OTUs and keeping untouched the "unshared" ones (Supplementary Table 1).

Testing the performance of HOME with simulations:

To test the effect of low substitution rates (μ) on the performance of HOME, we simulated OTU alignments with very low numbers of segregating sites (μ=0.1; Supplementary Figures 5) similar to those of the empirical OTU alignments (Supplementary Figure 2). Compared to more variable alignments (μ=1.5; Supplementary Figure 5), we found as expected that the ability to recover simulated parameters (*i.e.* the number of host-switches (ξ) and the substitution rate (μ)) decreases when the simulated substitution rate is low (Supplementary Figures 6 & 7). The approach tends to underestimate the inferred number of host-switches when there are many ($F_{1,68}$=11.7, p=0.001; Supplementary Figure 8a), however we still found a positive correlation between the number of simulated (ξ) and estimated host-switches ($\hat{\xi}$). Similarly, HOME correctly estimates the simulated low substitution rate, but tends to overestimate it when ξ is large (t_{69}=3.0, p=0.004; Supplementary Figure 8c). $\hat{\xi}$ and $\hat{\mu}$ values estimated from OTU alignments simulated independently from the host phylogeny were significantly higher than those estimated from transmitted OTUs (Supplementary Figure 8), and the null hypothesis of host-independent evolution was never rejected for these alignments (Figure I.2.3 and Supplementary Figure 8d). Conversely, the null hypothesis of host-independent evolution was on average rejected for 50% of the OTU alignments transmitted with no or few (less than 15) host-switches (intermediate statistical power; Supplementary Figure 8d); this statistical power decreased with the number of simulated host-switches (Supplementary Figure 8d). These results also depend on the number of hosts in which the OTU is found, and the statistical power decreased at 40% for OTUs occurring in only 20 hosts and below 10% for OTUs occurring in only 5 hosts (Supplementary Figure 9). Given the statistical power of HOME in this system (50% when the number of segregating sites is low and the number of hosts in which the OTU is found is high), we can conclude with high confidence that, if any, there are at most 4 core OTUs that are transmitted (Supplementary Figure 4a) and no more than 6 OTUs occurring in only 20 hosts that are transmitted. Therefore, most bacterial OTUs from the *Ariamnes* microbiota are likely independently evolving from their host spiders.

Similarly, when testing whether preferential host-switches could affect the performances of HOME, we found a good ability to recover simulated parameter values (Supplementary Figure 6 & 7), especially when the simulated substitution rate was high (μ=1.5). Both types of preferential host-switches (phylogenetic relatedness or geographic dependencies) affected the estimation of the parameters in the same way: the estimated number of switches ($\hat{\xi}$) and the estimated substitution rates ($\hat{\mu}$) tend to decrease when the effect of preferential host-switches is higher (Supplementary Figure 6 & 7). Transmitted OTUs simulated under preferential host-switches tend to be inferred as strictly vertically transmitted OTUs (*i.e.* they tend to have estimated parameters similar to those of OTUs

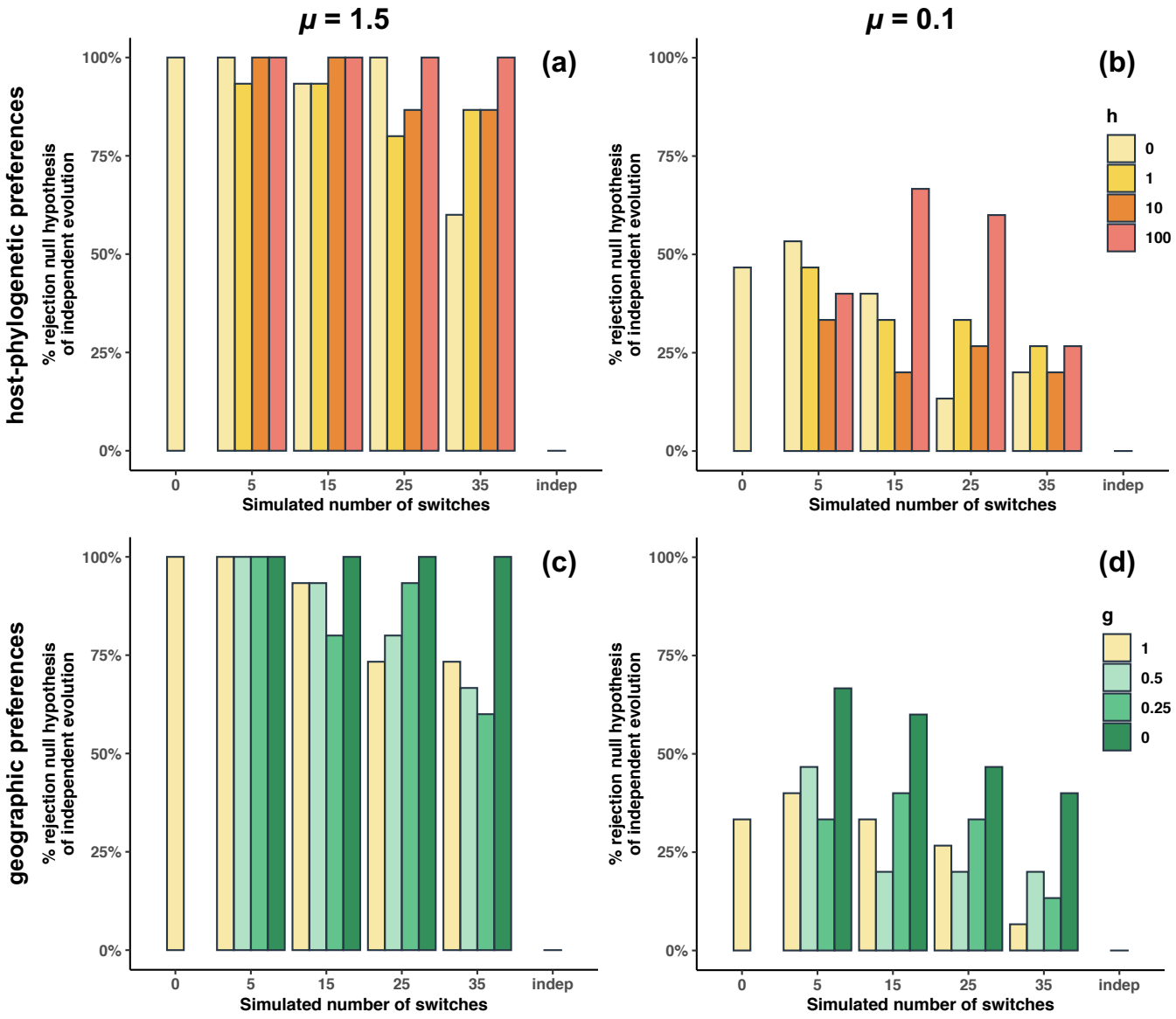


Figure I.2.3: Evaluating the effect of substitution rates and preferential host-switches on the rejection of the null hypothesis of host-independent evolution. Percentage of simulated OTU alignments that reject the null hypothesis of independent host-microbial evolution, according to the different evolutionary scenario simulated on the 63 tips *Ariamnes* tree: strict vertical transmission (ξ=0), vertical transmission with ξ host-switches (ξ>0), or independent evolution (referred to as "indep"). Four cases are tested: high substitution rate (a & c) relative to low substitution rate (b & d), and the possibility of host-phylogenetic preferences in the simulated switches (a & b) or geographic preferences (c & d). Simulated values h=0 and g=1 correspond to uniformly distributed host-switches, whereas h>0 and g<1 indicate a certain degree of preferential host-switches.

simulated under strict transmission: small $\hat{\xi}$ and $\hat{\mu}$). Similarly, the statistical power of the test of host-independent evolution significantly increased with preferential host-switches (Figure I.2.3), whatever the type of preferential host-switches or the simulated substitution rate: simulated transmitted OTUs that experienced preferential host-switches were more frequently inferred as transmitted OTUs than OTUs that experienced uniformly distributed host-switches (Figure I.2.3). Thus, preferential host-switches is very unlikely to negatively affect the ability of HOME to detect transmitted OTUs.

Discussion:

We used HOME, a cophylogenetic model-based approach well-adapted to deal with microbiota datasets, to assess whether phylosymbiosis can emerge without vertical microbial transmission in an empirical system. Indeed, we found that the significant phylosymbiosis pattern of recently-diverging spiders across the Hawaiian archipelago is likely not explained by vertically transmitted microbes. This result is not due to the low number of segregating sites nor to preferential host-switches.

Inferring transmitted symbionts:

Simulations with very low substitution rates showed that the statistical behavior of HOME remains acceptable when there are very few segregating sites in the microbial alignment: although the power to reject the null hypothesis of host-independent evolution is reduced, the type-I error rate remains very low (*i.e.* symbionts that evolved independently from the host phylogeny are rarely inferred as transmitted). This indicates that HOME can be applied to molecular microbial markers that evolve slowly (*e.g.* 16S rRNA gene) and when host divergences are recent, such as among the species of *Ariamnes* spiders included here that diverged only <2 million years ago (Gillespie *et al.*, 2018). The ability to study the mode of inheritance in microbiota over short time-scales is one of the main advantages of HOME compared to other models, such as Jane (Conow *et al.*, 2010) or ALE (Szöllősi *et al.*, 2013), that need reconstructed microbial trees.

The presence of preferential host-switches does not reduce the statistical power of HOME; on the contrary microbial alignments simulated under preferential host-switches are more likely to be inferred as strictly vertically transmitted symbionts. This result is not surprising as, if host-switches preferentially occurred between host lineages that are phylogenetically related, the microbial phylogeny tends to be more congruent with the host phylogeny than if host-switches occurred uniformly. Similarly, as phylogenetically related host lineages also tend here to occupy the same geographic area, we could expect the microbial phylogenies resulting from geographic-dependent host-switches to be more similar to the host phylogeny, than if host-switches occurred uniformly. Thus, in our simulated scenarios, preferential host-switches strengthen the host phylogenetic signal within the microbial alignment and increase the ability of HOME to detect transmitted OTUs.

Our test of the ability of HOME to detect preferential host-switches if they had occurred on the *Ariamnes* phylogeny showed that such detection is difficult, and that both phylogenetically- or geographically-driven host-switches leave similar signals in the microbial alignments and are therefore undifferentiable (Supplementary Results). This does not imply that preferential host-switches cannot be inferred from any host phylogeny, as the *Ariamnes* phylogeny has a strong geographic structure that renders this inference particularly challenging. Simulations on other hosts phylogenies are required to provide

definitive conclusions on these possibilities, as such detection would be informative on the host-switching processes (de Vienne *et al.*, 2013).

Absence of transmitted microbes in *Ariamnes* spiders:

The microbiota of the *Ariamnes* spiders showed a low proportion of core OTUs, which suggests that the bacterial turnover within spider microbiota is quite large. By applying HOME, we showed that these bacterial OTUs did not reject the null hypothesis of host-independent evolution. Given the statistical power of HOME when the number of segregating sites is low, we can conclude that there are likely less than five transmitted core OTUs in this system. The two *Bacillus* and *Erythrobacter* OTUs that rejected the null hypothesis of independent evolutions with HOME have high estimated number of host-switches, which resulted in incongruent cophylogenetic patterns, meaning that they are likely false-positives (Figure I.2.2). Instead, most bacteria are probably acquired in the environment at each generation by spider individuals. Although the ability of HOME to infer transmitted OTUs occurring in only few hosts is limited, the fact that these OTUs are absent in most of the host individuals suggests that these OTUs are facultative symbionts with a low specificity toward spiders, rather than specific vertically transmitted symbionts. The spider microbiota assembly is thus likely not determined by the vertical inheritance of microbial lineages in this system. Such results were partially expected given that spider microbiota can show a very high heterogeneity, and that feeding experiments have recently demonstrated the lability of the microbiota composition according to the spider's diet (Kennedy *et al.*, 2020). This corroborates the fact that the degree of conservatism and the functional relevance of the microbiota are highly variable across the animal kingdom, especially within arthropods in which the microbiota composition ranges from mainly transient microbes acquired from the environment (Hammer *et al.*, 2017, 2019) to striking examples of vertically transmitted microbes (Moran *et al.*, 2008).

One could argue that not detecting (many) transmitted bacterial OTUs in this study comes from the fact that the 16S rRNA marker evolves too slowly to have accumulated any segregating sites in the nucleotide alignment of transmitted OTUs. Therefore, nucleotide alignments without segregating sites could correspond to transmitted OTUs. However, the nucleotide substitution rate of the 16S rRNA gene is estimated to be 1% per 50 million years in bacteria (Ochman *et al.*, 1999). Given that the sum of the branch lengths of the phylogenetic tree of the *Ariamnes* spiders represents a total of 13 million years of nucleotide evolution, we expect on average at least one segregating site per 300 bp alignments of transmitted bacteria. Given that substitution rates of symbiotic bacteria are higher because of their small population size compared to free-living bacteria (Moran *et al.*, 1993, 2008), this would confer even more variability within the OTU alignments of transmitted bacteria. Therefore, it is unlikely that there are transmitted microbes among *Ariamnes* microbiota but that they do not have segregating sites. Using metabarcoding markers that evolve faster would help to confirm the absence of transmitted microbes among of *Ariamnes* spider microbiota.

Other drivers of phylosymbiosis:

Our study highlights an empirical system in which phylosymbiosis is likely explained by processes other than vertical transmission (Kohl, 2020). First, phylosymbiosis can emerge through the existence of a simple ecological-filtering during host colonization, as has been hypothesized before (Moran & Sloan, 2015) and demonstrated using simulations (Mazel *et al.*, 2018). Indeed, if the microbiota is entirely acquired from the environment and if its assembly is influenced by host traits (*e.g.* gut pH, diet…) that are phylogenetically conserved, then the microbiota will be more similar between closely-related than distantly-related species. Such mechanisms could be acting in the microbiota assembly of *Ariamnes* spiders and be responsible for the observed pattern of phylosymbiosis, but experimental works would be required to test whether differences in their microbiota could be linked to phylogenetically conserved host-filtering mechanisms. Second, a heterogeneous geographic structure in the environmental pools of available microbes can impact microbiota assembly and generate phylosymbiotic patterns if phylogenetically related host lineages tend to occupy similar geographic areas (Kohl, 2020). In *Ariamnes* spiders for example, each island is characterized by one dominant endosymbiont (Armstrong *et al.*, 2020): these intracellular bacteria, that generally colonize most spider tissues including the midgut (Sheffer *et al.*, 2019), are typical symbionts of arthropods. Here, these phylogenetically-conserved shifts in the presence/absence of abundant endosymbionts likely generate most of the phylosymbiosis in the *Ariamnes* system (Armstrong *et al.*, 2020), and might explain might why its significance decreased when not considering OTU abundances (Supplementary Figure 1). Interestingly, these bacterial endosymbionts are well-known to be transmitted from generation to generation through direct transfer in the maternal germline (Moran *et al.*, 2008), but our analyses suggest that they are not vertically transmitted over long timescales. Instead, this suggests that the temporal turnover of the endosymbionts is relatively high compared to the timescale of host diversification and that their epidemic spread is influenced by island structure. Altogether, this suggests that macro-organism diversification and microbiota evolution can happen at two decoupled timescales, even in the presence of phylosymbiosis.

Our results in *Ariamnes* spiders do not imply that phylosymbiosis in other host-microbiota systems is not (at least partly) explained by vertical transmission. For instance, bacterial vertical transmission occurs in the gut microbiota of mammals (Sanders *et al.*, 2014; Groussin *et al.*, 2017; Youngblut *et al.*, 2019; Article 1), where it generates stronger phylosymbiosis than host-filtering alone (Mazel *et al.*, 2018). Importantly, phylosymbiosis indicates a degree of host-phylogenetic conservatism in the many processes involved in microbiota assembly during host evolution (Song *et al.*, 2020), but does not by itself inform on the nature of these processes. Model-based approaches such as HOME can provide a more precise characterization of these non-exclusive processes, and more work in this direction is needed to improve our understanding of microbiota assembly and evolution.

References:

Amato KR, G. Sanders J, Song SJ, Nute M, Metcalf JL, Thompson LR, Morton JT, Amir A, J. McKenzie V, Humphrey G, *et al.* 2019. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. The ISME Journal 13: 576–587.

Armstrong EE, Perez-Lamarque B, Bi K, Chen C, Becking LE, Lim JY, Linderoth T, Krehenwinkel H, Gillespie R. 2020. A holobiont view of island biogeography: Unraveling patterns driving the nascent diversification of a Hawaiian spider and its microbial associates. bioRxiv.

Baldo L, Pretus JL, Riera JL, Musilova Z, Bitja Nyom AR, Salzburger W. 2017. Convergence of gut microbiotas in the adaptive radiations of African cichlid fishes. ISME Journal 11: 1975–1987.

Bogdanowicz D, Giaro K. 2013. On a matching distance between rooted phylogenetic trees. International Journal of Applied Mathematics and Computer Science 23: 669–684.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 8: 218–230.

Brooks AW, Kohl KD, Brucker RM, van Opstal EJ, Bordenstein SR. 2016. Phylosymbiosis: relationships and functional effects of microbial communities across host evolutionary history (D Relman, Ed.). PLOS Biology 14: e2000225.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. 2013. Stacks: An analysis tool set for population genomics. Molecular Ecology 22: 3124–3140.

Charleston MA, Perkins SL. 2006. Traversing the tangle: Algorithms and applications for cophylogenetic studies. Journal of Biomedical Informatics 39: 62–71.

Charleston MA, Robertson DL. 2002. Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny (M Sanderson, Ed.). Systematic Biology 51: 528–535.

Conow C, Fielder D, Ovadia Y, Libeskind-Hadas R. 2010. Jane: A new tool for the cophylogeny reconstruction problem. Algorithms for Molecular Biology 5: 1–10.

Funkhouser LJ, Bordenstein SR. 2013. Mom knows best: The universality of maternal microbial transmission. PLoS Biology 11: e1001631.

Gibson J, Shokralla S, Porter TM, King I, Van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences of the United States of America 111: 8007–8012.

Gillespie RG, Benjamin SP, Brewer MS, Rivera MAJ, Roderick GK. 2018. Repeated diversification of ecomorphs in Hawaiian stick spiders. Current Biology 28: 941-947.e3.

Gillespie RG, Rivera MAJ. 2007. Free-living spiders of the genus *Ariamnes* (Araneae, Theridiidae) in Hawaii. Journal of Arachnology 35: 11–37.

Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nature Communications 8: 14319.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, *et al.* 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host and Microbe 17: 603–616.

Hammer TJ, Janzen DH, Hallwachs W, Jaffe SP, Fierer N. 2017. Caterpillars lack a resident gut microbiome. Proceedings of the National Academy of Sciences of the United States of America 114: 9641–9646.

Hammer TJ, Sanders JG, Fierer N. 2019. Not all animals need a microbiome. FEMS Microbiology Letters 366: 69–73.

Kennedy SR, Dawson TE, Gillespie RG. 2018. Stable isotopes of Hawaiian spiders reflect substrate properties along a chronosequence. PeerJ 2018.

Kennedy SR, Tsau S, Gillespie R, Krehenwinkel H. 2020. Are you what you eat? A highly transient and prey-influenced gut microbiome in the grey house spider *Badumna longinqua*. Molecular

Ecology 29: 1001–1015.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120.

Kohl KD. 2020. Ecological and evolutionary mechanisms underlying patterns of phylosymbiosis in host-associated microbial communities. Philosophical Transactions of the Royal Society B: Biological Sciences 375: 20190251.

Lim SJ, Bordenstein SR. 2020. An introduction to phylosymbiosis. Proceedings of the Royal Society B: Biological Sciences 287: 20192900.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2015. Swarmv2: Highly-scalable and high-resolution amplicon clustering. PeerJ 2015: 1–12.

Mazel F, Davis KM, Loudon A, Kwong WK, Groussin M, Parfrey LW. 2018. Is host filtering the main driver of phylosymbiosis across the Tree of Life? (H Bik, Ed.). mSystems 3: 1–15.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey H V., Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, *et al.* 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences 110: 3229–3236.

Minich JJ, Sanders JG, Amir A, Humphrey G, Gilbert JA, Knight R. 2019. Quantifying and understanding well-to-well contamination in microbiome research (N Segata, Ed.). mSystems 4.

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

Moeller AH, Suzuki TA, Phifer-Rixey M, Nachman MW. 2018. Transmission modes of the mammalian gut microbiota. Science 362: 453–457.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annual Review of Genetics 42: 165–190.

Moran NA, Munson MA, Baumann P, Ishikawa H. 1993. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proceedings of the Royal Society of London. Series B: Biological Sciences 253: 167–171.

Moran NA, Sloan DB. 2015. The Hologenome concept: Helpful or hollow? PLoS Biology 13: e1002311.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32: 268–274.

Nyholm S V., McFall-Ngai M. 2004. The winnowing: establishing the squid–vibrio symbiosis. Nature Reviews Microbiology 2: 632–642.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proceedings of the National Academy of Sciences 96: 12638–12643.

Ochman H, Worobey M, Kuo CH, Ndjango JBN, Peeters M, Hahn BH, Hugenholtz P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biology 8: 3–10.

Parfrey LW, Knight R. 2012. Spatial and temporal variability of the human microbiota. Clinical Microbiology and Infection 18: 5–7.

Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host-microbiota evolution: Application to the great apes gut microbiota. Molecular Ecology Resources 19: 1659–1671.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596.

Revell LJ. 2012. phytools: An R-package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016: e2584.

Sanders JG, Powell S, Kronauer DJC, Vasconcelos HL, Frederickson ME, Pierce NE. 2014. Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Molecular

Ecology 23: 1268–1283.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Molecular Biology and Evolution 19: 101–109.

Schloss PD, Gevers D, Westcott SL. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16s rRNA-based studies (JA Gilbert, Ed.). PLoS ONE 6: e27310.

Selosse MA, Baudoin E, Vandenkoornhuyse P. 2004. Symbiotic microorganisms, a key for ecological success and protection of plants. Comptes Rendus - Biologies 327: 639–648.

Sheffer MM, Uhl G, Prost S, Lueders T, Urich T, Bengtsson MM. 2019. Tissue- and population-level microbiome analysis of the wasp spider *Argiope bruennichi* identified a novel dominant bacterial symbiont. Microorganisms 8: 8.

Song SJ, Sanders JG, Delsuc F, Metcalf J, Amato K, Taylor MW, Mazel F, Lutz HL, Winker K, Graves GR, *et al.* 2020. Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. mBio 11: 1–14.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013. Efficient exploration of the space of reconciled gene trees. Systematic Biology 62: 901–912.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21: 2045–2050.

Tkacz A, Cheema J, Chandra G, Grant A, Poole PS. 2015. Stability and succession of the rhizosphere microbiota depends upon plant type and soil composition. ISME Journal 9: 2349–2359.

de Vienne DM, Giraud T, Shykoff JA. 2007. When can host shifts produce congruent host and parasite phylogenies? A simulation approach. Journal of Evolutionary Biology 20: 1428–1438.

de Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: Methods for testing, evidence from natural associations and relation to coevolution. New Phytologist 198: 347–385.

Yeoh YK, Dennis PG, Paungfoo-Lonhienne C, Weber L, Brackin R, Ragan MA, Schmidt S, Hugenholtz P. 2017. Evolutionary conservation of a core root microbiome across plant phyla along a tropical soil chronosequence. Nature Communications 8: 215.

Youngblut ND, Reischer GH, Walters W, Schuster N, Walzer C, Stalder G, Ley RE, Farnleitner AH. 2019. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nature Communications 10: 1–15.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3mNpeln` or by scanning:

Article 3: Comparing different approaches for detecting vertical transmission in host-associated microbiota

Authors: Benoît Perez-Lamarque[1,2] & Hélène Morlon[1]

[1] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

Abstract

Long-term transmissions of gut bacteria are thought to be frequent and functionally important in mammals. Several approaches have been proposed to detect, among species-rich microbiota, the bacteria that have been vertically transmitted during the host clade radiation. Applied to mammal microbiota, these methods have sometimes led to quite conflicting results, and it remains unclear how these different approaches cope with the slow evolution of the genes used to characterize bacterial microbiota, like the 16S rRNA gene. Here, we use simulations to test the statistical performances of two widely-used global-fit approaches (ParaFit and PACo) and two event-based approaches (ALE and HOME). We find that these approaches have different advantages and weaknesses according to the amount of variation in the bacterial DNA sequences and may therefore be complementary. We apply them to the gut microbiota of primates and find that at most 10% of their bacteria are transmitted. We also provide recommendations for future studies looking at vertical transmission among host-associated microbiota.

Keywords: vertical transmission, cophylogeny, microbiota, primates.

Author contributions: BPL and HM designed the study. BPL performed the analyses and wrote the first draft of the manuscript.

Data availability: A tutorial on how to use the different approaches is available on `https://github.com/BPerezLamarque/HOME/`. ParaFit, PACo, and HOME are available as R functions. Conversely, ALE is less user-friendly as it requires the installation of PhyloBayes and the software ALE (`https://github.com/ssolo/ALE/`) and is executable on the terminal.

Citation: Perez-Lamarque B, Morlon H. Comparing different approaches for detecting vertical transmission in host-associated microbiota. ***in preparation***.

Introduction:

Most mammals strongly rely on their associated microbial communities, called microbiota, for various functions like their nutrition, protection, or development (Selosse *et al.*, 2004; McFall-Ngai *et al.*, 2013; Hacquard *et al.*, 2015). To ensure that new mammal hosts are colonized by beneficial microbes, animals have evolved a range of strategies to efficiently transmit their microbes at each generation, including direct transmissions at birth, during parental care, or through social contact (Moran *et al.*, 2019). If these transmissions are stable and faithful, host-microbe interactions are conserved in the host lineage over long-time scales and we refer to this process as vertical transmission (Groussin *et al.*, 2017). At host speciation, we expect transmitted microbes to be inherited by the two daughter host species and to separately evolve in each host lineage, resulting in a pattern of cophylogeny where the phylogeny of the transmitted microbe mirrors that of the host (de Vienne *et al.*, 2013). Conversely, if a microbe is acquired from the environmental pool of microbes at each host generation, we do not expect any cophylogenetic patterns between the microbe and the host.

Proofs of long-term vertical transmissions among the bacterial gut microbiota of mammals are numerous (Sanders *et al.*, 2014; Moeller *et al.*, 2016; Groussin *et al.*, 2017; Gaulke *et al.*, 2018; Youngblut *et al.*, 2019; Article 1). They mainly come from DNA metabarcoding datasets, where the whole bacterial communities are characterized using the 16S rRNA gene, a short and slowly evolving region (Ochman *et al.*, 2010). Usually, one clustered the 16S rRNA sequences into operational taxonomic units (OTUs) based on sequence similarity and inferred which bacterial OTUs present a cophylogenetic pattern with the host, suggesting that they may be vertically transmitted. However, what is the exact proportion of transmitted bacteria remains unclear. For instance, Groussin *et al.* (2017) estimated that more than 50% of the bacterial OTUs have been vertically transmitted in mammals, but Gaulke *et al.* (2018) only found 14% of transmitted bacterial clades. Similarly, we only estimated that $\sim$8% of the gut bacteria were transmitted in the gut microbiota of great apes (Article 1). These discrepancies likely come from the fact that different quantitative approaches have been used and we do not have a clear idea of the advantages and disadvantages of each approach to detect vertical transmission when applied to metabarcoding datasets. Indeed, transmitted bacteria characterized with 16S rRNA metabarcoding only accumulate very few substitutions in their DNA sequences since they co-diverged with their mammal hosts. Therefore, the low amount of information within the 16S rRNA sequences prevents us to robustly reconstruct the bacterial phylogeny. How phylogenetic uncertainty in bacterial evolution affects the results of the different approaches to detect transmitted bacteria remains unknown yet.

Roughly, these approaches can be divided into two categories (de Vienne *et al.*, 2013). First, the global-fit approaches measure a global congruence between the host and bacteria evolutionary histories. For instance, ParaFit (Legendre *et al.*, 2002) and PACo (Bal-

buena *et al.*, 2013) are two widely used approaches based on the fourth-corner statistic or Procrustes superimposition respectively. They do not need to reconstruct the OTU tree as they can be directly applied to the bacterial genetic distances, and they also accept multiple OTU strains per extant host. However, they only provide a measure of cophylogenetic pattern and do not inform on the processes at play.
Second, event-based approaches directly model the events of cospeciation, like host-switches, losses, or duplications, to reconciliate the host and OTU phylogenies, while taking into account the uncertainly in the OTU evolution. For instance, the ALE approach (Szöllősi *et al.*, 2013a) uses a posterior distribution of OTU phylogenetic trees to fit reconciliation events. It simultaneously considers that host-switches likely come from unsampled or extinct host lineages and that the OTU could have been absent in ancestors of all host and only secondarily acquired (Szöllősi *et al.*, 2013b). Recently, the HOME approach (Article 1) has been developed with the goal of inferring transmitted OTUs even when they only have accumulated few substitutions in their 16S rRNA gene sequences (*i.e.* when the host clade has recently diverged). Compared to ALE, HOME is a much more simplistic model, that only considers cospeciations and host-switches events and only allows one OTU strain per extant host. However, HOME does not need a phylogenetic reconstruction of the OTU tree but instead directly models the bacterial DNA substitution process on the host phylogeny (considering eventual host-switches).
Both global-fit and event-based approaches rely on randomizations for generating null expectations under independent host-OTU evolutions (*e.g.* when microbes are acquired from the environment): one can reject or not the null hypothesis of independent evolutions by comparing the estimated scenario to null expectations, and therefore, conclude whether an OTU is vertically transmitted.

Here, we performed simulations of bacterial evolutions on the primate phylogeny to investigate the statistical performances of different approaches to detect vertically transmitted bacteria in metabarcoding datasets. We simulated the evolution of the 16S rRNA gene sequences of vertically transmitted bacteria and independently evolving bacteria and measured the statistical power (the proportion of transmitted bacteria inferred as being transmitted) and the type-I error rate (the proportion of independently evolving bacteria inferred as being transmitted; *i.e.* false-positives) of ParaFit, PACO, ALE, and HOME. Finally, we applied these different methods to the gut bacterial microbiota of primates (Amato *et al.*, 2019), discussed the pros and the cons of each approach, and proposed future developments.

Methods:

Primate phylogeny:

Given that our final goal was to apply approaches to detect vertical transmission in the gut microbiota of primates from Amato *et al.* (2019), we performed all the simulations

on the primate phylogeny. We obtained the primate phylogenetic tree of Dos Reis *et al.* (2018), which corresponds to a nearly complete phylogenetic tree of extant primates (367 species) reconstructed using phylogenomic data and fossil calibrations. The crown age of primates was estimated at ∼74 million years (Myr).

Simulations:

We simulated different scenarios of host-microbiota evolution on the complete primate phylogenetic tree: (i) strict vertical transmissions where each microbial OTU evolves on the host phylogeny, (ii) vertical transmissions with a given number of horizontal host-switches (5, 10, 15, or 20), or (iii) environmental acquisition, where the microbes evolved on an independent phylogeny and are randomly acquired by the extant host species (Figure I.3.1). We first considered that each host species was only associated with a single OTU strain. For each scenario, we simulated DNA evolution of the 16S rRNA gene on the OTU phylogeny: DNA substitutions were assumed to follow a K80 model (Kimura, 1980) with different substitution rates (μ): 1.5 (many substitutions), 1, 0.5, 0.1, and 0.05 (very few substitutions). These substitution rates are relative rates that were chosen in order to obtain numbers of segregating sites and haplotypes in the simulated OTU alignment that are consistent with the empirical OTU alignments (see Results). For each OTU, we thus obtained a DNA alignment by taking the OTU sequence in each extant host species. These simulations were performed using the function *sim_microbiota* in the R-package HOME (Article 1; R Core Team, 2020). In particular, we considered that host-switches can happen uniformly on the host phylogeny from a donor branch to a receiving branch where it replaces the previous OTU strain. Once simulating the OTU alignments on the complete primate phylogeny, we only retained the OTU sequences present in the 18 primate species sampled in Amato *et al.* (2019), in order to insert our simulations in a context of under-sampling (that is very frequent when studying host-associated microbiota in a given host clade). We referred to these OTU alignments as the simulations with host-switches.

Second, we considered that OTUs can be lost during primate evolution or not detected in the extant host-associated microbiota using metabarcoding technics, such that we only randomly sampled the OTU sequences in 10 extant host species among each OTU alignment. We referred to these OTU alignments as the simulations with host-switches and losses.

Third, we considered that intra-host OTU duplications can happen stochastically on the host phylogeny (Figure I.3.1), such that multiple OTU strains can persist in a host lineage. We simulated duplications using a continuous-time Markov process, *i.e.* duplications can happen any time on the host branches, with a rate $\kappa = 2$. Similarly, we simultaneously simulated host-switches with the same scenarios as above. We obtained OTU alignments by selecting the OTU haplotypes present in each of the 18 primate species, *i.e.* we only retained the duplicated OTU strains that have experienced at least one sub-

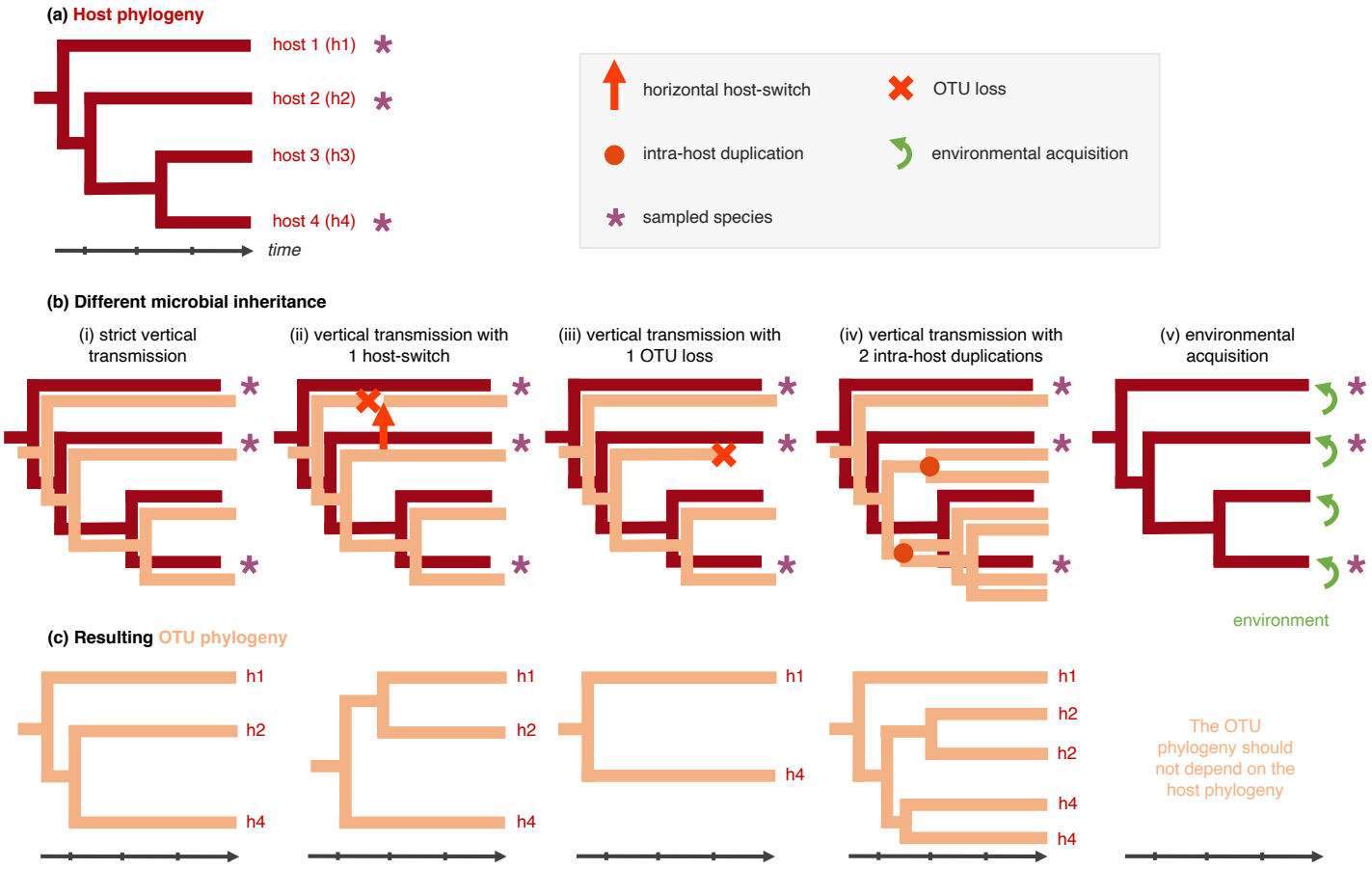


Figure I.3.1: Different modes of inheritance of a given host-associated operational taxonomic unit (OTU) and their consequences on the microbial phylogenies. On a phylogenetic tree of 4 host species (a), we represented the different modes of inheritance for a given OTU (b) and the resulting phylogeny (c): extreme scenarios correspond to strict vertical transmission (i; perfect cophylogenetic pattern) or environmental acquisition (v; no cophylogenetic pattern expected). Under vertical transmission, other punctual processes can result in a loss of perfect congruence between the host and the OTU phylogenies: punctual events of horizontal transmissions (ii; the horizontal transfer from one donor host to a receiver host with OTU replacement), microbial loss (iii) or intra-host duplication (iv). We also represented a sampling process where only some extant host species are sampled to study their microbiota.

stitution since their duplication. We referred to these OTU alignments as the simulations with host-switches and duplications.

Fourth, we also simulated losses and/or non-detection in the simulations with host-switches and duplications, by randomly keeping only 10 extant host species. We thus obtained simulations with host-switches, losses, and duplications.

For each simulated scenario and combination of parameters, we performed 50 simulations, except for $\mu = 0.05$, where we performed 100 simulations given that many of the resulting OTU alignments contained no segregating sites. We therefore obtained a total of 7,200 simulated OTU alignments.

Inferring vertically transmitted symbionts:

We considered four different approaches for detecting vertical transmission: two global-fit approaches, ParaFit and PACo, and two event-based approaches, ALE and HOME. Other approaches exist for detecting vertical transmission, like the global-fit approaches proposed by Hommola *et al.* (2009) that is a generalization of the Mantel tests, but we chose to only focused on the four approaches that we considered as frequently used (Groussin *et al.*, 2017; Gaulke *et al.*, 2018; Youngblut *et al.*, 2019; Article 1).

ParaFit and PACo were run between the host phylogenetic distances (directly obtained from the primate phylogeny) and the microbial genetic distances, obtained by computing pairwise distances from the DNA sequences (assuming a K80 model of substitution). We amended the functions $parafit$ and $PACo$ from the R-packages ape (Paradis *et al.*, 2004) and paco (Hutchinson *et al.*, 2017) respectively, to handle computations of the tests when the number of OTU haplotypes was low. ParaFit and PACo statistics were both computed using a Cailliez correction for negative eigenvalues. To evaluate the significance of the statistic of each test, we compared its value to a null distribution under a hypothesis of independent host-OTU evolution. To do so, we performed 10,000 randomizations by permuting for each host species its associated OTU haplotype(s); in other words, these randomizations kept the same number of OTU haplotype per host species, but permuted their identity (Legendre *et al.*, 2002; Balbuena *et al.*, 2013). To avoid issues during the randomizations of the host-OTU associations, we only ran ParaFit and PACo for the OTU alignments containing at least 3 haplotypes.

To run ALE, one needs first to generate a posterior distribution of OTU phylogenetic trees using Bayesian phylogenetic inference. Following Groussin *et al.* (2017), we reconstructed phylogenetic trees for each OTU alignment using PhyloBayes (Lartillot & Philippe, 2004) with a GTR model and a discrete gamma distribution with four categories. PhyloBayes was run for 4,000 generations, sampling at every generation after an initial burn-in of 1,000 generations. With the host phylogeny and the distribution of OTU trees as inputs, ALE was run using the ALEml program available at `https:`

//github.com/ssolo/ALE: it estimated by maximum likelihood the rates of host-switch, duplication, and loss, and generated a set of 100 host-OTU reconciliations, which gave the average numbers of cospeciations, host-switches, duplications, and losses. To evaluate the significance of these estimated reconciliations, we shuffled the primate species in the phylogenetic tree and re-ran ALE to obtain a distribution of the numbers of reconciliation events under a null hypothesis of independent host-OTU evolution. We considered that an OTU was vertically transmitted if the estimated number of cospeciations was higher than 95% of the null expectations and if the estimated number of host-switches was lower than 95% of the null expectations (Dorrell *et al.*, 2021). We performed 100 randomizations per OTU, except when simulating intra-host duplications, we only performed 50 randomizations because of the long computation time of ALE.

Finally, HOME was run using the function *HOME_model* in the R-package HOME (Article 1). For each OTU alignment, HOME outputs the maximum-likelihood estimates of the number of host-switches and the substitution rate. Likelihood computations were performed using Monte Carlo simulations with 5,000 trees and the tested numbers of host-switches were picked in a grid from 1 to 35. As for ALE, we assessed the significance of these estimations by performing 100 randomizations shuffling the host-OTU associations. We considered that an OTU was vertically transmitted if both the estimated substitution rate and the observed number of host-switches were lower than 95% of the null expectations. Because HOME does not tolerate multiple OTU strains per host tip at present, when simulations included duplications, we randomly picked one single haplotype per host species.

ALE and HOME were only run for OTU alignments presenting at least one segregating site. Because ALE inferences were too long when the numbers of segregating sites in the OTU alignments were low (see Results), we did not use ALE for OTU alignments simulated with $\mu = 0.05$. Conversely, because HOME inferences were too long when the numbers of segregating sites in the OTU alignments were high, we did not use HOME for OTU alignments simulated with $\mu > 0.5$ and duplications.

We computed the statistical power as the percentage of simulated transmitted OTUs (strictly vertically transmitted or transmitted with host-switches) that were inferred as being transmitted. Conversely, the type-I error rate was the ratio of OTUs simulated as independently evolving that were inferred as being transmitted. Note that we did not include in these computations the simulated OTUs for which we could not apply the approaches (*i.e.* OTUs with no segregating site for event-based approaches or OTUs with less than 3 haplotypes for global-fit ones).

Empirical application:

We downloaded the dataset from Amato *et al.* (2019) characterizing the gut bacterial microbiota of 153 primates belonging to 18 species using the V4 region of the 16S rRNA

gene available in https://www.ebi.ac.uk/ena/data/view/PRJEB22679. The demultiplexed Illumina reads were processed using a pipeline based on VSEARCH (Rognes *et al.*, 2016) available in https://github.com/BPerezLamarque/HOME/. In short, after quality filtering, the reads were clustered into OTUs using either Swarm clustering (Mahé *et al.*, 2015) or classical OTU clustering methods with 95% or 97% similarity thresholds. Chimeras were filtered out de novo and taxonomy was assigned to each OTU using the Silva database (Quast *et al.*, 2013). We only kept non-chimeric bacterial OTUs longer than 150 base pairs and represented by at least 5 reads in at least 2 samples. Finally, we assumed that if an OTU had less than 5 reads in a sample, it was likely a cross-contamination and set its abundance to 0.

We only tested the support for vertical transmission for OTUs being present in at least 10 species. We merged all the primate samples from the same species together and for each OTU and we built the OTU alignment by picking the most abundant sequence assigned to this OTU within each host species. In other words, we considered that there is only one OTU strain per host species. OTU sequences were aligned using MAFFT (Katoh & Standley, 2013). We looked at the number of segregating sites and haplotypes in the resulting OTU alignments and we applied ParaFit, PACo, ALE, and HOME to detect vertically transmitted OTUs.

Next, we relaxed the hypothesis of a single OTU strain per host species, by considering the possibility of intra-host duplications: we picked up to 3 OTU haplotypes per host species by selecting the 3 most abundant ones when available. Given that HOME cannot tolerate multiple OTU sequences per host, we only ran ParaFit, PACo, and ALE.

Finally, we performed model validation. Amato *et al.* (2019) highlighted that a cophylogenetic pattern in primate microbiota could arise because of the geographic split of the primates between the Old World (Africa and Asia) and the New World (Americas). Thus, rather than being due to vertical transmission, cophylogenetic patterns would be explained by the heterogeneous environmental pools of microbes combined with the fact that closely related primate species tend to be present in the same area. Therefore, for the OTUs that presented a significant cophylogenetic pattern according to the different approaches, we randomized the primate-OTU associations within the Old World and New World respectively, and re-run the approaches: if a significant cophylogenetic pattern is still found, it means that we cannot reject the hypothesis of heterogeneous pools of microbes between the Old World and the New World, whereas if no significant cophylogenetic pattern is recovered anymore when randomizing, we can conclude that the cophylogenetic pattern is likely linked to vertical transmissions.

Results:

Simulations with host-switches and losses:

The OTU alignments simulated with only host-switches contained a mean number of segregating sites >20 when the simulated substitution rate (μ) equaled 1.5 and <5 when $\mu = 0.05$ (with many OTU alignments presenting no segregating sites; Supplementary Figure 1). Similarly, the number of haplotypes went from >15 when $\mu = 1.5$ (almost one OTU haplotype for each host species) to <5 when $\mu = 0.05$, meaning that our simulations comprised a large range of variation within the OTU alignments.

Global-fit approaches (ParaFit and PACo) presented a very high statistical power ($\geq$ 98%) when $\mu \geq 0.5$, which decreased at ~70% when $\mu = 0.05$ (Figure I.3.2). However, they displayed a rather elevated type-I error rate when $\mu = 0.05$ (type-I error rate ~10% for both ParaFit and PACo). We also noticed that PACo tends to have a type-I error rate >5% even when μ was high (Figure I.3.2b).

Similarly, ALE had a very high power (>95%) and a low type-I error (<5%) when $\mu \geq 0.5$ (Figure I.3.3a). In terms of the estimated numbers of events, it correctly inferred

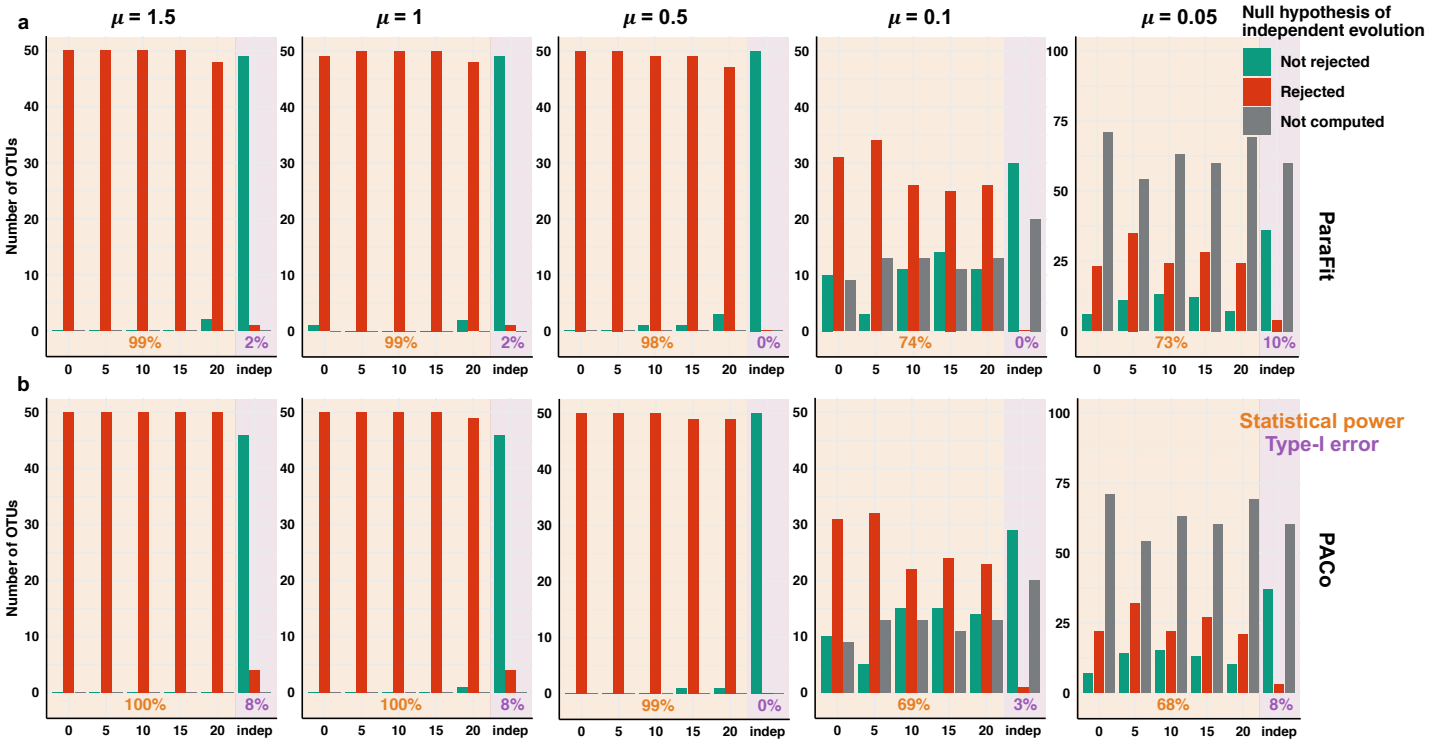


Figure I.3.2: Statistical performances of the global-fit approaches, ParaFit (a) and PACo (b). Numbers of simulated OTUs rejecting the null hypothesis of independent evolution (rejected in red, not rejected in green, and not computed in grey) represented as a function of the simulated scenario: either strict vertical transmission (0 host-switch), vertical transmission with host-switches (5, 10, 15, or 20 switches), or independently evolving ("indep."). Scenarios showing the statistical power of the approach are highlighted in yellow, whereas the ones indicating the type-I error rate are in purple: these performances are indicated as percentages at the bottom of the panels. Each panel corresponds to the different simulated substitution rates (μ).

only cospeciation events when simulating strict vertical transmissions and host-switches when simulated (Figure I.3.3b). Conversely, when independent evolution was simulated, the estimated number of cospeciation events was lower and the number of host switches importantly increased (Figure I.3.3b). Thus, when $\mu \geq 0.5$, ALE had similar statistical performances as global-fit approaches, but also correctly inferred reconciliation events, meaning that ALE is better to reconstruct the evolutionary history of the host-associated microbial OTUs. However, when the number of segregating sites decreased ($\mu = 0.1$), the power of ALE was only <50% and the type-I error increased to 6%. In addition, ALE tends to overfit reconciliation events: it estimated many losses and hosts-switches that

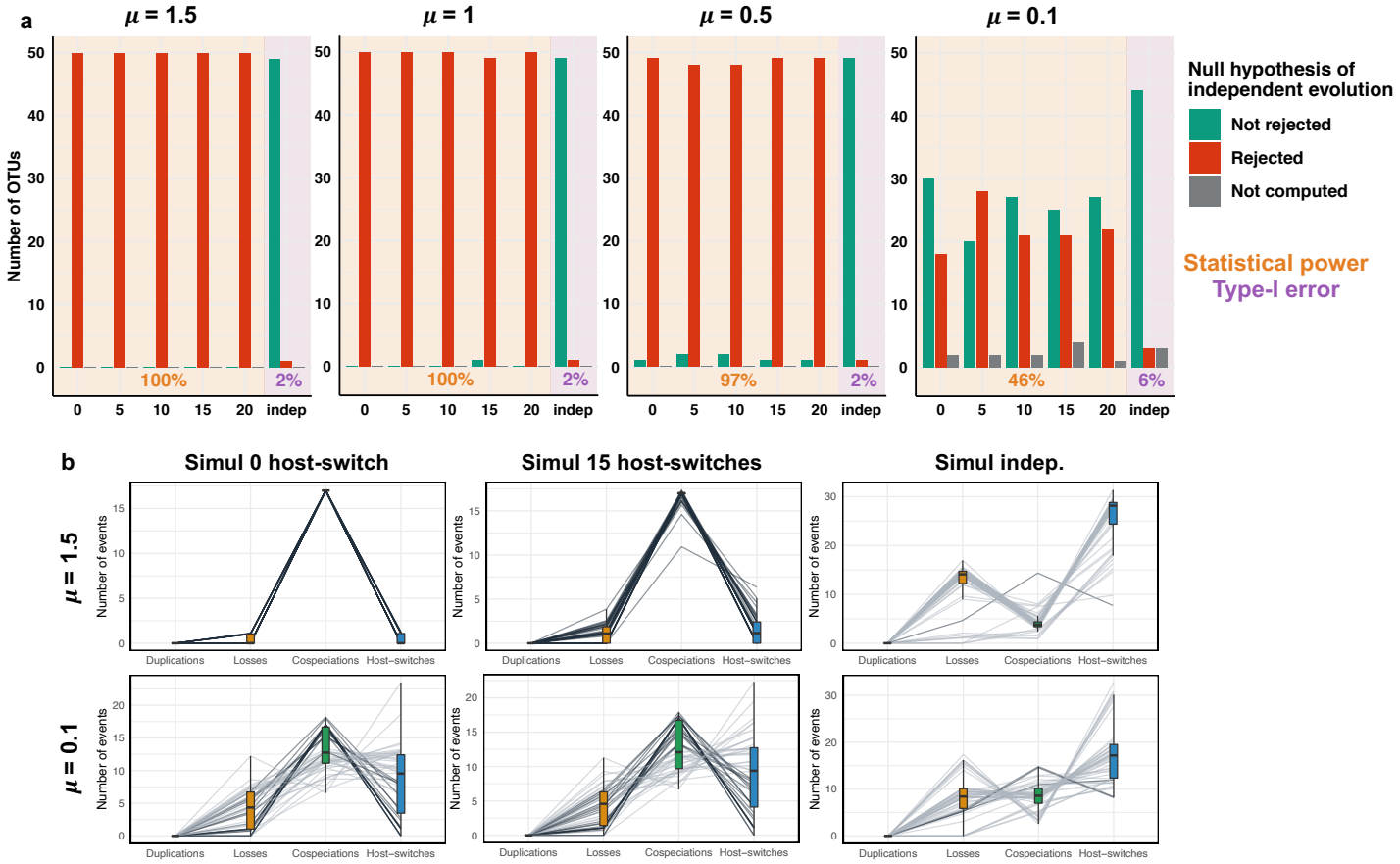


Figure I.3.3: Statistical performances of the event-based approach ALE. (a) Numbers of simulated OTUs rejecting the null hypothesis of independent evolution (rejected in red, not rejected in green, and not computed in grey) represented as a function of the simulated scenario: either strict vertical transmission (0 host-switch), vertical transmission with host-switches (5, 10, 15, or 20 switches), or independently evolving ("indep."). Scenarios showing the statistical power of the approach are highlighted in yellow, whereas the ones indicating the type-I error rate are in purple: these performances are indicated as percentages at the bottom of the panels. Each panel corresponds to the different simulated substitution rates (μ), except $\mu = 0.05$, which was too long to be computed. (b) Estimated parameters (numbers of duplications, losses, cospeciations, or host-switches) as a function of the simulated substitution rates ($\mu = 1.5$ or $\mu = 0.1$) and the simulated scenario: either strict vertical transmission (0 host-switch), vertical transmission with 15 host-switches, or independently evolving ("indep."). Dark grey lines represent OTUs that are inferred to be transmitted, whereas light grey lines represent OTUs acquired from the environment.

were not simulated (Figure I.3.3b). Given that the inferred reconciliation events were untrusty and that the statistical performances of ALE decreased, global-fit approaches are better than ALE when the number of segregating sites is low in the OTU alignments.

Finally, HOME also performed well when μ was high (Figure I.3.4a), but its statistical power tends to importantly decrease with μ: when $\mu = 0.1$, the power of HOME was $\sim$40% and only 23% for $\mu = 0.05$ (which was $\sim$15% lower than global-fit approaches when comparing the absolute number of OTUs inferred as transmitted; Figure I.3.2). However, compared with other approaches, HOME kept a type-I error rate very low ($< 5\%$) in all conditions, especially when μ was low (0% of type-I error; Figure I.3.4a). In terms of infer parameters, HOME correctly estimated the substitution rate as well as the number of host-switches (Figure I.3.4b), although it tends to be noisier when μ was

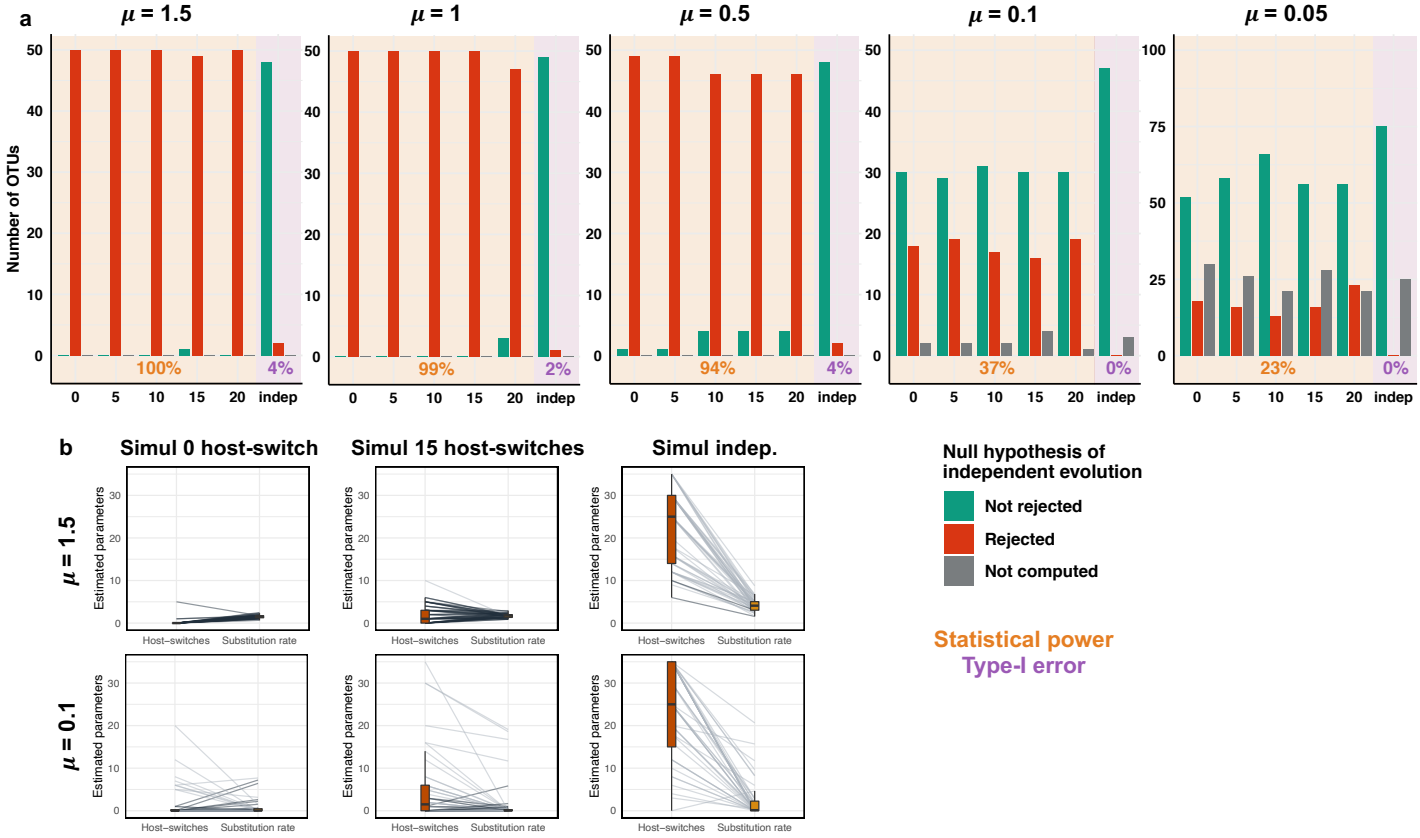


Figure I.3.4: Statistical performances of the event-based approach HOME. (a) Number of simulated OTUs rejecting the null hypothesis of independent evolution (rejected in red, not rejected in green, and not computed in grey) represented as a function of the simulated scenario: either strict vertical transmission (0 host-switch), vertical transmission with host-switches (5, 10, 15, or 20 switches), or independently evolving ("indep."). Scenarios showing the statistical power of the approach are highlighted in yellow, whereas the ones indicating the type-I error rate are in purple: these performances are indicated as percentages at the bottom of the panels. Each panel corresponds to the different simulated substitution rates (μ). (b) Estimated parameters (number of host-switches and substitution rates) as a function of the simulated substitution rates ($\mu = 1.5$ or $\mu = 0.1$) and the simulated scenario: either strict vertical transmission (0 host-switch), vertical transmission with 15 host-switches, or independently evolving ("indep."). Dark grey lines represent OTUs that are inferred to be transmitted, whereas light grey lines represent OTUs acquired from the environment.

low. When independent evolution was simulated, both the estimated substitution rate and the number of host-switches increased (Figure I.3.4b).

In terms of computation time, global-fit approaches were the fastest (especially ParaFit) and their computation time only slightly increased with the simulated substitution rate (*i.e.* more information within the alignment; Supplementary Figure 2). Conversely, both event-based approaches were much slower: Like global-fit approaches, the computation time of HOME increased with μ (from only a few minutes when the number of segregating sites was very low, to several days when there were many of them), whereas the computation time of ALE had an opposite trend. Indeed, when reconciliating the host and OTU trees, ALE considers the phylogenetic uncertainty in the OTU trees. When there were many segregating sites in the OTU alignments, the phylogenetic uncertainty was rather low and ALE was rather fast to run, but when the segregating sites were scarce, the phylogenetic uncertainty was important and therefore, ALE could take several days to run for a single OTU. Thus, to save time and energy, we avoided running ALE when the simulated substitution rates were $\mu < 0.05$ and HOME when $\mu \geq 1$.

When simulating losses (or non-detection within hosts), the statistical power of all the approaches decreased (Supplementary Figures 3-5), especially for HOME ($\sim$10% when $\mu = 0.05$). In addition, the type-I error rate also strongly increased ($> 10\%$) for global-fit and ALE when μ were low, but it remained very low for HOME (0% when $\mu = 0.05$).

Simulations with host-switches, losses, and duplication:

When simulating duplications (and host-switches), we globally increased the number of segregating sites and haplotypes in the simulated OTU alignments (Supplementary Figure 6).
Simulating duplications and allowing multiple strains per host species did not impact the statistical power of global-fit approaches that remained very high (>60% for all μ; Supplementary Figure 7). However, the type-I error rate importantly increased, and it even reached 20% for PACo when $\mu = 1$.
Conversely, ALE handled very well duplications and conserved a high power (>95%) and a low type-I error (5-10%; Supplementary Figure 8). However, the computation time of the approach importantly increased, which complicated the use of the method when the number of segregating sites in the OTU alignment was low.
Finally, HOME, which cannot consider multiple OTU strains per extant host, was not that affected by the sampling at random of a single OTU strain per host: it kept an intermediate power and had still no type-I errors (Supplementary Figure 9).

When also simulating losses (or non-detection within hosts), we observed similar trends with an overall decrease of the power for all the approaches (Supplementary Figures 10-11-12). In addition, we noticed that the type-I error rate of ALE increased >5%.

Empirical application:

A total of 149 95% OTUS, 86 97% OTUs, and 47 Swarm OTUs were tested for vertical transmission. They were on average present in 12 host species and had a number of segregating sites and haplotypes similar to those of the OTUs simulated with substitution rates from $\mu = 0.05$ to $\mu = 0.5$ (Supplementary Figures 1 and 13). The majority of these OTUs were inferred as being vertically transmitted when using global-fit approaches or ALE (Figure I.3.5a). Conversely, according to HOME, only 20% of the tested OTUs were transmitted. More than 60% of the OTUs found as being transmitted were simultaneously found as being transmitted by at least three approaches (Figure I.3.5b): these different approaches tend to agree on a core of transmitted OTUs (Figure I.3.5b), but many OTUs were also inferred as being transmitted with only one approach.

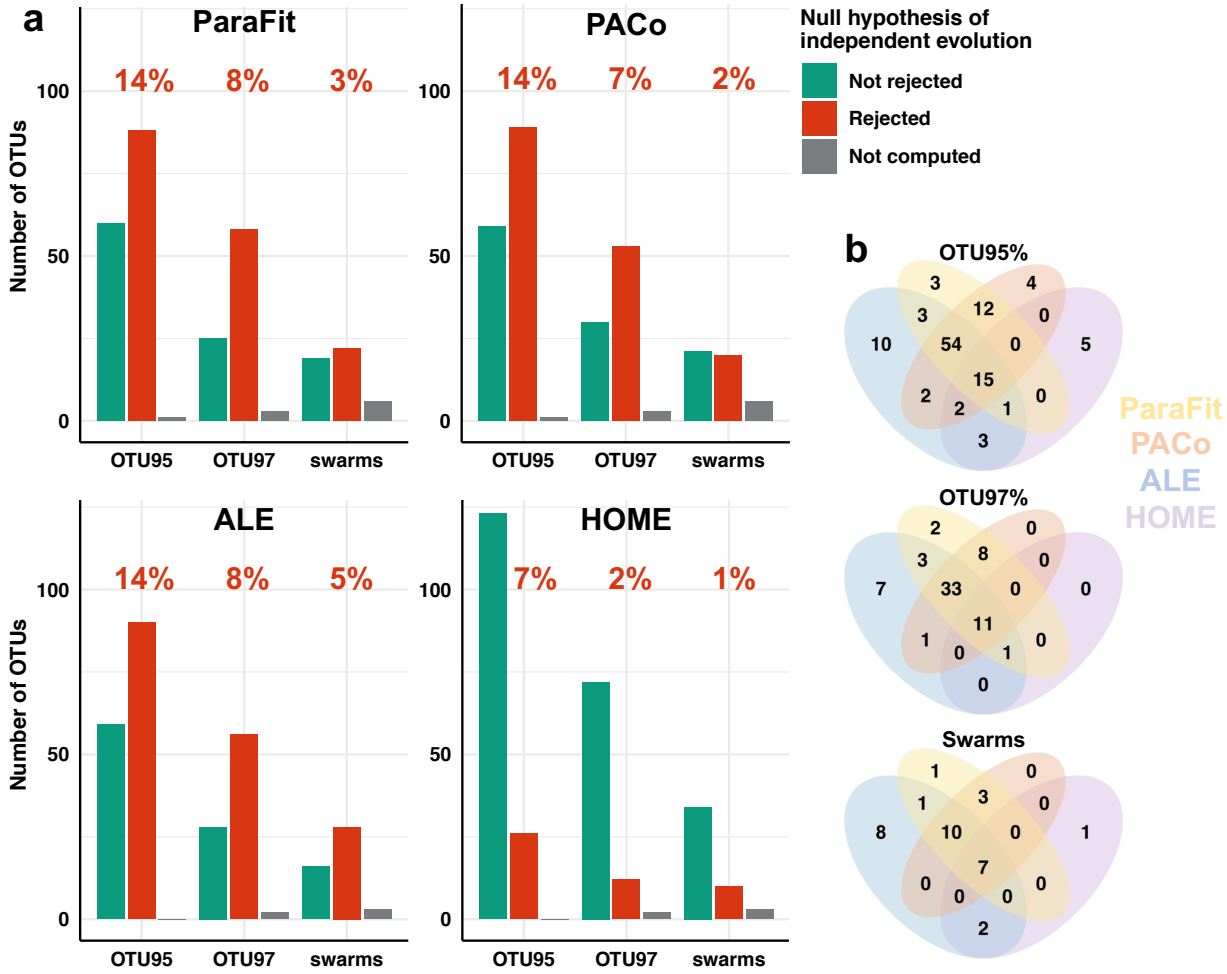


Figure I.3.5: Vertical transmission in primate gut microbiota . (a) Number of OTUs from the gut microbiota of primates rejecting (in red) or not (in green) the null hypothesis of independent evolutions according to the different approaches tested: ParaFit, PACo, ALE, or HOME. In other words, OTUs colored in red represent transmitted OTUs. At the top of each bar, we indicated the percentage of reads corresponding to these transmitted OTUs in the whole primate gut microbiota. OTUs were either clustered as 95%, 97%, or as Swarm OTUs. (b) Venn diagrams indicating the OTUs that are simultaneously found as transmitted based on the different approaches.

In terms of estimated reconciliation events, ALE inferred many host-switches and losses in transmitted OTUs, whereas as expected, non-transmitted ones presented a lot of host-switches compared to cospeciations (Supplementary Figure 14). Similarly, HOME tends to infer lower substitution rates (estimated $\mu < 5$) and lower numbers of host-switches (<10) in transmitted OTUs than in non-transmitted ones (Supplementary Figure 15).

When looking at the numbers of segregating sites and haplotypes of the OTUs being inferred as transmitted according to the different approaches, we observed two opposite trends (Supplementary Figure 13): global-fit approaches and ALE tend to infer more frequently transmitted OTUs when the numbers of segregating sites and haplotypes were low in the alignments, whereas HOME tends to instead infer less transmitted OTUs when OTUs contained less nucleotide variation. Given that (i) the statistical power of all these approaches decreases when there is less nucleotide variation (Figures I.3.2-4) and more losses (Supplementary Figures 3-5), and that (ii) the type-I error rate of global-fit approaches and ALE also increases in these conditions, we suggested that many of the OTUs inferred as being transmitted by global-fit approaches and ALE were likely to be false positives.

Similarly, when selecting several strains per OTUs and applying global-fit approaches and ALE, more than 75% of OTUs are inferred as being vertically transmitted (Supplementary Figure 15). Similarly, given that our simulations suggested that the type-I error rate of these approaches (especially PACo; Supplementary Figures 7-12) increases when considering duplications, we suggested that many of these OTUs might correspond to false positives.

The model validations to assess the role of the split between Old and New Worlds in the cophylogenetic signals showed contrasting results (Figure I.3.6). Indeed, with ParaFit, PACo, and ALE, we still recovered a cophylogenetic signal in most of the OTUs when randomizing the primate-OTU associations within the Old and New Worlds, suggesting that we cannot exclude that the signal in these OTUs came from heterogeneous environmental pools of bacteria rather than vertical transmission. Conversely, most of the OTUs inferred with HOME did not reject the null hypothesis of independent evolution when randomizing based on the geography, suggesting that these OTUs are likely to have been transmitted.

According to the global-fit approaches or ALE, transmitted OTUs correspond to 2% (based on Swarm OTUs) to 14% (based on 95% OTUs) of the total number of reads of the primate gut microbiota (Figure I.3.5). Conversely, based on HOME, they only represent less than 7% of the microbiota. When removing the OTUs that are likely exhibiting a signal of geographic isolation (New *versus* Old Worlds), all the approaches estimated that at most 5% of the reads of the primate gut microbiota were vertically transmitted

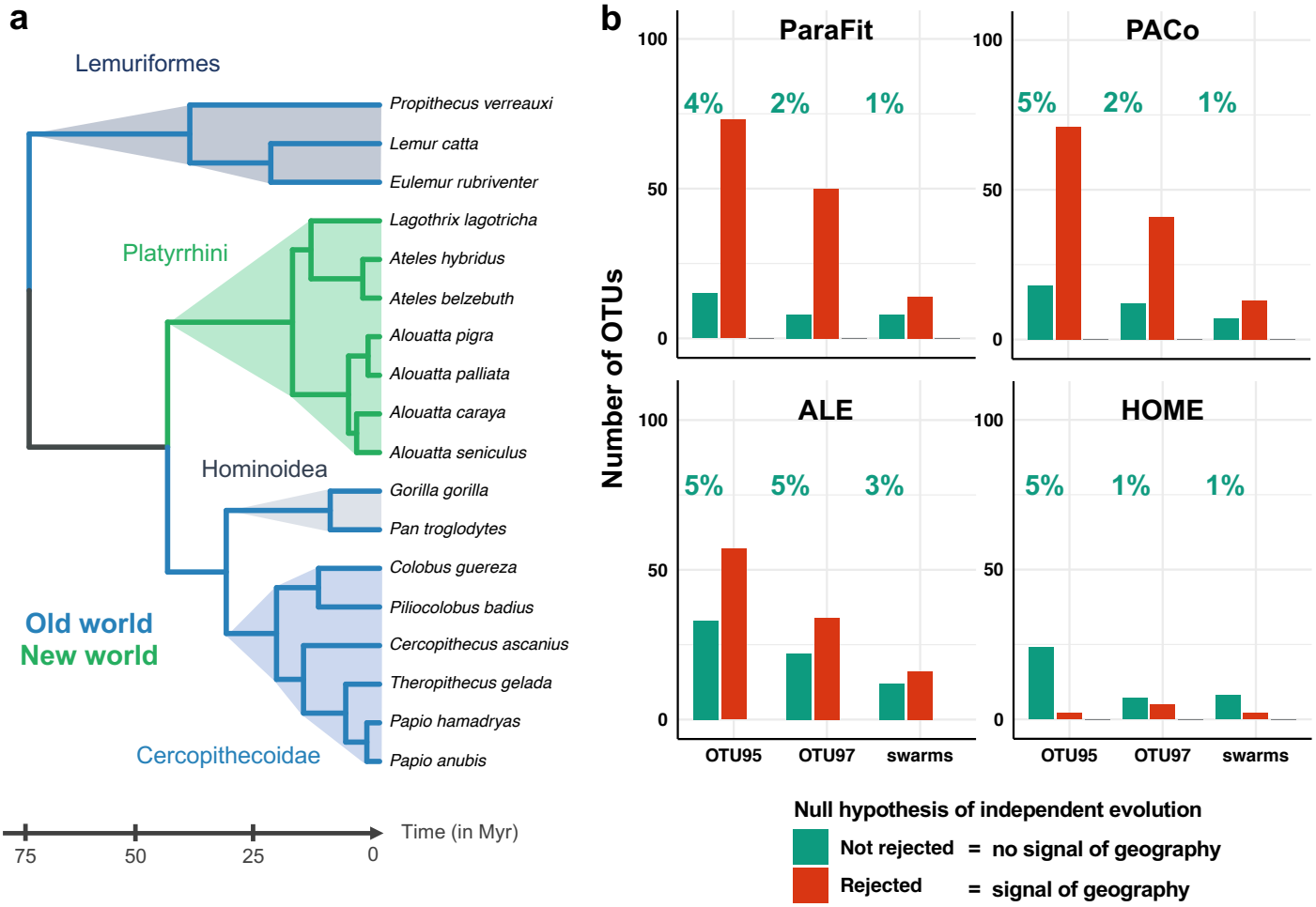


Figure I.3.6: Model validation: Geography can explain a large number of the cophylogenetic patterns. (a) Phylogenetic tree of the 18 primates with branches colored according to the New or Old Worlds. (b) For each OTU being vertically transmitted according to ParaFit, PACo, ALE, or HOME, we re-ran the inference after randomizing the primate-OTU associations within the Old and New Worlds respectively. In other words, if a cophylogenetic pattern is still significant for some OTUs, it means that the original cophylogenetic pattern was mainly driven by a split between the Old and New Worlds and that we cannot exclude that geography alone (*i.e.* heterogeneous environment pools of bacteria) explains alone the patterns of cophylogeny, rather than vertical transmission. Conversely, if the null hypothesis is rejected when randomizations are performed, it means that the cophylogenetic signal is likely generated by vertical transmissions. The panels indicate the number of OTUs from the gut microbiota of primates rejecting (in red) or not (in green) the null hypothesis according to the different approaches tested: ParaFit, PACo, ALE, or HOME. In other words, OTUs colored in red represent transmitted OTUs. At the top of each bar, we indicated the percentage of reads corresponding to these transmitted OTUs in the whole primate gut microbiota. OTUs were either clustered as 95%, 97%, or as Swarm OTUs.

(Figure I.3.6). In terms of taxonomy, the transmitted bacteria mostly belonged to the class Clostridia (phylum Firmicutes), especially the orders Lachnospirales and Oscillospirales, and to the class Bacilli (phylum Firmicutes) to a lesser extent.

Discussion:

In this study, we used simulations to compare the statistical performances of different global-fit and event-based approaches to detect vertical transmission among DNA metabarcoding datasets. We found that the different approaches we tested are rather complementary and we applied them to identify the transmitted bacterial OTUs (operational taxonomic units) in primate guts.

The pros and the cons of the quantitative approaches to detect vertical transmission:

When there is little information in the OTU sequences across the host clade (only a few segregating sites), global fit-methods have high statistical power, but an elevated type-I error rate. Given that PACo tends to always have a higher type-I error than ParaFit and takes more time to run (although it remains faster than event-based approaches), we recommend using ParaFit to detect vertically transmitted OTUs. In terms of event-based approaches, ALE is very slow to run when the number of segregating sites in the OTU alignments was low (<10) because there was too much uncertainty in the reconstructed OTU trees. Given that ALE also tends to overfit reconciliation events in these conditions, we do not recommend using it when the amount of variation in the OTU sequences is too low. Indeed, when there is so little information, ALE might be a too complex model to fit. Finally, HOME performs correctly in terms of type-I errors, but has limited power. In other words, HOME is very unlikely to infer false positives, but likely misses many transmitted OTUs. Thus, when there is little variation within the OTUs, we recommend to use ParaFit (high power, many false positives) in combination with HOME (low power, almost no false positives) to detect vertical transmission.

Conversely, when the OTU sequences have accumulated more divergences (>10 segregating sites), ALE performs better than all the other approaches, as it has high power, a low type-I error rate, and accurately fit reconciliation events (host-switches, duplications, and losses) between the hosts and the OTU phylogenies. To evaluate the significance of the reconciliated scenarios, we recommend separately comparing the numbers of cospeciations and host-switches against the null expectations, rather than looking at the differences between the numbers of cospeciations and host-switches (like in Groussin *et al.*, 2017), as the latter strategy seems to decrease the statistical power of the approach (Figure I.3.3). Therefore, when there is a lot of variation within the OTUs, we recommend to ALE to detect vertical transmission.

When simulating intra-host duplications and considering multiple OTU strains in the extant host species, it did not particularly increase the power of the approaches, but strongly increased the type-I error rates in many cases, especially for PACo. Therefore, if one is not so sure that the multiple OTU strains present in an extant host are biological units (and not resulting from PCR and sequencing errors), we suggest only picking the most abundant one for testing support of vertical transmissions.

Randomizations and model validation:

Global-fit approaches and event-based approaches are based on two different randomization technics: Global-fit approaches randomize which OTU(s) is/are present within each host species, whereas ALE and HOME shuffled the host species names. Thus, in event-based approaches, the structure of the interactions is conserved, and only the phylogenetic relationships of the hosts are randomized, while in global-fit approaches, all the interactions are randomized and only the number of OTU strains per host species is conserved, which is overall less conservative. This may explain why the type-I error rates of global-fit approaches tend be more important than event-based ones and using a more conservative randomization strategy might decrease the tendency of detecting false positives. Alternatively, one can opt for correcting for multiple testing when applying these approaches to empirical data; we did not choose this option here because according to the conditions, the type-I error rate of the approaches can be very low (<1%) such that multiple testing is unlikely to generate a large number of false positives, but would probably strongly decrease the statistical power of the approaches.

All the approaches we tested actually only assess a cophylogenetic pattern between the host and the OTU phylogenies. Such a pattern can be due to multiple processes (de Vienne *et al.*, 2013) and vertical transmission is only one of them. Here, we secondarily investigated the effect of heterogeneous pools of microbes in the host's environments due to geographical isolation (between the New World and Old Worlds) on patterns of cophylogeny. We excluded or not this process by randomizing the OTU-host associations within the main geographic area in the empirical application: if a cophylogenetic pattern was no longer significant when such randomization was performed, we concluded that vertical transmission likely explained the cophylogeny. An alternative way would be to use post-processing of the inferences to test, for instance, whether the host-switches inferred by ALE and HOME tend to be more frequent between host lineages present on the same continents (Article 2). We found that HOME was less sensitive to pick OTUs that present a strong geographical signal than global-fit approaches or ALE. This might be due to the fact that HOME directly models DNA substitutions on the host phylogeny and might thus be less prone to detect a cophylogenetic signal when there is actually only a phylogenetic signal in the environmental pools of the available OTU haplotypes; in other words, if OTU haplotypes differ between the New and Old Worlds, they are likely not particularly well modelled by a substitution process on the host tree. Ideally, we should complexify our simulations of host-OTU independent evolution and perform simulations to test how the different approaches cope with heterogeneous pools of environmental microbes (and phylogenetic signals in the host geographic distributions). Such model validations would help to more robustly link cophylogenetic patterns to the generating processes.

Bacterial transmissions in the primate gut microbiota:

We observed quantitative differences according to the different OTU clustering we performed. In particular, core OTUs (the OTUs present in at least 10 host species) were quite rare when using Swarm clustering, maybe because this clustering method is too stringent and tend to over-split into separated OTUs the vertically transmitted bacteria that have accumulated too many divergences (Article 1). Consequently, we detected in proportion less transmitted bacteria using Swarm clustering. One way to avoid an arbitrary clustering method would be to use phylogenetic based-approaches, like the clade-based taxonomic units (ClaaTU; Gaulke *et al.*, 2018).

Ideally, to assess whether cophylogenetic patterns were generated by vertical transmission, one has also to test whether the divergence times for the hosts and the OTU are matching (de Vienne *et al.*, 2013). Though we could not robustly reconstruct the OTU phylogenetic tree here, we can at least look at the number of segregating sites: the bacterial OTUs from the primate gut presented mostly between 2 and 15 segregating sites across the primate clade. Given that on average the 16S rRNA gene diverges by 1% every 50 million years (Myr) (Ochman *et al.*, 1999), and that the primates are >65 Myr old, the observed number of segregating sites in the OTU alignments are compatible with a process of vertical transmissions when the number of segregating site is low. When it exceeds 10 (especially for 95% OTUs), it might either correspond to conglomerates of several transmitted ones (Article 1) or to fast-evolving bacteria, as it can be the case for some transmitted bacteria with small population sizes (Moran *et al.*, 1993).

When removing the OTUs that are exhibiting a signal of geographic isolation, we estimated that <5% of the reads of the primate gut microbiota were vertically transmitted. Given that the statistical power of our approaches can be low (<50% in some conditions), we may conclude that at most 10% of the bacterial gut microbiota of primates have been vertically transmitted. This estimate is likely more realistic than larger ones (*e.g.* from Groussin *et al.*, 2017) given that mammal gut microbiota can be composed of a large proportion of transient food-derived and/or environment-specific microbes that are unlikely to be faithfully transmitted over more than 50 Myr (Nishida & Ochman, 2019; Amato *et al.*, 2019). Among the transmitted bacteria, we found a large proportion of the order Clostridia (phylum Firmicutes), as previously found in previous analyses (Groussin *et al.*, 2017; Gaulke *et al.*, 2018; Article 1).

Conclusion:

Looking at vertically transmitted OTUs using metabarcoding datasets is challenging because of the low amount of information contained in metabarcoding genes. The different approaches to do so have complementary advantages and weaknesses. When having limited amount of variations in the OTU alignments, we recommend combining HOME, which has very infrequent type-I errors but limited power, with ParaFit, which

has higher power but many type-I errors. The 'right' number of transmitted OTUs is likely between the estimates of the two approaches. We also recommend performing further model validations (*e.g.* randomizing the associations within the main geographic area) to check whether the detected cophylogenetic patterns may have been generated by other processes than vertical transmissions. Applied to the gut microbiota of primates, we confirmed that vertically transmitted bacteria are frequent (up to 10%). Future works should particularly focus on the roles on these bacteria in primate microbiota.

References:

Amato KR, G. Sanders J, Song SJ, Nute M, Metcalf JL, Thompson LR, Morton JT, Amir A, J. McKenzie V, Humphrey G, *et al.* 2019. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. The ISME Journal 13: 576–587.

Balbuena JA, Míguez-Lozano R, Blasco-Costa I. 2013. PACo: A novel procrustes application to cophylogenetic analysis (CS Moreau, Ed.). PLoS ONE 8: e61048.

Dorrell RG, Villain A, Perez-Lamarque B, Audren de Kerdrel G, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR, *et al.* 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. Proceedings of the National Academy of Sciences 118: e2009974118.

Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. 2018. Ecophylogenetics clarifies the evolutionary association between mammals and their gut microbiota (A Martiny and DA Relman, Eds.). mBio 9: 1–14.

Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nature Communications 8: 14319.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, *et al.* 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host and Microbe 17: 603–616.

Hommola K, Smith JE, Qiu Y, Gilks WR. 2009. A permutation test of host-parasite cospeciation. Molecular Biology and Evolution 26: 1457–1468.

Hutchinson MC, Cagua EF, Balbuena JA, Stouffer DB, Poisot T. 2017. paco: implementing Procrustean Approach to Cophylogeny in R. Methods in Ecology and Evolution 8: 932–940.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120.

Lartillot N, Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Molecular Biology and Evolution 21: 1095–1109.

Legendre P, Desdevises Y, Bazin E. 2002. A statistical test for host-parasite coevolution (RDM Page, Ed.). Systematic Biology 51: 217–234.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2015. Swarmv2: Highly-scalable and high-resolution amplicon clustering. PeerJ 2015: 1–12.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey H V., Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, *et al.* 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences 110: 3229–3236.

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

Moran NA, Munson MA, Baumann P, Ishikawa H. 1993. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. Proceedings of the Royal Society of London. Series B: Biological Sciences 253: 167–171.

Moran NA, Ochman H, Hammer TJ. 2019. Evolutionary and ecological consequences of gut microbial communities. Annual Review of Ecology, Evolution, and Systematics 50: 451–475.

Nishida AH, Ochman H. 2019. A great-ape view of the gut microbiome. Nature Reviews Genetics 20: 195–206.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proceedings of the National Academy of Sciences 96: 12638–12643.

Ochman H, Worobey M, Kuo CH, Ndjango JBN, Peeters M, Hahn BH, Hugenholtz P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biology 8: 3–10.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host-microbiota evolution: Application to the great apes gut microbiota. Molecular Ecology Resources 19: 1659–1671.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596.

R Core Team. 2020. R: A language and environment for statistical computing.

Dos Reis M, Gunnell GF, Barba-Montoya J, Wilkins A, Yang Z, Yoder AD. 2018. Using phylogenomic data to explore the effects of relaxed clocks and calibration strategies on divergence time estimation: Primates as a test case (S Ho, Ed.). Systematic Biology 67: 594–615.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016: e2584.

Sanders JG, Powell S, Kronauer DJC, Vasconcelos HL, Frederickson ME, Pierce NE. 2014. Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Molecular Ecology 23: 1268–1283.

Selosse MA, Baudoin E, Vandenkoornhuyse P. 2004. Symbiotic microorganisms, a key for ecological success and protection of plants. Comptes Rendus - Biologies 327: 639–648.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013a. Efficient exploration of the space of reconciled gene trees. Systematic Biology 62: 901–912.

Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013b. Lateral gene transfer from the dead. Systematic Biology 62: 386–397.

de Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: Methods for testing, evidence from natural associations and relation to coevolution. New Phytologist 198: 347–385.

Youngblut ND, Reischer GH, Walters W, Schuster N, Walzer C, Stalder G, Ley RE, Farnleitner AH. 2019. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nature Communications 10: 1–15.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3gc8SRP` or by scanning:

Chapter II.

Measuring the interplay between host-microbiota evolutions:

Host-microbiota interactions are essential for most animal and plant functioning. These interactions resulted from billions of years of evolution and such interplay between host and microbes have likely drastically shaped the evolutionary histories of both clades of hosts and microbes. In this chapter, we examined the interplay between the evolutionary histories of host and host-associated microbial clades (Figure 0.3.13b). We focus on two specific questions: "To what extent does evolutionary history influence which microbial species interact with which host species?" and "How does the evolutionary history of hosts influence the diversification of host-associated microbial clades?".

The first question leads us to compare different existing methods for estimating phylogenetic signals in host-microbiota interaction networks, *i.e.* tools to measure whether closely related species share similar sets of partners (Article 4). Thanks to simulations generated using a recently-developed individual-based model of network evolution (Maliet *et al.*, 2020), we compared the performances of two widely-used approaches: the Mantel tests (Mantel, 1967) and the Phylogenetic Bipartite Linear Model (PBLM; Ives & Godfray, 2006). We found that the PBLM has a high tendency at detecting phylogenetic signals when it should not (false-positives) and should therefore not be used to estimate phylogenetic signals in species interactions. Conversely, Mantel tests performed rather well and we proposed a robust way to investigate clade-specific phylogenetic signals. We provided general guidelines for measuring phylogenetic signals in interaction networks that we applied to a mycorrhizal network from La Réunion island (Martos *et al.*, 2012).

We explored the second question by studying the diversification of the arbuscular mycorrhizal fungi (Glomeromycotina) in the past 500 million years and evaluating how land plants might have affected the diversification of their obligate mycorrhizal symbionts (Article 5). To do so, we used the MaarjAM database (Öpik *et al.*, 2010) that gathered worldwide plant-Glomeromycotina interactions with fungal characterization based on the 18S SSU rRNA gene. We reconstructed the phylogenetic trees of the Glomeromy-

cotina, estimated their global diversity, and applied a range of diversification models. Given that inferring the past diversification of microbial clades can be challenging, we performed a range of model validations to take into account the various sources of uncertainty (in species delineations, phylogenetic reconstructions, and global diversity estimations). We found that overall Glomeromycotina have low diversification rates. After a diversification peak around 150 Myr ago, they experienced an important diversification slowdown toward the present. Such a slowdown could be at least in part related to a shrinking of their mycorrhizal niches, due to the recent acquisition of alternative nutritive strategies in many plant lineages.

Contents of Chapter II

Chapitre II : Mesurer les liens entre les histoires évolutives des hôtes et de leurs microbiotes

Les interactions hôtes-microbiotes sont essentielles pour le fonctionnement de la plupart des animaux et végétaux. Ces interactions résultent de milliards d'années d'évolution et les effets des hôtes sur leurs microbes et *vice versa* ont vraisemblablement façonné leurs histoires évolutives de manière drastique. Dans ce chapitre, nous avons examiné les liens entre les histoires évolutives des hôtes et de leurs microbes associés (Figure 0.3.13b). Nous nous sommes intéressés à deux questions en particulier : « Dans quelle mesure les patrons d'interactions hôtes-microbiotes sont influencés par leurs histoires évolutives ? » et « comment l'histoire évolutive des hôtes influence-t-elle la diversification de leurs microbes associés ? »

La première question nous a amené à comparer différentes méthodes pour estimer le signal phylogénétique dans les réseaux d'interactions hôtes-microbiotes, c'est-à-dire les différents outils à disposition pour mesurer si des espèces proches ont tendance à partager des partenaires similaires (Article 4). Grâce à des simulations générées via un modèle individu-centré d'évolution de réseaux (Maliet *et al.*, 2020), nous avons comparé les performances de deux approches fréquemment utilisées : les tests de Mantel (Mantel, 1967) et le modèle linéaire phylogénétique bipartite (PBLM ; Ives & Godfray, 2006). Nous avons trouvé que le PBLM a une tendance importante à détecter du signal lorsque qu'il n'y en a pas (faux positifs) et ne devrait donc pas être utilisé pour estimer le signal phylogénétique dans les interactions entre espèces. À l'inverse, les tests de Mantel fonctionnent plutôt bien, et nous avons proposé une façon robuste d'analyser le signal phylogénétique dans certains sous-clades spécifiques uniquement. Nous avons de plus énoncé des recommandations générales pour mesurer le signal phylogénétique dans des réseaux d'interactions, que nous avons appliquées à un réseau mycorhizien de l'île de la Réunion (Martos *et al.*, 2012).

Nous avons exploré la seconde question en étudiant la diversification des champignons endomycorhiziens à arbuscules (Glomeromycotina) au cours des 500 derniers millions d'années et évalué comment les plantes terrestres auraient influencé la diversification de leur symbiontes mycorhiziens obligatoires (Article 5). Pour se faire, nous avons utilisé la base de données MaarjAM (Öpik *et al.*, 2010) qui rassemble des interactions plantes-Glomeromycotina à l'échelle mondiale et caractérise les champignons via leur gène de l'ARN ribosomal 18S. Nous avons reconstruit l'arbre phylogénétique des Glomeromycotina, estimé leur diversité globale et appliqué une série de modèles de diversification. Sachant qu'inférer la diversification d'un clade microbien à l'aide d'un seul gène peut être difficile, nous avons réalisé une série de validations de modèles afin de prendre en compte les diverses sources d'incertitudes (dans la délimitation d'espèces, les reconstructions phylogénétiques et l'estimation de la diversité globale) ainsi que des simulations. Nous avons trouvé que les Glomeromycotina ont en moyenne des taux de diversification particulièrement bas. Après avoir connu un pic dans leur diversification il y a environ 150 millions d'années, ils ont récemment subi un fort ralentissement. Un tel ralentissement de leur diversification pourrait être en partie lié à la réduction globale de leurs niches mycorhizennes, due aux récentes acquisitions d'autres alternatives nutritives par de nombreuses lignées de plantes.

Article 4: Do closely related species interact with similar partners? Testing for phylogenetic signal in ecological networks

Authors: Benoît Perez-Lamarque[1,2], Odile Maliet[1], Marc-André Selosse[2,3], Florent Martos[2], Hélène Morlon[1]

[1] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

[3] Department of Plant Taxonomy and Nature Conservation, University of Gdansk, Wita Stwosza 59, 80-308 Gdansk, Poland

Abstract

Whether interactions between species are conserved on evolutionary time scales is a central question in ecology and evolution. This question has spurred the development of both correlative and model-based approaches for testing phylogenetic signal in interspecific interactions: do closely related species interact with similar sets of partners? Here we test the performances of some of these approaches using simulations. We find that one of the most widely used model-based approach often detects phylogenetic signal when it should not. Conversely, simple Mantel tests investigating the correlation between phylogenetic distances and dissimilarities in sets of interacting partners have low type-I error rates and satisfactory statistical power, especially when using weighted interactions and phylogenetic dissimilarity metrics; however, they often artifactually detect anti-phylogenetic signals. Partial Mantel tests, which are used to partial out the phylogenetic signal linked to similarity in numbers of partners, actually fail at correcting for the confounding effect of the numbers of partners. We instead propose the use of sequential Mantel tests. We also explore the ability of simple Mantel tests to analyze clade-specific phylogenetic signal. We provide general guidelines and an illustration on an orchid-fungus mycorrhizal network.

Keywords: ecological networks, phylogenetic constraint, species interactions, specialization, mycorrhiza.

Author contributions: BPL, FM, MAS, and HM designed the study. BPL performed the analyses and FM gathered the data. BPL and HM wrote the first draft of the manuscript,

and all authors contributed to revisions.

Acknowledgments: The authors acknowledge M. Elias and D. de Vienne for helpful discussions. They also thank I. Overcast, S. Lambert, I. Quintero, and A. Silva for comments on an early version of the manuscript. This work was supported by a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL and the École Doctorale FIRE – Programme Bettencourt. Funding of the research of FM was from the Agence Nationale de la Recherche (ANR-19-CE02-0002). HM acknowledges support from the European Research Council (grant CoG-PANDA).

Data availability: All the R functions used to measure phylogenetic signals in bipartite interaction networks, including (simple, partial, and clade-specific) Mantel tests and PBLM, are available in the R-package RPANDA (Morlon *et al.*, 2016; functions *phylosignal_network* and *phylosignal_sub_network*).
A tutorial can be also be found in `https://github.com/BPerezLamarque/Phylosignal_network`. Amended functions of BipartiteEvol are also included in RPANDA.

Citation: Perez-Lamarque B, Selosse MA, Martos F, Morlon H. Do closely related species interact with similar partners? Testing for phylogenetic signal in ecological networks. ***in preparation***.

Introduction:

Species in ecological communities engage in diverse types of interspecific interactions, such as pollination, mycorrhizal symbiosis, predatory, or parasitism (Bascompte *et al.*, 2003; Fontaine *et al.*, 2011; Bascompte & Jordano, 2013; Chagnon, 2016). Understanding the processes that shape interaction networks, including the role of evolutionary history, is a major focus of ecology and evolution (Rezende *et al.*, 2007; Vázquez *et al.*, 2009; Thébault & Fontaine, 2010; Krasnov *et al.*, 2012; Elias *et al.*, 2013; Rohr & Bascompte, 2014; Fontaine & Thébault, 2015). One way to assess the role of evolutionary history in shaping contemporary interactions is to test for phylogenetic signal in species interactions, *i.e.* whether closely related species interact with similar sets of partners (Peralta, 2016).

Testing for phylogenetic signal in a unidimensional trait (*i.e.* whether a trait is phylogenetically conserved) for a given species group is mainstream (Felsenstein, 1985; Blomberg *et al.*, 2003; Münkemüller *et al.*, 2012). One approach (the 'correlative' approach) is to perform a Mantel test between species phylogenetic and trait distances (Mantel, 1967); another approach (the 'model-based' approach) relies on trait evolution models such as Pagel's λ (Pagel, 1999) or Blomberg's κ (Blomberg *et al.*, 2003). The model-based approach has a higher ability to detect an existing phylogenetic signal (power) and a lower propensity to infer a phylogenetic signal when it should not (type-I error) (Harmon &

Glor, 2010). The correlative approach should therefore only be used when the model-based approach is not applicable, for example if the 'trait' data is expressed in terms of pairwise distances (Harmon & Glor, 2010).

Testing for phylogenetic signal in species interactions falls in the category of cases where the 'trait' data is expressed in terms of pairwise distances, here the between-species dissimilarity in interaction partners. Simple Mantel tests have therefore been widely used in this context (*e.g.* Rezende *et al.*, 2007; Elias *et al.*, 2013; Fontaine & Thébault, 2015). Partial Mantel tests have also been used to test whether the phylogenetic signal is really in the identity of the interacting partners and not in the degree of generalism, as similarity in the number of partners can increase the value of similarity metrics (Rezende *et al.*, 2007; Jacquemyn *et al.*, 2011; Aizen *et al.*, 2016). Conversely, another very widely used approach is the Phylogenetic Bipartite Linear Model (PBLM) to investigate the phylogenetic signal in interaction networks (Ives & Godfray, 2006). This approach has been used to test for phylogenetic signal in species interactions in a variety of empirical networks, including host-parasite, plant-fungus, and pollination networks (Martos *et al.*, 2012; Martín González *et al.*, 2015; Xing *et al.*, 2020).

Here, we consider weighted and unweighted bipartite interaction networks represented by a matrix of interaction between species from two guilds A and B (Figure II.4.1). We aim to evaluate the statistical performances of the correlative (simple or partial) Mantel tests and of the model-based PBLM (Box 1) and use simulations to perform a comparative analysis of the different approaches. Our results lead us to propose an alternative approach for measuring phylogenetic signal in the identity of the interacting partners. We also investigate the ability of Mantel tests to detect the presence of clade-specific phylogenetic signal. Finally, we provide general guidelines and apply them to a mycorrhizal network between orchids and their fungal partners from La Réunion island.

Box 1: Methods for measuring phylogenetic signal in species interactions

Mantel tests (Mantel, 1967) were introduced to examine the correlation between two dissimilarity matrices. They have been used to measure the phylogenetic signal in species interactions by computing the correlation between the matrix of phylogenetic distances (for species pairs from guild A for example, Figure II.4.1) and the matrix of 'ecological dissimilarities' between the sets of interacting partners (*i.e.* species from guild B interacting with species pairs from guild A). The correlation (R) with -1<R<1 is often evaluated using Pearson correlation, that is the mean of the products of the corresponding elements in the two standardized dissimilarity matrices. Alternatively, R can be evaluated using Spearman correlation (computed from dissimilarities transformed into ranks), or Kendall correlation (computed by counting the pairs of observations that have the same rank). The parametric Pearson correlation is statistically more powerful, but makes stronger hypotheses (it assumes a linear relationship) than the non-parametric Spearman and Kendall correlations (which assume only a monotonic relationship). A positive (resp. negative) correlation indicates a phylogenetic (resp. anti-phylogenetic) signal in species interactions. Its significance is evaluated using randomizations by repeatedly permuting one of the original dissimilarity matrices: one-tailed p-values are obtained by comparing the rank of the original correlation R to the randomized correlations. Ecological dissimilarities of interacting partners between two species from guild A can be measured with various indices. Two classical indices are the Jaccard distance, defined as the number of their unshared partners from guild B divided by their total number of partners, and the UniFrac distance, which incorporates phylogenetic relatedness between partners, computed as the fraction of unshared branch length in the phylogenetic tree of their partners from guild B. Both indices also have a weighted version that accounts for interaction strength. Partial Mantel tests examine the correlation between two dissimilarities matrices while accounting for a third dissimilarity matrix (Smouse *et al.*, 1986). When testing for phylogenetic signal in bipartite interaction networks, these tests are useful for controlling the phylogenetic signal in degree of generalism; indeed similarity in the number of partners can decrease the value of ecological dissimilarity metrics independently of the identity of the partners, such that a phylogenetic signal in the degree of generalism can generate a phylogenetic signal in species interactions that is not linked to an evolutionary conservatism of interacting partners. Partial Mantel tests therefore investigate the correlation between phylogenetic and ecological dissimilarities while controlling for the absolute difference in degrees between species pairs (Rezende *et al.*, 2007).

Box 1 (end)

The Phylogenetic bipartite linear model (PBLM; Ives & Godfray 2006) assumes that interaction strengths between species from guilds A and B are determined by (unobserved) traits that evolve on the two phylogenies each following a simplified Ornstein-Uhlenbeck process parametrized by d_A and d_B (Blomberg *et al.*, 2003). The strength of interaction between two species is assumed to be given by the product of their two traits. Under these assumptions, d_A and d_B can be estimated from the two phylogenies and the matrix of interaction strengths using generalized least squares (Ives & Godfray 2006). d_A and d_B are then interpreted as a measure of phylogenetic signal in species interactions. If $d = 1$, the traits evolved as Brownian motions, if d=0, there is no effect of the phylogenies (similar than evolving on star phylogenies), whereas $0 < d < 1$ would represent stabilizing selection and $d > 1$ disruptive selection. Ives & Godfray (2006) proposed two approaches to assess the significance of the signal. The simplest consists in comparing the mean square errors (MSE) of the generalized least squares regression to the same MSE obtained using star phylogenies (MSEstar). MSE<MSEstar is interpreted as a significant phylogenetic signal in species interactions. The second approach uses a bootstraping strategy to build 95% confidence intervals around the estimated d_A and d_B values: the null hypothesis (absence of phylogenetic signal in guild A, resp. B) is rejected if the confidence interval around d_A (resp. d_B) does not include 0. While designed primarily for applications to bipartite networks characterized by matrices of interaction strengths (*e.g.* net attack rate of a parasitoid on its hosts), PBLM has been applied to weighted networks characterized by matrices of interaction abundance (*i.e.* the number of times the interaction has been observed) and unweighted (binary) networks, using 1 for the interaction strength when species interact and 0 otherwise (Ives & Godfray, 2006; Vázquez *et al.*, 2009; Jacquemyn *et al.*, 2011; Martos *et al.*, 2012; Xing *et al.*, 2020).

Methods:

Simulating interaction networks with or without phylogenetic signal in species interactions:

We used a recently-developed model, BipartiteEvol, to generate interaction networks with or without phylogenetic signal (Maliet *et al.*, 2020). This approach, available in the R-package RPANDA (Morlon *et al.*, 2016; R Core Team, 2020) is an individual-based eco-evolutionary model of two guilds interacting in a mutualistic, antagonistic, or neutral way. Each individual from guild A (resp. B) is characterized by a multidimensional continuous trait and interacts with one individual from guild B (resp. A). The effect of this interaction on the fitness of each individual from guilds A or B is determined by

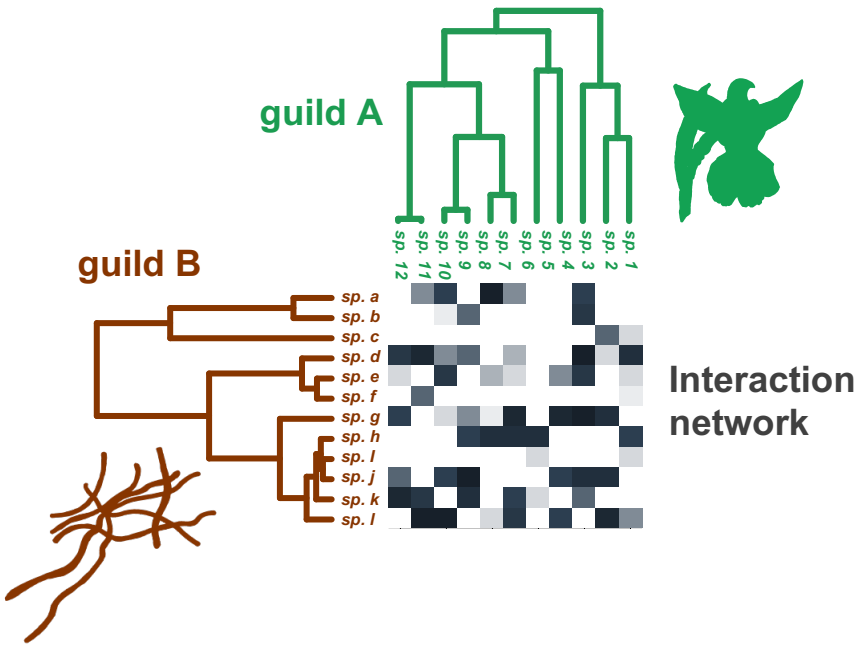


Figure II.4.1: Illustration of the data used to test for phylogenetic signal in species interactions. Toy example of an interaction network between orchids (in green) and mycorrhizal fungi (in brown) with associated phylogenetic trees. The bipartite interaction network between two guilds A (here the orchids) and B (the fungi) is represented by a matrix which elements indicate either whether or not species interact (*i.e.* 1 if they do and 0 otherwise, 'unweighted' or 'binary' network) or the frequency of the interaction ('weighted' network; for example here we indicated the number of times a given pairwise interaction has been observed using shades of grey from white (no interaction) to dark grey (many interactions)). Each guild is also characterized by a rooted phylogenetic tree, used to compute phylogenetic distances between pairs of species.

the distance in trait space of the two interacting individuals, according to a classical trait matching expression parametrized by two parameters α_A and α_B (Supplementary Methods 1, Maliet *et al.* 2020). These parameters determine the nature and specificity of the interaction: positive α_A and α_B correspond to mutualistic interactions, negative α_A and positive α_B to antagonistic interactions (with guild A representing hosts or preys and guild B parasites or predators), high $|\alpha|$ values to scenarios with strong fitness effects (*i.e.* highly specialized interactions), and $|\alpha|$ values close to 0 to more neutral scenarios. At each time step, one individual from guild A is killed at random and replaced by another individual from guild A proportionally to its fitness given its interaction with the individual from guild B. The new individual has a probability μ to mutate, in which case its new trait is drawn independently in each dimension in a normal distribution centered on the parent trait (and a variance of 1). This death/birth/mutation process is replicated for guild B, and these two processes are repeated for a large number of time steps. Here, we amended the species delineation of the original BipartiteEvol model (Maliet *et al.*, 2020) and considered that each combination of traits forms a new species. Under this process, closely related species tend to interact with similar sets of partners (*i.e.* there is a phylogenetic signal in species interactions) if (and only if): (i) closely related species have similar traits (*i.e.* there is a phylogenetic signal in species traits) and (ii) these traits

determine who interacts with whom, *i.e.* $\alpha \neq 0$. Similarly, an anti-phylogenetic signal in species interactions (*i.e.* the tendency for closely related species to associate with dissimilar partners) is expected if there is anti-phylogenetic signal in species traits and $\alpha \neq 0$.

We simulated a total of 2,400 interaction networks with individuals characterized by a six-dimensional trait. In order to obtain networks covering a wide range of sizes, we considered a total number of 500, 1,000, 2,000, 3,000, 4,000, or 5,000 pairs of interacting individuals per simulated network. For each total number of interactions, we simulated the evolution of 100 neutral networks ($\alpha_A = 0$; $\alpha_B = 0$), 120 mutualistic networks ((i) $\alpha_A = 1$; $\alpha_B = 1$; (ii) $\alpha_A = 0.1$; $\alpha_B = 0.1$; (iii) $\alpha_A = 0.01$; $\alpha_B = 0.01$; (iv) $\alpha_A = 1$; $\alpha_B = 0.1$; (v) $\alpha_A = 1$; $\alpha_B = 0.01$; and (vi) $\alpha_A = 0.1$; $\alpha_B = 0.01$) and 180 antagonistic networks ((i) $\alpha_A = -1$; $\alpha_B = 1$; (ii) $\alpha_A = -0.1$; $\alpha_B = 0.1$; (iii) $\alpha_A = -0.01$; $\alpha_B = 0.01$; (iv) $\alpha_A = -1$; $\alpha_B = 0.1$; (v) $\alpha_A = -1$; $\alpha_B = 0.01$; (vi) $\alpha_A = -0.1$; $\alpha_B = 1$; (vii) $\alpha_A = -0.1$; $\alpha_B = 0.01$; (viii) $\alpha_A = -0.01$; $\alpha_B = 1$; (ix) $\alpha_A = -0.01$; $\alpha_B = 0.1$). We used a mutation rate $\mu = 0.01$ and followed the evolution of the interacting individuals during 50^6 death events. At the end, we extracted for each guild a species tree from its genealogy by randomly selecting one individual per species (Supplementary Figure 1) and reconstructed the corresponding weighted interaction network by counting the number of occurrences of each interspecific interaction.

We separated the $2,400$ simulated networks between those for which we should expect a phylogenetic signal in species interactions and those for which we should not. We did not expect phylogenetic signal in species interactions in neutral networks and in non-neutral networks with no phylogenetic signal in species traits. Conversely, we expected phylogenetic signal in non-neutral networks with phylogenetic signal in species traits. For simplicity and consistency with the rest of the paper, we tested for phylogenetic signal in species traits using Mantel tests (Pearson correlation) between phylogenetic distance matrices and trait distance matrices computed as the Euclidian distances between trait values for each species pair.

Computing phylogenetic signals in species interactions:

Mantel tests: We evaluated the phylogenetic signal in species interactions in guilds A and B separately using simple Mantel tests between phylogenetic and ecological distances. Ecological distances were measured both without accounting for evolutionary relatedness of the interacting partners, using (weighted or unweighted) Jaccard, and accounting for relatedness using (weighted or unweighted) UniFrac distances (Box 1; Chen *et al.*, 2012). We used the Pearson, Spearman, and Kendall correlations by extending the existing *mantel* function in the R-package ecodist (Goslee & Urban, 2007); we evaluated the significance of the correlation using $10,000$ permutations, except for the computationally intensive Kendall correlation for which we used 100 permutations. For each network, we defined the "upper p-value" as the percentage of randomized correlations above the original value ($p<0.05$ is interpreted as a significant phylogenetic signal),

and the "lower p-value" as the percentage of randomized correlations below the original value (p<0.05 is interpreted as a significant anti-phylogenetic signal).

PBLM: To estimate phylogenetic signal based on PBLM, we modified the function *pblm* from the R-package picante (Kembel *et al.*, 2010) to more efficiently perform matrix inversions and handle large interaction networks. We followed Ives & Godfray (2006) (Box 1) by considering that the phylogenetic signal is significant when the mean square error (MSE) of the model is smaller than that obtained using star phylogenies (MSEstar), *i.e.* MSE<MSEstar; we also used a more stringent criterion by considering that the signal is significant when the MSE is at least 5% lower than MSEstar, *i.e.* (MSEstar-MSE)/MSEstar>5%. Finally, for the smallest networks (500 pairs of interacting individuals), we applied the bootstrapping method of Ives & Godfray (2006) (Box 1). We did not perform these analyses on larger networks because of their computational cost.

Confounding effect of the phylogenetic signal in degrees of generalism:

To test the performances of the partial Mantel test at measuring phylogenetic signal in species interactions while controlling for signal in degrees of generalism (Box 1), we first performed partial Mantel tests between phylogenetic and ecological distances, while controlling for the absolute differences in degrees, on the networks simulated with BipartiteEvol. There is no reason to produce a phylogenetic signal in degrees of generalism in the BipartiteEvol simulations, and we verified this by performing Mantel tests between phylogenetic distances and degree differences. These analyses were performed to assess whether partial Mantel test loose power compared to simple Mantel tests. If they do not suffer power loss, partial Mantel tests applied to BipartiteEvol simulations should be significant when simple Mantel tests are significant.

Second, we tested whether partial Mantel tests successfully correct for phylogenetic signal in degrees of generalism using networks simulated under a process that generate phylogenetic conservatism in the number, but not the identity, of interacting partners. If partial Mantel tests successfully correct for phylogenetic signal in degrees of generalism, they should not be significant when applied to such networks. We thus simulated networks with only phylogenetic conservatism in the number of interacting partners in guild A: We first simulated phylogenetic trees for guilds A and B using the *pbtree* function (R-package phytools; Revell, 2012) with a number of species uniformly sampled between 40 and 150 species by guild. Second, we simulated the degree of generalism of the species from guild A on the phylogenetic tree using an Ornstein-Uhlenbeck process with an attraction toward 0, a variance of 0.1 (noise of the Brownian motion), and a selection strength (a_A) ranging from 5 (strong stabilizing effect, weak phylogenetic signal) to 0 (Brownian motion, strong phylogenetic signal). We computed the number of partners per species by calibrating the simulated degree values between 1 and the number of species in guild B and taking the integer part. For each a_A value (5, 1, 0.5, 0.05, or 0), we performed 100 simulations using the function *mvSIM* (R-package mvMORPH; Clavel *et al.*,

2015). Third, for each species in A, we attributed the corresponding number of interacting partners in B at random to obtain binary networks. We checked that our simulations indeed generated a signal in degrees of generalism by performing simple Mantel tests between phylogenetic and degree difference distances. Finally, we performed on each simulated network a partial Mantel test between phylogenetic and ecological distances while controlling for the degree difference distances.

Given the poor performances of partial Mantel tests (see Results), we tested whether using sequential Mantel tests would provide a good alternative: based on simple Mantel tests, we consider that there is a phylogenetic signal in the identity of the partners if there is a phylogenetic signal in species interactions and there is no phylogenetic signal in degrees of generalism. We applied this successive testing to all our simulated networks.

Testing the robustness of our results to phylogenetic uncertainty, sampling asymmetry, and network heterogeneity:

The BipartiteEvol simulation framework lacks at reproducing some challenging aspects encountered in the empirical networks, such as the phylogenetic uncertainty, sampling asymmetry, and network heterogeneity. We thus performed additional analyses to investigate the effect of these aspects on the measure of phylogenetic signal.

First, we tested the effect of phylogenetic uncertainty in the partners' tree on the measure of phylogenetic signals when evolutionary relatedness is accounted for (*i.e.* using UniFrac distances). To add some variability in the phylogenetic tree of guild B (resp. A) used to compute the UniFrac distances between species pairs from guild A (resp. B), we first simulated, on the original partners tree, the evolution of a short DNA sequence and then reconstructed the tree from the simulated DNA alignment using neighbor-joining (*nj* function, R-package APE; Paradis *et al.*, 2004). Given that shorter fragments should result in noisier phylogenies, we used the function *simulate_alignment* (R-package HOME; Article 1) to simulate sequences of length (N) 75, 150, 300, 600, or 1,200 base pairs, with 30% of variable sites, and a substitution rate of 1.5. For each of the 2,400 simulations and each N, we obtained a "noisy" tree of guild B (resp. A) for computing the UniFrac distances and the phylogenetic signal in guild A (resp. B), while keeping the original phylogenetic tree of guild A (resp. B).

Second, we tested the influence of sampling asymmetry on measures of phylogenetic signal. Empirical networks are often an incomplete representation of the actual interactions between two guilds because they are under-sampled, and frequently, in an asymmetrical way: for instance, by sampling targeted species from guild A, observed networks are constituted by a few species from guild A which have the complete set of their partners and by often more species from guild B which have an incomplete set of their partners (as they likely interact with unsampled species from guild A). We tested the influence of such sampling asymmetry by selecting only 10% of the most abundant species

from guild A in the simulated network (while retaining at least 10 species) and similarly computed the phylogenetic signals in these asymmetrical subsampled networks.

Third, both Mantel tests and PBLM neglect the heterogeneity within networks. Indeed, a non-significant phylogenetic signal at the level of the entire network can potentially hide a sub-clade presenting significant phylogenetic signal. Alternatively, a phylogenetic signal in the entire network may be driven by only two subclades of guilds A and B. To explore the potential heterogeneity of the phylogenetic signal within one guild, one possibility is to apply Mantel tests to the sub-network formed by a given sub-clade (*e.g.* Song *et al.*, 2020). In order to test this approach, for each node of the tree of guild A having at least 10 descendants, we estimated the clade-specific phylogenetic signal using a Mantel test investigating whether closely related species from this sub-clade of A tend to interact with similar partners (and *vice versa* for guild B). Using UniFrac distances, we performed the Mantel tests with 100,000 permutations, and introduced a Bonferroni correction for multiple testing to keep a global alpha-risk of 5%. To test the power of such an approach for detecting signal in subclades, we generated synthetic networks with known subclade signal by artificially combining networks simulated under neutrality with networks simulated with the set of mutualistic parameters (v) (see Results), such that it creates a separate module. We grafted each "mutualistic" phylogenetic tree from guilds A and B within a "neutral" phylogenetic tree by randomly selecting a branch, such that it creates a separate module with strong phylogenetic signal. Such simulations could correspond to the evolution of a different niche, *e.g.* terrestrial *versus* epiphytic plants associating with different mycorrhizal fungi (Martos *et al.*, 2012). We then performed our clade-specific analysis of phylogenetic signal and investigated whether we recover significant phylogenetic signals at the nodes where mutualism originated.

General guidelines and illustration with application on the orchid-fungus mycorrhizal network from La Réunion:

We used our results and other empirical considerations to provide general guidelines for researchers interested in detecting phylogenetic signal in interaction networks. We illustrated these guidelines by applying them in a network between orchids and mycorrhizal fungi from La Réunion island (Martos *et al.*, 2012). This network encompasses 70 orchid species (either terrestrial or epiphytic species) and 93 associated fungal species (defined according to 97% sequence similarity; see Martos *et al.* (2012) for details). We gathered the maximum-likelihood plant and fungal phylogenies on TreeBASE (Study Accession no. S12721), calibrated the orchid phylogeny using a relaxed clock with the R function *chronos* (Paradis, 2013), and obtained a species-level phylogeny of the orchids by arbitrarily adding 10 million-years-old polytomies in unresolved genera.

Results:

Expected phylogenetic signals in species interactions in BipartiteEvol networks:

The networks simulated using BipartiteEvol gave realistic ranges of sizes for guilds A and B (from less than 50 to more than 250 species; Supplementary Figure 2) and connectance values (*i.e.* ratios of realized interactions, between 5 and 20%; Supplementary Figure 3).

We found a significant phylogenetic signal in species traits for most antagonistic and neutral simulations (Supplementary Figure 4). In contrast, for many mutualistic sets of parameters, closely related species often did not tend to have similar traits, except when $\alpha_B = 0.01$ (*i.e.* mutualistic sets (iii), (v) and (vi); Supplementary Figure 4). Conversely, when α_B were higher (*i.e.* mutualistic sets (i), (ii) and (iv)), we suspect stabilizing selection to occur and erase the phylogenetic signal in the traits (Maliet *et al.*, 2020): as a consequence, we expected no phylogenetic signal in species interactions for these simulations. In addition, we found an anti-phylogenetic signal in species traits in less than 1% of the simulations (Supplementary Figure 4): these networks were removed when evaluating the performance of the different approaches, we therefore do not expect anti-phylogenetic signal in species interactions for the remaining networks.

Computing phylogenetic signals in species interaction:

Using Mantel tests, as expected, we did not find significant phylogenetic signal in species interactions for most neutral networks or for networks with no signal in species traits (Figure II.4.2, Supplementary Figures 5-6-7): type-I error rate was below 5%, corresponding to the alpha-risk of the permutation test (Supplementary Table 1), with one notable exception for small networks when using the weighted Jaccard ecological distances and the Pearson correlation ($\sim$8% type-I error). Conversely, many mutualistic or antagonistic networks where we expected phylogenetic signal in species interactions (*i.e.* non-neutral networks with signal in species traits) presented no significant signals (Figure II.4.2, Supplementary Figures 5-6-7), in particular those simulated with parameters α_A and α_B close to 0 (*e.g.* antagonism (vii)), which tend toward a neutral effect of the traits. In mutualism, phylogenetic signals in species interactions were only present when there was a large asymmetry in the effects of trait matching on the finesses of the species from guilds A or B (case (v) $\alpha_A = 1$; $\alpha_B = 0.01$), *i.e.* when only one guild was specialized. Conversely, in antagonism, phylogenetic signals were mainly found when trait matching had a strong impact on the fitness of guild B (the parasites/predators - $\alpha_B \geq 0.1$), mainly because species from guild B obligately interact with species from guild A, whereas species from guild A have a fitness of 1 when there is a mismatch (Supplementary Methods 1). Additionally, when the phylogenetic signal was significant in one guild, it was generally also significant in the other, although in antagonism, it was usually higher in guild A compared to guild B (Supplementary Figures 5-6-7).

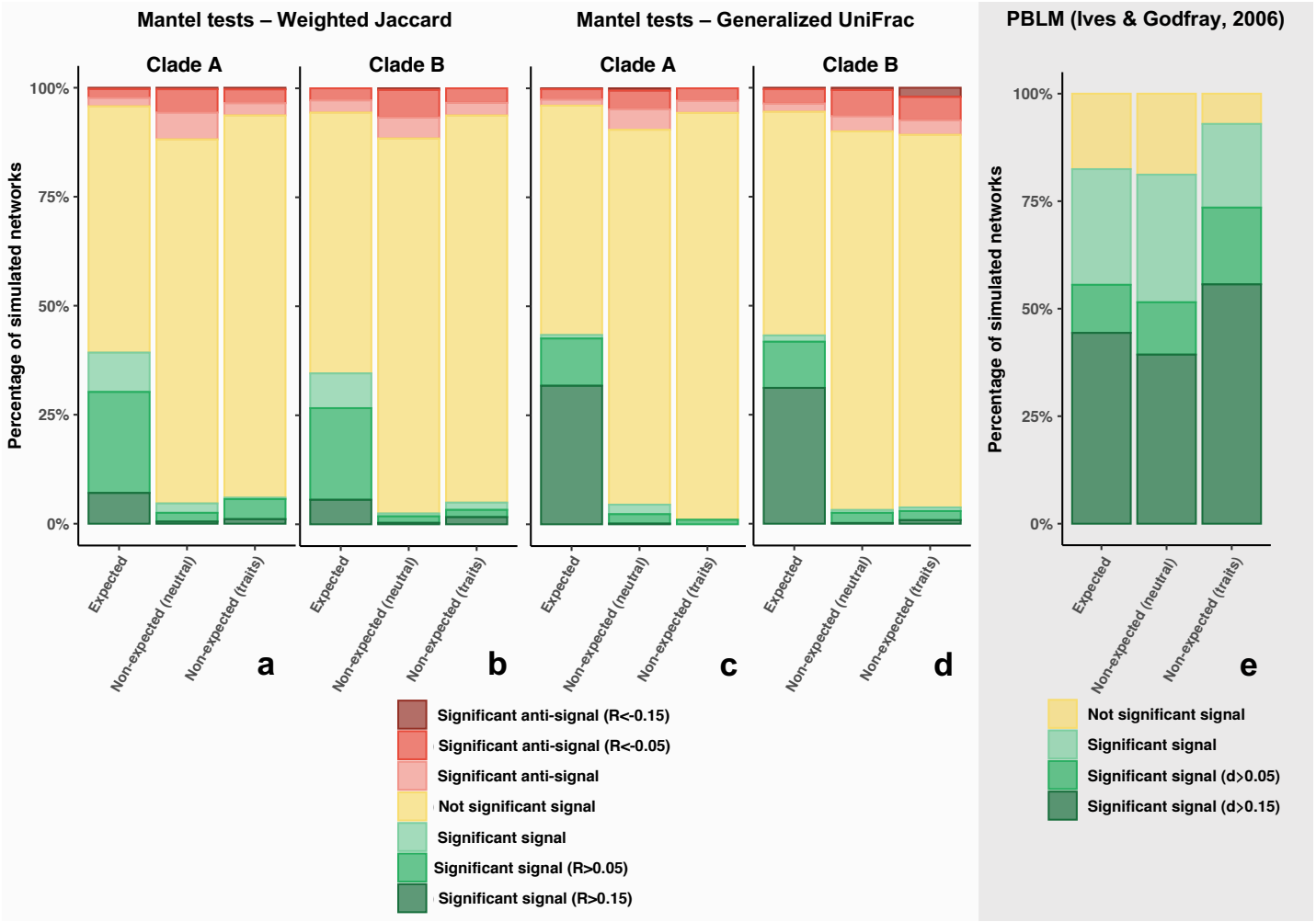


Figure II.4.2: Statistical performances of the simple Mantel tests and the Phylogenetic bipartite linear model (PBLM; Ives & Godfray, 2006) evaluated using BipartiteEvol simulations (Maliet *et al.*, 2020). For each panel, the simulations are divided between networks where phylogenetic signal in species interactions is expected (*i.e.* networks (i) simulated with an effect of the traits on individual fitness - antagonistic and mutualistic simulations - and (ii) presenting traits that are phylogenetically conserved - see Supplementary Figure 2) and networks where phylogenetic signal in species interactions is not expected: *i.e.* neutral simulations ($\alpha = 0$) or simulated networks where we observed no phylogenetic signal in the traits). a-d: Phylogenetic signals in species interactions estimated using simple Mantel tests with Pearson correlation (R) in the guilds A (a, c) and B (b, d). The different panels correspond to the 2 tested ecological distances: weighted Jaccard (a, b) or generalized UniFrac (c, d) distances. One-tailed Mantel tests between phylogenetic distances and ecological distances were performed using $10,000$ permutations. In each panel, the bars indicate the percentage of simulated networks that present a significant positive correlation (in green; upper p-value>0.05), a significant negative correlation (in red; lower p-value>0.05), or no significant correlation (in yellow; both p-values>0.05). e: Phylogenetic signals estimated using PBLM. The bar indicates the percentage of simulated networks that present no significant (in yellow; MSE≥MSEstar) or a significant (green; MSE<MSEstar) phylogenetic signals. Phylogenetic signals are shaded from light green to dark green according to the strength of the signal (*e.g.* in dark green if $d_A > 0.15$ or $d_B > 0.15$). PBLM were run on the weighted networks. In each panel, the first bar indicates the statistical power of the test, whereas the second and third bar indicate the type-I error rate of the test.

The statistical power of Mantel tests measuring phylogenetic signal in species interactions (significant positive correlation) seems to be modulated by many factors, including the network size, as phylogenetic signals were less often significant but generally stronger in smaller networks (Supplementary Figures 5-6-7). Moreover, the Mantel tests based on Pearson correlations had higher power than Spearman and Kendall correlations (Supplementary Figures 5-6-7) and the generalized UniFrac distances outperform the other ecological distances (Supplementary Figures 5-6-7; Supplementary Table 2). Finally, we surprisingly detected significant anti-phylogenetic signals in >10% of simulated networks, in particular in the small ones (Figure II.4.2, Supplementary Figures 5-6-7).

When using mean square errors (MSE) to evaluated the significance of PBLM, we found a significant phylogenetic signal in most of the simulated networks including when we did not expect any (Figure II.4.2e). The propensity of PBLM to detect phylogenetic signal decreased in large unweighted networks, but the type-I errors remained >30%, including when using a more stringent significance cutoff (Supplementary Figures 8 & 9). Similar results were obtained when bootstrapping to evaluate the significance of the phylogenetic signals (Supplementary Figure 10).

Testing the confounding effect of phylogenetic signal in degrees of generalism:

As expected, tests of phylogenetic signals in degrees of generalism were non-significant in the large majority of the BipartiteEvol networks, especially the larger ones (Supplementary Figure 11), even if we observed a correlation between ecological distances and degree difference distances (Supplementary Figure 12). When testing for phylogenetic signal in species interactions, partial Mantel tests had similar type-I error and power compared to simple Mantel tests (Supplementary Figure 5-13; Supplementary Table 2). Finally, performing sequential Mantel tests barely decreased the statistical power by <2% (Supplementary Table 2).

Networks simulated with phylogenetic conservatism in the number, but not the identity, of partners covered a realistic range of sizes and connectance values (Supplementary Figure 14). As expected, Mantel tests revealed significant phylogenetic signal in degrees of generalism in many of these networks (>60%), with an increasing percentage of significant tests with decreasing a_A (*i.e.* increasing simulated conservatism in the degrees of generalism; Supplementary Figure 15). We found a correlation between degree differences and ecological distances in most of these simulated networks (Supplementary Figure 16) and as a result of this confounding effect, when testing for phylogenetic signal in species interactions, simple Mantel tests were frequently significant (type-I error>30%; Supplementary Figure 17; Supplementary Table 3). Partial Mantel tests controlling for degree differences slightly decreased the proportion of false-positives, but it remained high (type-I error>25%; Supplementary Figure 18). In addition, partial Mantel tests detected a spurious significant anti-phylogenetic signal in species interactions in >15% of the networks (Supplementary Figure 18). Conversely, only a few networks with a signif-

icant simple Mantel test in species interactions did not have a significant simple Mantel test in degrees of generalism, such that sequential Mantel tests had only a type-I error rate ∼7% (Supplementary Table 3).

Testing the robustness of our results to phylogenetic uncertainty, sampling asymmetry, and network heterogeneity:

First, when testing for phylogenetic uncertainty, as expected, the statistical power of the Mantel tests using UniFrac distances decreased when the length of the simulated DNA sequences decreased (*i.e.* when phylogenetic uncertainty increased; Supplementary Figure 19). However, even when the simulated DNA sequences were the shortest (N=75 base pairs), resulting in very noisy reconstructed partners' tree (Supplementary Figure 20), the statistical power of the Mantel tests using UniFrac distances was still larger than when using Jaccard distances (Supplementary Figure 19).

Second, when considering sampling asymmetry, the obtained asymmetrical networks had also realistic (but higher) connectances (Supplementary Figures 21 & 22) and we found very similar trends when measuring phylogenetic signal (Supplementary Figures 23 & 24): PBLM spuriously detect phylogenetic signal when it should not, and Mantel tests had decent statistical performances, especially when using generalized UniFrac distances. In addition, the correlations of the Mantel tests in guild A were generally higher when significant (Supplementary Figure 23).

Third, when measuring clade-specific phylogenetic signals by performing separate Mantel tests while correcting for multiple testing, we recovered significant phylogenetic signals in 82% of the nodes where mutualism originated (Supplementary Figure 25).

General guidelines and illustration with application on the orchid-fungus mycorrhizal network from La Réunion:

Box 2 and Figure II.4.3 provide general guidelines based on our results and empirical considerations to measure phylogenetic signal in interaction networks, that we applied on the orchid-fungus mycorrhizal network from La Réunion. First (step 1), we computed the phylogenetic signals in species interactions for fungi and orchids using Mantel tests and found a significant but low phylogenetic signal (R<0.10) in orchid interactions with Jaccard distances as ecological distances, but its significance disappeared with UniFrac distances (Supplementary Table 4). Similarly, marginally not-significant and low phylogenetic signals were detected for mycorrhizal fungi (R<0.04; Supplementary Table 4). Next (step 2), we found no phylogenetic signal in degrees of generalism (one-tailed simple Mantel test: p-value>0.05).

Box 2: Recommended guidelines to accurately measure the phylogenetic signal in species interactions within bipartite ecological networks.

This guideline is composed of two fixed steps and following by two optional ones. It considered that a bipartite interaction network (with or without abundances) and at least the phylogenetic tree of guild A are available.

Step 1: The first step corresponds to testing the phylogenetic signal in species interactions of guild A (*i.e.* whether closely related species from guild A tend to interact with similar partners from guild B) using one-tailed simple Mantel test. This step requires to pick an ecological distance (UniFrac distances are recommended compared to Jaccard distances) and a type of correlation (Pearson correlation by default).

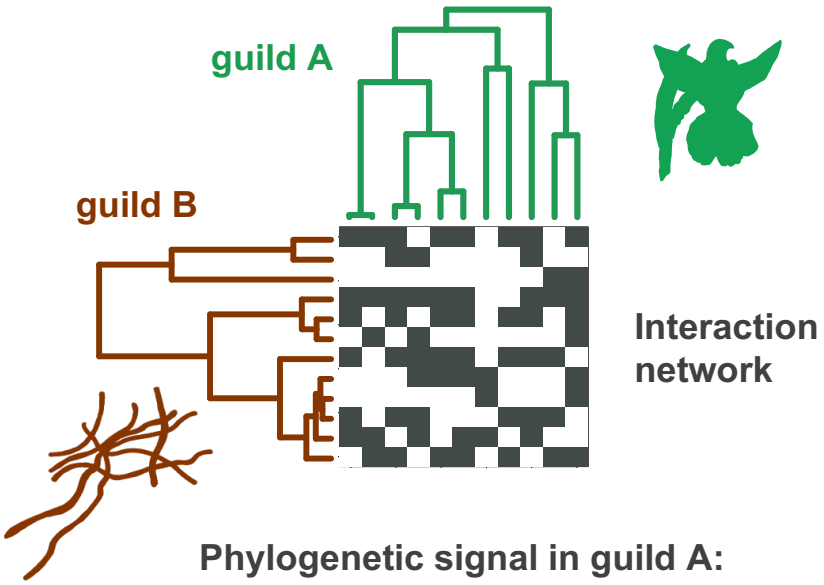
Step 2: Next, to assess whether a phylogenetic signal in species interactions really comes from the species identity, the second step consists in testing whether there is phylogenetic signal in degrees of generalism of guild A (*i.e.* whether closely related species from guild A tend to interact with the same number of partners from guild B) using a one-tailed simple Mantel test.

Option 1: Then, the first option proposes to test for the presence of clade-specific phylogenetic signal in some sub-clades of guild A using simple Mantel tests while correcting for multiple testing (*e.g.* Bonferroni correction).

Option 2: Finally, a last step can be to validate the robustness of the findings by looking at how the conclusions might be affected by phylogenetic uncertainty (*e.g.* using a Bayesian posterior of tree, by investigating the influence of polytomy, ...) or sampling bias. For the latter point, for instance, the overrepresentation of some sub-clades can cause a spurious phylogenetic signal only in this clade whereas signals in other under-sampled clades are non-significant: therefore, such a bias can be verified by subsampling all clades to the same degree.
Then, all these steps can be replicated for testing for phylogenetic signal in species interaction in guild B.

When investigating clade-specific phylogenetic signal in the orchid phylogeny (option 1), we found a significant phylogenetic signal in Angraecinae, a sub-tribe composed of epiphytic species (R=0.37; Bonferroni-corrected p-value=0.016; Figure II.4.4), suggesting that closely related Angraecinae tend to interact with more similar mycorrhizal fungi.

In addition, to check the robustness of the significant phylogenetic signal detected in Angraecinae (option 2), we (i) replaced the well-resolved *Angraecum* clade by a polytomy and (ii) subsampled down to 10 species of Angraecinae (instead of 34) and still recovered significant signal in species interactions in both cases (Supplementary Figure 26).



Phylogenetic signal in guild A:

Step 1: test the phylogenetic signal in the **species interactions** (simple Mantel test)
(i) choice of ecological distances (Jaccard, UniFrac…)
(ii) with or without interaction abundances

phylosignal_network(network, tree_A, tree_B, method = "GUniFrac", correlation = "Pearson")

Step 2: test the phylogenetic signal in the **degree of generalism** (simple Mantel test)

phylosignal_network(network, tree_A, method = "degree", correlation = "Pearson")

Option 1: investigate clade-specific phylogenetic signals (simple Mantel tests with Bonferroni correction)

phylosignal_sub_network(network, tree_A, tree_B, method = "GUniFrac", correlation = "Pearson")

Option 2: test the robustness of the findings to phylogenetic uncertainty and/or sampling bias

(repeat for guild B)

Figure II.4.3: Illustration of the recommended guidelines for accurately measuring the phylogenetic signal in species interactions within bipartite ecological networks (Box 2). At the top, a toy example of an interaction network between orchids (in green) and mycorrhizal fungi (in brown) is informed with the phylogenetic trees of each guild. For each step of the guidelines (Box 2), an example of the corresponding function available in the R-package RPANDA is indicated in grey. Note that the phylogenetic tree does not need to be binary, rooted, or ultrametric.

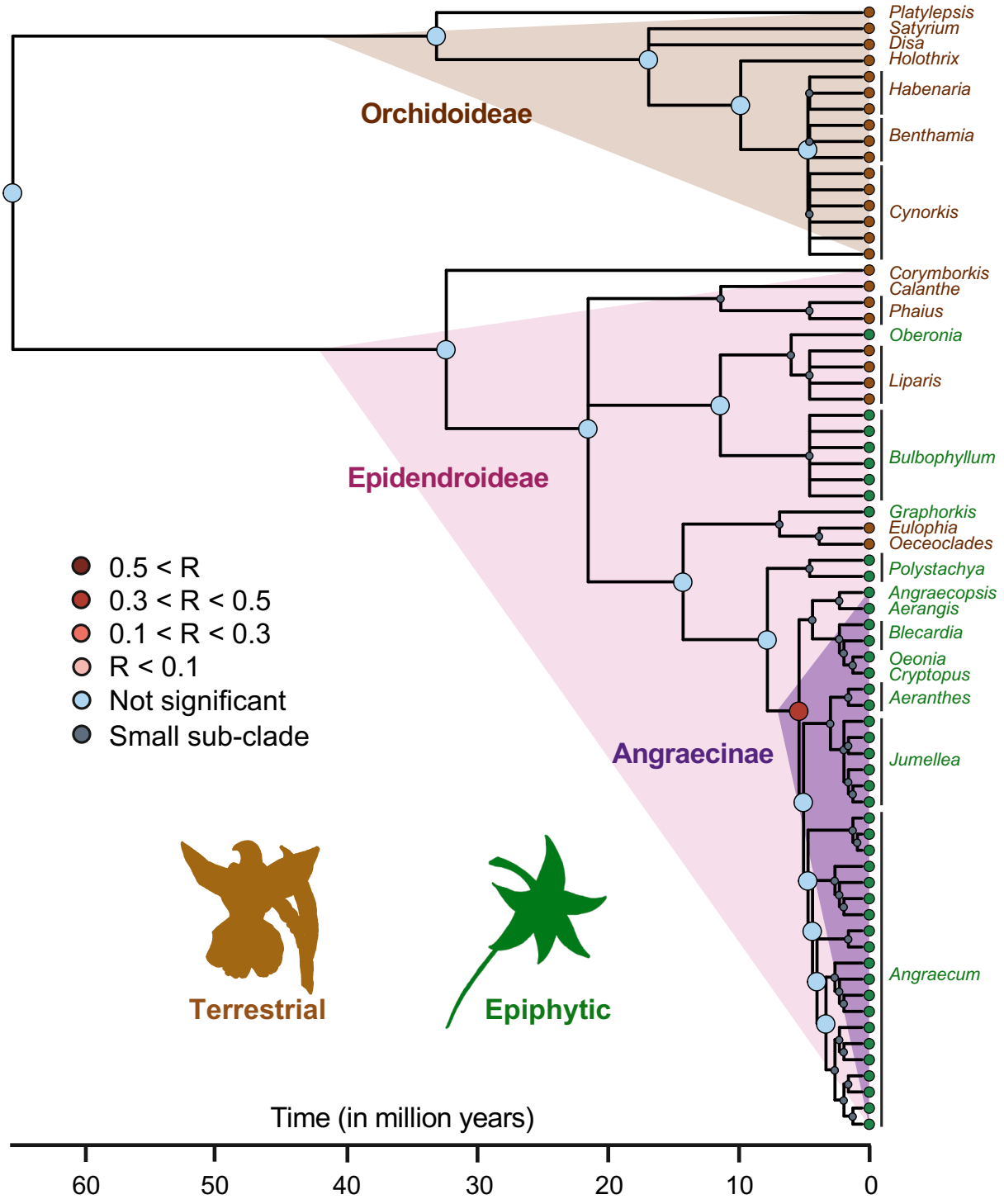


Figure II.4.4: Empirical application on an orchid-fungus interaction network from La Réunion island (Martos *et al.*, 2012): the clade-specific analyses of phylogenetic signal in species interactions revealed a significant phylogenetic signal in the epiphytic subtribe Angraecinae. The orchid phylogeny (Martos *et al.*, 2012) is represented with its nodes colored according to the results of the Mantel test performed on the corresponding sub-network: in blue if non-significant, in grey when the node has less than 10 descendent species (the Mantel test was not performed), and in red when the phylogenetic signal is significant. Each one-tailed simple Mantel test was performed using the Pearson correlation and 100,000 permutations and its significance was evaluated while correcting for multiple testing (Bonferroni correction). For each species, its habitat (terrestrial or epiphytic) is indicated at the tips of the tree and the main orchid clades are highlighted in colors. Only the genera are indicated at the tips of the tree (see Supplementary Figure 26 for the species list).

Discussion:

Here, we used simulations to perform a comparative analysis of Mantel tests and the Phylogenetic bipartite linear model (PBLM; Ives & Godfray 2006) used for measuring the phylogenetic signals in species interactions. Our results highlighted the weaknesses of partial Mantel tests and PBLM. Conversely, we argue that simple Mantel tests present rather good statistical performances.

Simple Mantel tests appeared to accurately measure phylogenetic signals in species interactions, with a decent type-I error rate and a moderate statistical power. Although correlations between phylogenetic and ecological distances are not particularly expected to be linear, Pearson correlations performed better than Spearman and Kendall correlations which only test for monotonic correlation (Box 1), but probably loose information. Among ecological distances, considering interaction abundances and phylogenetic relatedness of the partners using generalized UniFrac distances significantly improved the detection of phylogenetic signal, even when the reconstructed partners trees were not robust. In addition, given that species delineations may be somewhat arbitrary, especially in microbial interactors, and that Jaccard distances are directly sensitive to species delineation (Sanders *et al.*, 2014), we advocate for using generalized UniFrac distances. However, if one suspects recent speciations of the partners to be the main differences in the community composition, one should better use Jaccard distances, as UniFrac distances put more weight on differences in long branches instead of recent splits (Sanders *et al.*, 2014). In addition, given that we expected anti-phylogenetic signal in species interactions only in less than 1% of the BipartiteEvol simulations, we highlighted the high proportion of false-positives (5-10%) when testing for anti-phylogenetic signals in empirical networks using Mantel tests and encourage interpreting them cautiously. Finally, although our simulations accounting for network heterogeneity were limited and would require further testing, we found that using simple Mantel tests to investigate clade-specific phylogenetic signals perform rather well and that a Bonferroni correction for multiple testing was not too stringent. Such approach can therefore be valuable for measuring local phylogenetic signal in large "meta-networks" which likely presents high heterogeneity, *e.g.* when measuring host-microbiota phylosymbiosis (Song *et al.*, 2020).

At best, we only retrieved significant phylogenetic signals in species interactions in 40% of the BipartiteEvol networks where signal was expected (Supplementary Table 2). The power was particularly low when the effect of traits on species fitness is low (low α values), meaning that when phylogenetically-conserved trait matching only slightly impacts the finesses of the interacting species, current interactions do not retain any signal of past interactions. We also confirmed the propensity of phylogenetic signal in antagonisms compared to mutualisms (Rohr & Bascompte, 2014; Nuismer & Harmon, 2015), which tend to be driven by the specialization of the predators/parasites on phylogenetically related preys/hosts (Fontaine & Thébault, 2015). However, we acknowledge that

these simulations are limited by many aspects, including the facts that they considered a concomitant single origin of both interacting lineages that then evolve sympatrically, and neglected the heterogeneity within each guild; this might impact the generality of our findings towards all interaction networks. Nevertheless, our results seem robust to different sampling strategies such as the asymmetrical sampling of the guilds, which is particularly frequent when studying microbial symbiosis (Jacquemyn *et al.*, 2011; Martos *et al.*, 2012; Song *et al.*, 2020). The generality of correlative approaches such as the Mantel tests that do not rely on strong hypotheses likely makes them robust to various network structures and sampling strategies.

Although partial Mantel tests are frequently used when investigating phylogenetic signal in species interactions while controlling for degrees of generalism, our simulations demonstrated that their type-I error rate was very high (they detected significant signals in species interactions when we only simulated signals in degrees of generalism) and that their statistical power was moderate (similar to the power of a regular Mantel test). Thus, partial Mantel tests fail at discerning whether phylogeny strictly affects the identity of partners, independently of the total number of partners associated with each species (Rezende *et al.*, 2007). This corroborates the poor statistical performances of partial Mantel tests frequently observed in other contexts (Harmon & Glor, 2010; Guillot & Rousset, 2013). To reliably assess whether the phylogenetic signal is only in the identity of species interactions and not in degrees of generalism, we rather suggest to perform sequential simple Mantel tests testing first for phylogenetic signal in species interactions, and if significant, testing for phylogenetic signal in degrees of generalism. Such an approach has a low type-I error rate and a very limited power decrease. However, sequential simple Mantel tests do not allow testing if there is still a signal in species identity when there is a signal in degrees of generalism. This could be tested separately by selecting only sets of species having similar degrees (*e.g.* specialist species) and applying a simple Mantel tests measuring their phylogenetic signal in species interactions.

The Phylogenetic bipartite linear model (PBLM) unreliably assessed the phylogenetic signal in species interactions in networks simulated using BipartiteEvol. As explained in Box 1, PBLM assumes that the interaction strength between the two species is determined by the product of two unobserved traits evolving on the phylogenies of species from guilds A and B respectively, according to two independent OU processes with the selection strengths d_A and d_B: PBLM tests the significance of d_A and d_B, which measure the phylogenetic signal of the unobserved traits. A species with a high trait value will have high interaction strengths with many partner species (generalist), while a species with a low trait value will have low interaction strengths with most partner species, except with the few species with high trait values (specialist). Therefore, instead of measuring phylogenetic signal in species interactions, we argue that d_A and d_B measure the phylogenetic signals in degrees of generalism. However, the strong hypotheses made by PBLM to explain how the degrees of generalism evolve and how interactions assemble might pre-

vent its validity in a general context (model misspecification). Therefore, PBLM should not be used as a routine for measuring phylogenetic signals in empirical networks. As a future direction, the validity of PBLM should be investigated when using specific measures of interaction strengths (*e.g.* parasitic attack rates) or when using other strategies to evaluate the model significance, given that bootstrapping for instance seems incorrect when interaction networks have low connectances (Ives & Godfray, 2006). It exists other model-based approaches (Rafferty & Ives, 2013; Hadfield *et al.*, 2014) that are extensions of PBLM and propose to infer more parameters describing the phylogenetic structure of interactions networks, while also offering the possibility to control for heterogeneity in sampling effort and spatial variations (Hadfield *et al.*, 2014). Although the ability of such approaches to correctly infer phylogenetic signal should also be tested with simulations, their very computationally intensive inference prohibited their incorporation in our comparative analyses. Given that phylogenetic signals only measure general patterns (Losos, 2008), the advances of such integrative model-based approaches should pave the way toward a better understanding of the ecological or evolutionary processes playing a role in the assembly of interaction networks (Harmon *et al.*, 2019). In the meantime, phylogenetic signals measured using Mantel tests represent a quite reliable and very rapid analysis that can easily help to understand the importance of evolutionary processes in structuring empirical networks.

In the mycorrhizal network from La Réunion, we found non-significant or weak phylogenetic signals in species interactions at the level of the entire orchid-fungus network, suggesting these interactions are generally poorly conserved over long evolutionary timescales. Conversely, clade-specific Mantel tests detected a significant phylogenetic signal in the Angraecinae epiphytic clade. This signal is likely produced by the different orchids genera in Angraecinae associating with specific fungal clades (Martos *et al.*, 2012). Thus, our results corroborate a trend toward mycorrhizal specialization in epiphytic orchids compared with terrestrial species (Xing *et al.*, 2019), as the epiphytic habitats might require particular adaptations and stronger dependences toward specific mycorrhizal fungi.

Interaction networks are increasingly being analyzed to unravel the evolutionary processes shaping their structure and to predict their stability in the context of global changes. Currently-used tools for measuring phylogenetic signals are clearly misleading. We provide an alternative approach based on sequential Mantel tests, and by emphasizing the limits of current model-based approaches, we hope to stimulate new developments in model-based tests of phylogenetic signal.

References:

Aizen MA, Gleiser G, Sabatino M, Gilarranz LJ, Bascompte J, Verdú M. 2016. The phylogenetic structure of plant-pollinator networks increases with habitat size and isolation. Ecology Letters 19: 29–36.

Bascompte J, Jordano P. 2013. Mutualistic networks. Princeton: Princeton University Press.

Bascompte J, Jordano P, Melián CJ, Olesen. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences of the United States of America 100: 9383–9387.

Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57: 717–745.

Chagnon PL. 2016. Seeing networks for what they are in mycorrhizal ecology. Fungal Ecology 24: 148–154.

Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28: 2106–2113.

Clavel J, Escarguel G, Merceron G. 2015. mvMORPH: An R-package for fitting multivariate evolutionary models to morphometric data (T Poisot, Ed.). Methods in Ecology and Evolution 6: 1311–1319.

Elias M, Fontaine C, Frank Van Veen FJ. 2013. Evolutionary history and ecological processes shape a local multilevel antagonistic network. Current Biology 23: 1355–1359.

Felsenstein J. 1985. Phylogenies and the comparative method. American Naturalist 125: 1–15.

Fontaine C, Guimarães PR, Kéfi S, Loeuille N, Memmott J, van der Putten WH, van Veen FJF, Thébault E. 2011. The ecological and evolutionary implications of merging different types of networks. Ecology Letters 14: 1170–1181.

Fontaine C, Thébault E. 2015. Comparing the conservatism of ecological interactions in plant-pollinator and plant-herbivore networks. Population Ecology 57: 29–36.

Goslee SC, Urban DL. 2007. The ecodist package for dissimilarity-based analysis of ecological data. Journal of Statistical Software 22: 1–19.

Guillot G, Rousset F. 2013. Dismantling the Mantel tests. Methods in Ecology and Evolution 4: 336–344.

Hadfield JD, Krasnov BR, Poulin R, Nakagawa S. 2014. A tale of two phylogenies: Comparative analyses of ecological interactions. The American Naturalist 183: 174–187.

Hansen TF, Martins EP. 1996. Translating between microevolutionary process and macroevolutionary patterns: The correlation structure of interspecific data. Evolution 50: 1404–1417.

Harmon LJ, Andreazzi CS, Débarre F, Drury J, Goldberg EE, Martins AB, Melián CJ, Narwani A, Nuismer SL, Pennell MW, *et al.* 2019. Detecting the macroevolutionary signal of species interactions. Journal of Evolutionary Biology 32: 769–782.

Harmon LJ, Glor RE. 2010. Poor statistical performance of the Mantel test in phylogenetic comparative analyses. Evolution 64: 2173–2178.

Ives AR, Godfray HCJ. 2006. Phylogenetic analysis of trophic associations. The American Naturalist 168: E1–E14.

Jacquemyn H, Merckx VSFT, Brys R, Tyteca D, Cammue BPAA, Honnay O, Lievens B. 2011. Analysis of network architecture reveals phylogenetic constraints on mycorrhizal specificity in the genus *Orchis* (Orchidaceae). New Phytologist 192: 518–528.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26: 1463–1464.

Krasnov BR, Fortuna MA, Mouillot D, Khokhlova IS, Shenbrot GI, Poulin R. 2012. Phylogenetic signal in module composition and species connectivity in compartmentalized host-parasite networks. American Naturalist 179: 501–511.

Losos JB. 2008. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecology Letters 11: 995–1003.

Maliet O, Loeuille N, Morlon H. 2020. An individual based model for the eco-evolutionary emergence of bipartite interaction networks (T Poisot, Ed.). Ecology Letters: ele.13592.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209–220.

Martín González AM, Dalsgaard B, Nogués-Bravo D, Graham CH, Schleuning M, Maruyama PK, Abrahamczyk S, Alarcón R, Araujo AC, Araújo FP, *et al.* 2015. The macroecology of phylogenetically structured hummingbird-plant networks. Global Ecology and Biogeography 24: 1212–1224.

Martos F, Munoz FF, Pailler T, Kottke I, Gonneau C, Selosse MA. 2012. The role of epiphytism in architecture and evolutionary constraint within mycorrhizal networks of tropical orchids. Molecular Ecology 21: 5098–5109.

Morlon H, Lewitus E, Condamine FL, Manceau M, Clavel J, Drury J. 2016. RPANDA: An R-package for macroevolutionary analyses on phylogenetic trees. Methods in Ecology and Evolution 7: 589–597.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W. 2012. How to measure and test phylogenetic signal. Methods in Ecology and Evolution 3: 743–756.

Nuismer SL, Harmon LJ. 2015. Predicting rates of interspecific interaction from phylogenetic trees. Ecology Letters 18: 17–27.

Pagel M. 1999. Inferring the historical patterns of biological evolution. Nature 401: 877–884.

Paradis E. 2013. Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. Molecular Phylogenetics and Evolution 67: 436–444.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Peralta G. 2016. Merging evolutionary history into species interaction networks. Functional Ecology 30: 1917–1925.

Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host-microbiota evolution: Application to the great apes gut microbiota. Molecular Ecology Resources 19: 1659–1671.

R Core Team. 2020. R: A language and environment for statistical computing.

Rafferty NE, Ives AR. 2013. Phylogenetic trait-based analyses of ecological networks. Ecology 94: 2321–2333.

Revell LJ. 2012. phytools: An R-package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rezende EL, Lavabre JE, Guimarães PR, Jordano P, Bascompte J. 2007. Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 448: 925–928.

Rohr RP, Bascompte J. 2014. Components of phylogenetic signal in antagonistic and mutualistic networks. The American Naturalist 184: 556–564.

Sanders JG, Powell S, Kronauer DJC, Vasconcelos HL, Frederickson ME, Pierce NE. 2014. Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Molecular Ecology 23: 1268–1283.

Smouse PE, Long JC, Sokal RR. 1986. Multiple regression and correlation extensions of the mantel test of matrix correspondence. Systematic Zoology 35: 627–632.

Song SJ, Sanders JG, Delsuc F, Metcalf J, Amato K, Taylor MW, Mazel F, Lutz HL, Winker K, Graves GR, *et al.* 2020. Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. mBio 11: 1–14.

Thébault E, Fontaine C. 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329: 853–856.

Vázquez DP, Chacoff NP, Cagnolo L. 2009. Evaluating multiple determinants of the structure of plant–animal mutualistic networks. Ecology 90: 2039–2046.

Xing X, Jacquemyn H, Gai X, Gao Y, Liu Q, Zhao Z, Guo S. 2019. The impact of life form on the architecture of orchid mycorrhizal networks in tropical forest. Oikos 128: 1254–1264.

Xing X, Liu Q, Gao Y, Shao S, Guo L, Jacquemyn H, Zhao Z, Guo S. 2020. The architecture of the network of orchid–fungus interactions in nine co-occurring *Dendrobium* species. Frontiers in Ecology and Evolution 8: 1–10.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3uQQDpd` or by scanning:

Article 5: Global drivers of obligate mycorrhizal symbionts diversification

Authors: Benoît Perez-Lamarque[1,2], Maarja Öpik[3], Odile Maliet[1], Ana C. Afonso Silva[1], Marc-André Selosse[1,4], Florent Martos[2], Hélène Morlon[1]

[1] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[2] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

[3] University of Tartu, 40 Lai Street, 51 005 Tartu, Estonia

[4] Department of Plant Taxonomy and Nature Conservation, University of Gdansk, Wita Stwosza 59, 80-308 Gdansk, Poland

Abstract

Arbuscular mycorrhizal fungi (AMF) are widespread microscopic fungi that provide mineral nutrients to most land plants by forming one of the oldest terrestrial symbioses. They have sometimes been referred to as an "evolutionary cul-de-sac" for their limited species diversity and their ecological niches restricted to plant-symbiotic life style. Here we use the largest global database of AMF to analyze their diversification dynamics in the past 500 million years (Myr) based on the small subunit (SSU) rRNA gene. We find that overall AMF have low diversification rates. After a diversification peak between 200 and 100 Myr ago, they experienced an important diversification slowdown toward the present. Such a slowdown could be at least partially related to a shrinking of their mycorrhizal niches and to their limited ability to colonize new (non-mycorrhizal) niches. Given that estimating the diversification history of a microbial clade using a single slowly-evolving marker gene can be problematic, we performed a range of sensitivity to assess the robustness of our results. Our results identify patterns and drivers of diversification in a group of obligate symbionts of major ecological and evolutionary importance.

Keywords: microbial diversification, arbuscular mycorrhiza, obligate symbiosis, ecological niche, fungi.

Author contributions: All the authors designed the study. MÖ gathered the data and BPL performed the analyses. OM and ACAS provided some codes. BPL and HM wrote

the first version of the manuscript and all authors contributed substantially to the revisions.

Acknowledgments: The authors acknowledge C. Strullu-Derrien, M. Elias, D. de Vienne, A. Vogler, J.-Y. Dubuisson, C. Quince, S.-K. Sepp, and M. Chase for helpful discussions. They also thank L. Aristide, S. Lambert, J. Clavel, I. Quintero, I. Overcast, and G. Sommeria for comments on an early version of the manuscript and David Marsh for English editing. BPL acknowledges B. Robira, F. Foutel-Rodier, F. Duchenne, E. Faure, E. Kerdoncuff, R. Petrolli, and G. Collobert for useful discussions and C. Fruciano and E. Lewitus for providing codes. This work was supported by a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL and the École Doctorale FIRE – Programme Bettencourt. MÖ was supported by the European Regional Development Fund (Centre of Excellence EcolChange) and University of Tartu (PLTOM20903). Funding of the research of FM was from the Agence Nationale de la Recherche (ANR-19-CE02-0002). HM acknowledges support from the European Research Council (grant CoG-PANDA).

Data availability: All of the data used in this study are available in the open-access MaarjAM database (`https://maarjam.botany.ut.ee`). Spore lengths of AMF were collected in the supplementary data of Davison *et al.* (2018).

Citation: Perez-Lamarque B, Öpik M, Maliet O, Afonso Silva AC, Selosse MA, Martos F, Morlon H. Global drivers of obligate mycorrhizal symbionts diversification. ***under review***.

Introduction:

Arbuscular mycorrhizal fungi (AMF - subphylum Glomeromycotina) are obligate symbionts that have been referred to as an "evolutionary cul-de-sac, albeit an enormously successful one" (Malloch, 1987; Morton, 1990). This alludes to their ecological success despite limited morphological and species diversities: they associate with the roots of >80% of land plants, where they provide mineral resources in exchange for photosynthates (Smith & Read, 2008). Present in most terrestrial ecosystems, AMF play key roles in plant protection, nutrient cycling, and ecosystem functions (van der Heijden *et al.*, 2015). Fossil evidence and molecular phylogenies suggest that AMF contributed to the emergence of land plants (Selosse & Le Tacon, 1998; Field *et al.*, 2015; Strullu-Derrien *et al.*, 2018; Feijen *et al.*, 2018) and coevolved with them for more than 400 million years (Myr) (Simon *et al.*, 1993; Lutzoni *et al.*, 2018; Strullu-Derrien *et al.*, 2018).

Despite the ecological ubiquity and evolutionary importance of AMF, large-scale patterns of their evolutionary history are poorly known. Studies on the diversification of AMF have been hampered by the difficulty of delineating species, quantifying global

scale species richness, and building a robust phylogenetic tree for this group. Indeed, AMF are microscopic soil- and root-dwelling fungi that are poorly differentiated morphologically and difficult to cultivate. Although their classical taxonomy is mostly based on the characters of spores and root colonization (Smith & Read, 2008; Stürmer, 2012), AMF species delineation has greatly benefited from molecular data (Krüger *et al.*, 2012). Experts have defined "virtual taxa" (VT) based on a minimal 97% similarity of a region of the 18S small subunit (SSU) rRNA gene and monophyly criteria (Öpik *et al.*, 2010, 2014). As for many other pragmatic species delineations, VT have rarely been tested for their biological relevance (Powell *et al.*, 2011), and a consensual system of AMF classification is still lacking (Bruns *et al.*, 2018). AMF are also poorly known genetically: the full SSU rRNA gene sequence is known in few species (Rimington *et al.*, 2018), other gene sequences in even fewer (James *et al.*, 2006; Lutzoni *et al.*, 2018), and complete genomes in very few (Venice *et al.*, 2020).

The drivers of AMF diversification are unknown. A previous dated phylogenetic tree of VT found that many speciations occurred after the last major continental reconfiguration around 100 Myr ago (Davison *et al.*, 2015), suggesting that AMF diversification is not linked to vicariant speciation during this geological event. Still, geographical speciation could play an important role in AMF diversification, as these organisms have spores that disperse efficiently (Egan *et al.*, 2014; Bueno & Moora, 2019; Correia *et al.*, 2019), which could result in frequent founder-event speciation (Templeton, 2008). Other abiotic factors include habitat: tropical grasslands have, for example, been suggested as diversification hotspots for AMF (Pärtel *et al.*, 2017). Besides abiotic factors, AMF are obligate symbionts and, although relatively generalist (Sanders, 2003; van der Heijden *et al.*, 2015; Article 6), their evolutionary history could be largely influenced by a diffuse coevolution with their host plants (Zanne *et al.*, 2014; Lutzoni *et al.*, 2018; Sauquet & Magallón, 2018). Over the last 400 Myr, land plants have experienced massive extinctions and radiations (Cleal & Cascales-Miñana, 2014; Zanne *et al.*, 2014), adaptations to various ecosystems (Bredenkamp *et al.*, 2002; Brundrett & Tedersoo, 2018), and associations with different soil microorganisms (Werner *et al.*, 2014, 2018). All these factors could have influenced diversification dynamics in AMF.

Here, we reconstruct several thoroughly sampled phylogenetic trees of AMF, considering several criteria of species delineations and uncertainty in phylogenetic reconstructions. We combine this phylogenetic data with paleoenvironmental data and data of current AMF geographic distributions, ecological traits, interaction with host plants, and genetic diversity to investigate the global patterns and drivers of AMF diversification in the last 500 Myr.

Methods:

Virtual taxa phylogenetic reconstruction:

We downloaded the AMF SSU rRNA gene sequences from MaarjAM, the largest global database of AMF gene sequences (Öpik *et al.*, 2010). We reconstructed several Bayesian phylogenetic trees of the 384 VT from the corresponding representative sequences available in the MaarjAM database (Öpik *et al.*, 2010) updated in June 2019 (Supplementary Methods 1). We used the full length (1,700 base pairs) SSU rRNA gene sequences from (Rimington *et al.*, 2018) to better align the VT sequences using MAFFT (Katoh & Standley, 2013). We selected the 520 base pair central variable region of the VT aligned sequences and performed a Bayesian phylogenetic reconstruction using BEAST2 (Bouckaert *et al.*, 2014). We obtained a consensus VT tree and selected 12 trees equally spaced in 4 independent Bayesian chains to account for phylogenetic uncertainty in the subsequent diversification analyses, hereafter referred to as the VT replicate trees. We set the crown root age at 505 Myr (Davison *et al.*, 2015), which is coherent with fossil data and previous dated molecular phylogenies (Lutzoni *et al.*, 2018; Strullu-Derrien *et al.*, 2018).

Delineation into Evolutionary Units (EUs):

We considered several ways to delineate AMF species based on the SSU rRNA gene. In addition to the VT species proxy, we delineated AMF de novo into evolutionary units (EUs) using 5 different thresholds of sequence similarity ranging from 97 to 99% and a monophyly criterion. We gathered 36,411 AMF sequences of the SSU rRNA gene from MaarjAM, mainly amplified by the primer pair NS31–AML2 (variable region) (Simon *et al.*, 1992; Lee *et al.*, 2008) (dataset 1, Supplementary Table 1), corresponding to 27,728 haplotypes. We first built a phylogenetic tree of these haplotypes and then applied to this tree our own algorithm (R-package RPANDA; Morlon *et al.*, 2016; R Core Team, 2020) that traverses the tree from the root to the tips, at every node computes the average similarity of all sequences descending from the node, and collapses the sequences into a single EU if their sequence dissimilarity is lower than a given threshold (Supplementary Methods 2). Finally, we performed Bayesian phylogenetic reconstructions of the EUs using BEAST2 (Supplementary Methods 1).

Coalescent-based species delineation analyses:

Finally, we considered the Generalized Mixed Yule Coalescent method (GMYC) (Pons *et al.*, 2006; Fujisawa & Barraclough, 2013), a species delineation approach that does not require specifying an arbitrary similarity threshold. GMYC estimates the time t in a reconstructed calibrated tree that separates species diversification (Yule process – before t) and intraspecific differentiation (coalescent process – after t). GMYC is too computationally intensive to apply on the 36,411 SSU sequences; we used it here on three smaller clades to investigate the ability of the SSU gene to delineate AMF species despite its

slow evolution (Bruns *et al.*, 2018), and as a way to evaluate the biological relevance of the VT and various EUs delineations. We selected the following AMF clades: the family Claroideoglomeraceae; the order Diversisporales; and an early-diverging clade composed of the orders Archaeosporales and Paraglomerales. For each clade, we reconstructed Bayesian phylogenetic trees of haplotypes (Supplementary Methods 1). We then ran GMYC analyses (splits R-package; Ezard *et al.*, 2009) on each of these trees and evaluated the support of the GMYC model compared to a null model in which all tips are assumed to be different species, using a likelihood ratio test (LRT). If the LRT supports the GMYC model, different SSU haplotypes belong to the same AMF species, *i.e.* the SSU rRNA gene has time to accumulate substitutions between AMF speciation events.

Total diversity estimates:

We evaluated how thoroughly sampled our species-level AMF phylogenetic trees are by estimating the total number of VT and EUs using rarefaction curves and the Bayesian Diversity Estimation Software (BDES; Quince *et al.*, 2008) (Supplementary Methods 3).

Diversification analyses:

We estimated lineage-specific diversification rates using ClaDS, a Bayesian diversification model that accounts for rate heterogeneity by modeling small rate shifts at speciation events (Maliet *et al.*, 2019). At each speciation event, the descending lineages inherit new speciation rates sampled from a log-normal distribution with an expected value $\log[\alpha \times \lambda]$ (where λ represents the parental speciation rate and α is a trend parameter) and a standard deviation σ. We considered the model with constant turnover ϵ (*i.e.* constant ratio between extinction and speciation rates; ClaDS2) and ran a newly-developed ClaDS algorithm based on data augmentation techniques which enables us to estimate mean rates through time (Maliet & Morlon, 2020). We ran ClaDS with 3 independent chains, checked their convergence using a Gelman-Rubin diagnostic criterion (Gelman & Rubin, 1992), and recorded lineage-specific speciation rates. We also recorded the estimated hyperparameters (α, σ, ϵ) and the value $m = \alpha \times exp(\sigma^2/2)$, which indicates the general trend of the rate through time (Maliet *et al.*, 2019).

In addition, we applied TreePar (Stadler, 2011), another diversification approach that does not consider rate variation across lineages, but models temporal shifts in diversification rates affecting all lineages simultaneously. We searched for up to ten shifts in diversification rates at every 2-million-year interval in each phylogenetic tree. We estimated the number of temporal shifts in AMF diversification rates using maximum likelihood inferences and likelihood ratio tests. We also used CoMET, its equivalent piecewise-constant model in a Bayesian framework (TESS R-package; Höhna *et al.*, 2016; May *et al.*, 2016). We chose the Bayesian priors according to maximum likelihood estimates from TreePar, disallowed mass extinction events, and ran the MCMC chains until convergence (minimum effective sample sizes of 500).

We also fitted a series of time-dependent and environment-dependent birth-death diversification models using RPANDA (Condamine *et al.*, 2013; Morlon *et al.*, 2016) to confirm the observed temporal trends and test the influence of temperature, pCO_2, and land plant fossil diversity on AMF diversification. For the time-dependent models, we considered models with constant or exponential variation of speciation rates through time and null or constant extinction rates (fit_bd function). As extinction is notoriously hard to estimate from reconstructed phylogenies (Rabosky, 2016), we tested the robustness of the inferred temporal trend in speciation when fixing arbitrarily high levels of extinction (Supplementary Methods 4). For the environment-dependent models, we considered an exponential dependency of the speciation rates with the environmental variable (env), *i.e.* speciation rate=b*exp(a*env), where a and b are two parameters estimated by maximum likelihood (fit_env function). Best-fit models were selected based on the corrected Akaike information criterion (AICc), considering that a difference of 2 in AICc indicates that the model with the lowest AICc is better.

The influence of temperature was tested on the complete AMF phylogenetic trees, using estimates of past global temperature (Royer *et al.*, 2004). As these temporal analyses can be sensitive to the root age calibration, we replicated them using the youngest (437 Myr) and oldest (530 Myr) crown age estimates from (Lutzoni *et al.*, 2018). We also carried a series of simulation analyses to test the robustness of our temperature-dependent results (Supplementary Methods 5). The influence of pCO_2 (Foster *et al.*, 2017) and of land plant fossil diversity was tested starting from 400 Myr ago, as these environmental data are not available for more ancient times. For these analyses we sliced the phylogenies at 400 and 200 Myr ago, and applied the diversification models to the sliced sub-trees larger than 50 tips. Estimates of land plant diversity were obtained using all available Embryophyta fossils from the Paleobiology database (`https://paleobiodb.org`) and using the shareholder quorum subsampling method (Supplementary Methods 6; Alroy, 2010).

Finally, because our results can be sensitive to incorrect species delineations toward the present (Moen & Morlon, 2014), we replicated the RPANDA analyses by excluding the last 50 Myr, following Lewitus *et al.* (2018).

All diversification analyses were performed for each delineation on the consensus and on the 12 replicate trees to account for phylogenetic uncertainty (we even used 100 replicate trees when the 12 trees gave different results). We considered missing species by imputing sampling fractions, computed as the number of observed VT or EUs divided by the corresponding BDES estimates of global AMF diversity (Supplementary Table 2). We also replicated all diversification analyses using lower sampling fractions down to 50%.

Testing for potential drivers of AMF diversification:

To further investigate the potential factors driving AMF diversification, we assessed the relationship between lineage-specific estimates of present-day speciation rates and characteristics of each AMF taxonomic unit, *i.e.* VT or EUs.

First, we characterized AMF relative niche width using a set of 10 abiotic and biotic variables recorded in MaarjAM database for each AMF unit. In short, among a curated dataset containing AMF sequences occurring only in natural ecosystems (dataset 2; Supplementary Table 3; Article 6), for each AMF unit, we reported the number of continents, ecosystems, climatic zones, biogeographic realms, habitats, and biomes where it was sampled, as well as its number of plant partners, their phylogenetic diversity, and its centrality in the plant-fungus bipartite network, and performed a principal component analysis (PCA; Supplementary Methods 7). For AMF units represented by at least 10 sequences, we tested whether these PCA coordinates reflecting AMF niche widths were correlated with the present-day speciation rates using both linear mixed-models (not accounting for phylogeny) or MCMCglmm models (Hadfield, 2010). For MCMCglmm, we assumed a Gaussian residual distribution, included the fungal phylogenetic tree as a random effect, and ran the MCMC chains for 1,300,000 iterations with a burn-in of 300,000 and a thinning interval of 500.

Next, we tested the relationship between speciation rates and geographic characteristics of AMF units. To test the effect of latitude, we associated each AMF unit with its set of latitudes and used similar MCMCglmm with an additional random effect corresponding to the AMF unit. To account for inhomogeneous sampling along the latitudinal gradient, we re-ran the model on jackknifed datasets (we re-sampled 1,000 interactions per slice of latitude of twenty degrees). Similarly, we tested the effect of climatic zone and habitat on the speciation rates.

Finally, to test the effect of dispersal capacity, we assessed the relationship between spore size and speciation rate for the few (n=32) VT that contain sequences of morphologically characterized AMF isolates (Davison *et al.*, 2018). We gathered measures of their average spore length (Davison *et al.*, 2018) and tested their relationship with speciation rate by using a phylogenetic generalized least square regression (PGLS).

Estimating genetic diversity:

As a first attempt at connecting AMF macroevolutionary diversification to microevolutionary processes, we measured intraspecific genetic diversities across AMF units. For each AMF unit containing at least 10 sequences, we computed genetic diversity using Tajima's estimator (Tajima, 1983; θ_π; Supplementary Methods 8). Using similar statistical tests as above, we investigated the correlation of AMF genetic diversity with speciation rate, niche width, geographic characteristics, and spore size. We tested the robustness

of the results to the minimal number of sequences per AMF unit (10, 15, or 20) used to compute genetic diversity and to perform the PCA.

These statistical models were replicated on the different phylogenetic trees (consensus or replicates) for each delineation and we reported p-values (P) corresponding to two-sided tests.

Assessing the robustness of our findings:

Investigating the diversification history of a clade of species that diverged more than 500 Myr ago using a single slowly evolving gene to delineate species and reconstruct the phylogenetic tree is challenging and can lead to bias (Moen & Morlon, 2014). Therefore, we assessed the robustness of our finding by (i) using simulations to show that diversification inferences can be performed in this context and (ii) by investigating whether similar trends were observed when using another AMF gene.

First, we simulated the diversification of clades of species in the last 505 Myr, according to three scenarios: (i) constant speciation rate and no extinction, (ii) constant speciation and extinction rates, and (iii) declining speciation rate and constant extinction (Supplementary Figure 1a). On the obtained trees, we used the function *simulate_alignment* (R-package HOME; Article 1) to simulate the evolution of short 520 bp DNA sequences with a substitution rate of 0.001 event per Myr and only 25% of variable sites, which mimicked the AMF SSU rRNA alignment. Next, we applied the same pipelines as above to reconstruct the phylogenetic tree of the simulated clades and infer their diversification trend through time.

Second, we replicated our analyses using the large subunit (LSU) rRNA gene. We downloaded the Glomeromycotina LSU database of Delavaux *et al.* (2020) as well as the LSU sequences available in MaarjAM. We obtained a total 2,044 sequences that we aligned using MAFFT and TrimAl. We retained the 1,760 unique haplotypes and reconstructed the phylogenetic tree of the LSU sequences using BEAST2 (same pipeline as above) and used the resulting calibrated tree to delineate Glomeromycotina LSU units with the GMYC model. Finally, we reconstructed the species-level phylogeny (still using BEAST2) and performed the diversification analyses (ClaDS and RPANDA) with a range of sampling fraction down to 50%.
Besides the SSU and the LSU rRNA genes, in fungi, the usual barcode is the ITS region, although the ITS data on AMF are currently less common (Lekberg *et al.*, 2018). However, we confirmed using the dataset of Lekberg *et al.* (2018) that the ITS sequences are very difficult to align, making them unsuitable for reconstructing a robust phylogeny for diversification analyses (Supplementary Figure 2).

Results:

AMF species delineations & phylogenetic reconstructions:

The EU97.5 and EU98 delineations (obtained using a threshold of 97.5% and 98% respectively) provided a number of AMF units (340 and 641) comparable to the 384 currently recognized VT, while the EU97 delineation had much less (182). Conversely, the EU98.5 and EU99 delineations yielded a much larger number of AMF units (1,190 and 2,647) that was consistent with the number obtained using GMYC analyses (Supplementary Tables 4, 5, & 6). This supports the idea that some VT might lump together several cryptic species (Bruns *et al.*, 2018; Supplementary Note 1), and that a 98.5 or 99% similarity threshold is more relevant for AMF species delineation. In addition, the GMYC analyses indicated that the level of genetic variation within the SSU marker is overall sufficient to separate AMF species-like units among SSU haplotypes (GMYC LRT: P<0.05; Supplementary Figure 3); on average, for one AMF unit delineated using GMYC, there are 10 SSU haplotypes with a mean intraspecific sequence similarity of 99% (Supplementary Table 6 & Supplementary Figure 3). Rarefaction curves as well as Bayesian and Chao2 estimates of diversity suggested that more than 90% of the total AMF diversity is represented in our dataset regardless of the delineation threshold (Figure II.5.1, Supplementary Tables 2, 6, & 7), which is consistent with the proportion of new AMF units detected in recent studies (Sepp *et al.*, 2019).

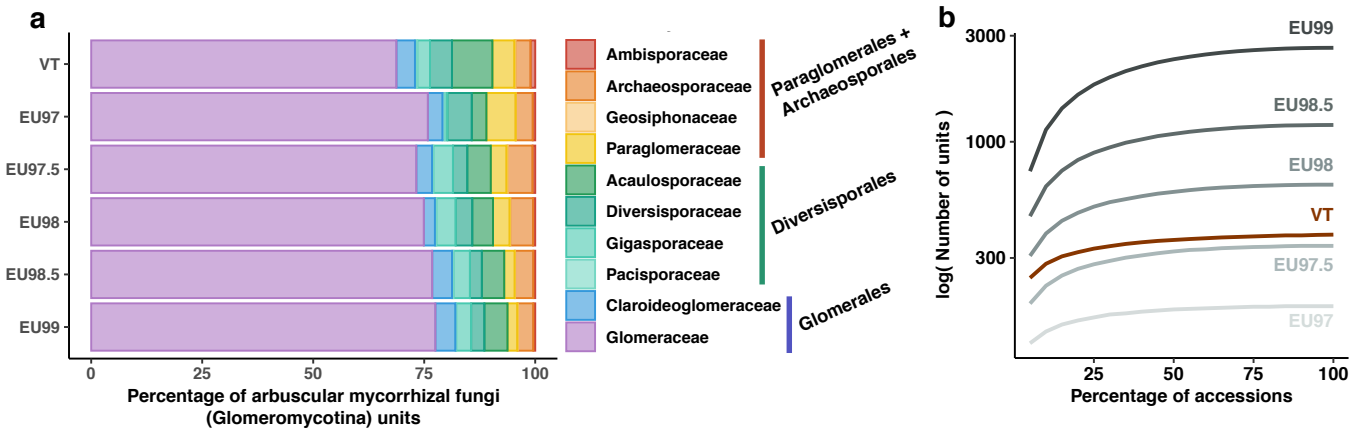


Figure II.5.1: Molecular-based species delineations of arbuscular mycorrhizal fungi (AMF) give consistent results and indicate a nearly complete sampling. We compared the virtual taxa (VT) delineation from (Öpik *et al.*, 2010) with newly-developed automatic delineations into evolutionary units (EUs) based on an average threshold of similarity and a criterion of monophyly. (a) The proportion of AMF units (VT or EUs) in each AMF family reveals constant proportions across delineations, although Glomeraceae tend to be relatively less abundant compared with the other AMF family in the VT delineation. The main AMF orders are indicated on the right of the charts: Paraglomerales + Archaeosporales, Diversisporales, and Glomerales (Glomeraceae + Claroideoglomeraceae). (b) Rarefaction curves indicating the number of AMF units as a function of the percentage of sampled AMF accession revealed that the AMF sampling in MaarjAM is close to saturation for all delineations (VT or EUs). Rarefactions were performed 100 times every 5 percent and the median of the 100 replicates is represented here.

The reconstructed Bayesian phylogenetic trees based on VT and EU delineations did not yield high support for the nodes separating the main AMF orders; yet, they had similar topologies and branching times of the internal nodes overall (Figure II.5.2, Supplementary Figure 4). As expected, finer delineations resulted in an increase in the number of nodes close to the present (Supplementary Figure 5). However, we observed a slowdown in the accumulation of new lineages close to the present in all lineage through time plots (LTTs), including those with the finest delineations (EU98.5 and EU99; Supplementary Figure 6).

Temporal diversification dynamics:

AMF speciation rates ranged from 0.005 to 0.03 events per lineage per Myr (Figure II.5.2; Supplementary Figure 7), and varied both within and among AMF orders, with Glomerales and Diversisporales having the highest present-day speciation rates (Supplementary Figure 8). As expected we observed higher present-day speciation rates for finer delineations, but at the level of the individuals we found a significant correlation of the speciation rates according to the different delineations (Supplementary Figure 9). Whatever the delineations, AMF experienced their most rapid diversification between 200 and 100 Myr ago according to estimates of diversification rates through time obtained with ClaDS (Figure II.5.2; Supplementary Figure 10), and 150-50 Myr ago according to diversification models with piecewise constant rates (TreePar and CoMET, Figure II.5.2; Supplementary Figures 11 & 12).

The fast diversification of AMF between 200 and 100 Myr ago was followed by a slowdown in the recent past (Figure II.5.2; Supplementary Figure 10), as suggested by the plateauing of the LTTs. A global decrease of the speciation rates through time was independently supported by ClaDS, TreePar, and CoMET analyses, as well as time-dependent models in RPANDA (Morlon *et al.*, 2011; Supplementary Figures 11, 12, 13 & 14). This slowdown was robust to all species delineations, the branching process prior (Supplementary Table 8), phylogenetic uncertainty, and sampling fractions down to 50%, except in ClaDS analyses where the trend disappeared in some EU99 trees and for sampling fractions lower than 70% (Supplementary Figures 15, 16 & 17). Finally, we still observed a significant decrease of the rates through time when not considering the last 50 Myr (Supplementary Figure 18).

We did not find a strong signal of extinction in our analyses: the turnover rate estimated from ClaDS was generally close to zero (Supplementary Figure 13b), and models including extinctions were never selected in RPANDA (Supplementary Figure 14). Similarly, the extinction rates estimated in piecewise-constant models were not significantly different from 0 (Supplementary Figure 19). Forcing the extinction rate to positive values did not modify the general trend of speciation rate slowdown (Supplementary Figures 20 & 21).

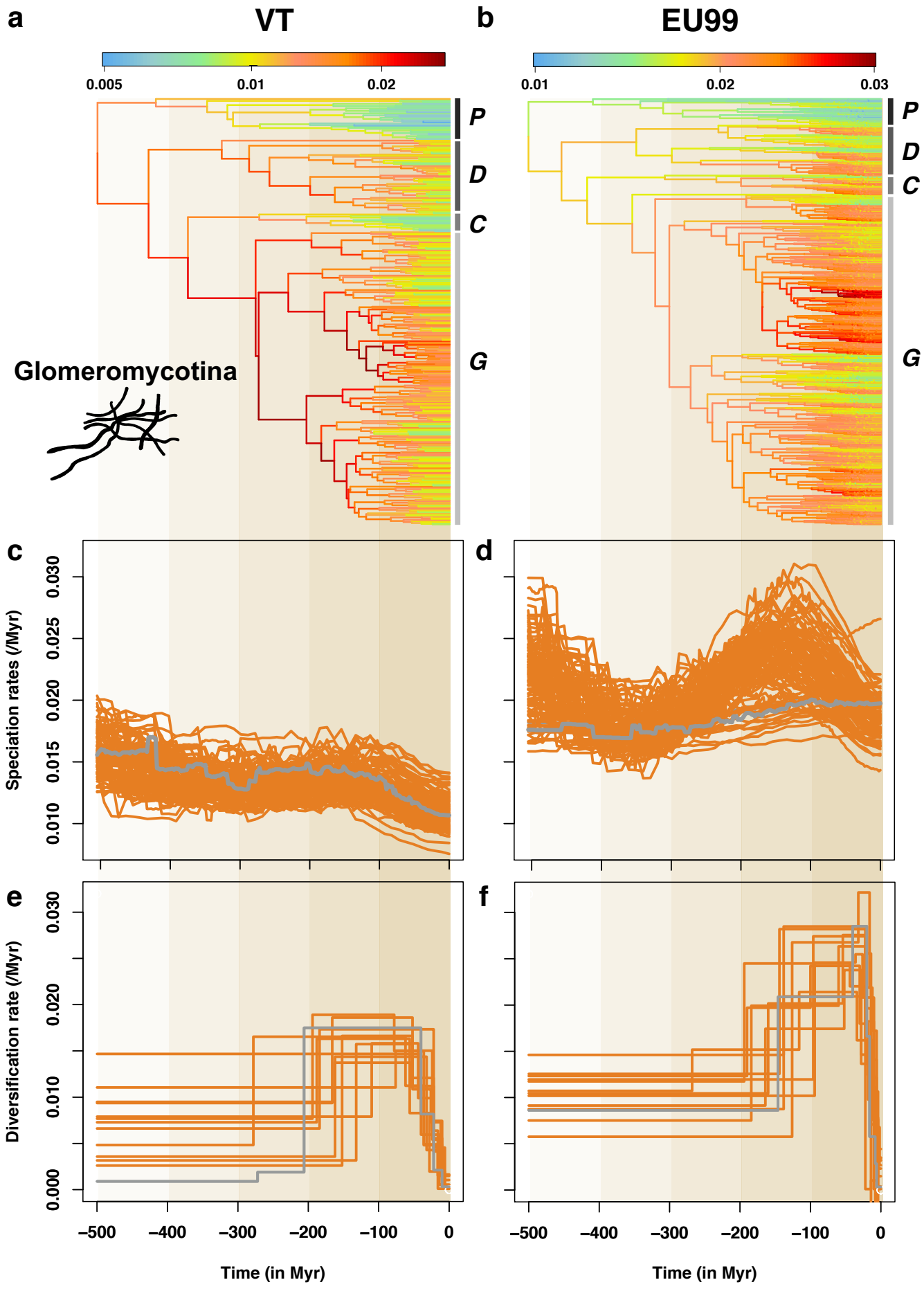

a
VT
0.005
0.01
0.02
P
D
C
G
Glomeromycotina
b
EU99
0.01
0.02
0.03
P
D
C
G
c
Speciation rates (/Myr)
0.010
0.015
0.020
0.025
0.030
d
e
Diversification rate (/Myr)
0.000
0.010
0.020
0.030
−500
−400
−300
−200
−100
0
Time (in Myr)
f
−500
−400
−300
−200
−100
0
Time (in Myr)

Figure II.5.2: The diversification dynamic of arbuscular mycorrhizal fungi (AMF) varies significantly through time and between lineages. (a-b) AMF consensus phylogenetic trees corresponding to the VT (a) and EU99 (b) species delineations. Branches are colored according to the lineage-specific speciation rates estimated by ClaDS using the BDES estimated sampling fraction: lineages with low and high speciation rates are represented in blue and red, respectively. The main AMF clades are indicated with the following letters: P = Paraglomerales + Archaeosporales, D = Diversisporales, C = Claroideoglomeraceae, and G = Glomeraceae.
(c-d) Mean speciation rates through time estimated by ClaDS, for the VT (c) and EU99 (d) delineations and using the BDES estimated sampling fraction. The mean speciation rate corresponds to the maximum *a posteriori* (MAP) of the mean speciation rate across all fungal lineages back in time (including extinct and unsampled lineages). Orange and grey lines represent the independent replicate trees and the consensus tree, respectively: because the 12 replicate trees showed different trends, we replicated ClaDS inferences using 100 replicate trees. Unlike most replicate trees, the EU99 consensus tree tends to present a limited diversification slowdown, which reinforces the idea that consensus trees can be a misleading representation (Janzen & Etienne, 2017).
(e-f): Net diversification rates (speciation rates minus extinction rates) through time estimated by TreePar, for the VT (c) and EU99 (d) delineations and using the BDES estimated sampling fraction. Orange and grey lines represent the 12 independent replicate trees and the consensus tree, respectively.

AMF diversification drivers:

When fitting environment-dependent models of diversification, we found high support for temperature-dependent models compared to time-dependent models for all AMF delineations, sampling fractions, and crown ages (Figure II.5.3; Supplementary Figures 22, 23, 24, 25, & 26), with the exception of some EU99 trees with a 50% sampling fraction (Supplementary Figure 26). This signal of temperature dependency was not due to a temporal trend (Supplementary Figures 27 & 28) nor to an artefact caused by rate heterogeneities (Supplementary Figure 29). Evidence for temperature dependency, however, decreased in some clades closer to the present, as small trees tend to be best fit by constant or time-depend models (Supplementary Figure 30). We detected a significant positive dependency of the diversification rates on CO_2 concentrations in some sub-trees, but rarely found a significant effect of plant fossil diversity (Supplementary Figure 30). Finally, this signal of temperature dependency was also observed in most of the trees when excluding the past 50 Myr (Supplementary Figure 18), suggesting that this signal is not artifactually driven the recent past.

The PCA of AMF niche width characteristics had a first principal component (PC1) that indicated the propensity of each AMF unit (VT or EUs) to be vastly distributed among continents, ecosystems and/or associated with many plant species and lineages, whereas the second principal component (PC2) indicated the propensity of a given AMF unit to associate with few plant species on many continents (Supplementary Figures 31, 32, & 33). Hence, PC1 reflects AMF niche width, whereas PC2 discriminates the width of the abiotic relatively to the biotic niche (Figure II.5.4a-b). We found a positive correlation between PC1 and lineage-specific speciation rates in the majority of the VT and

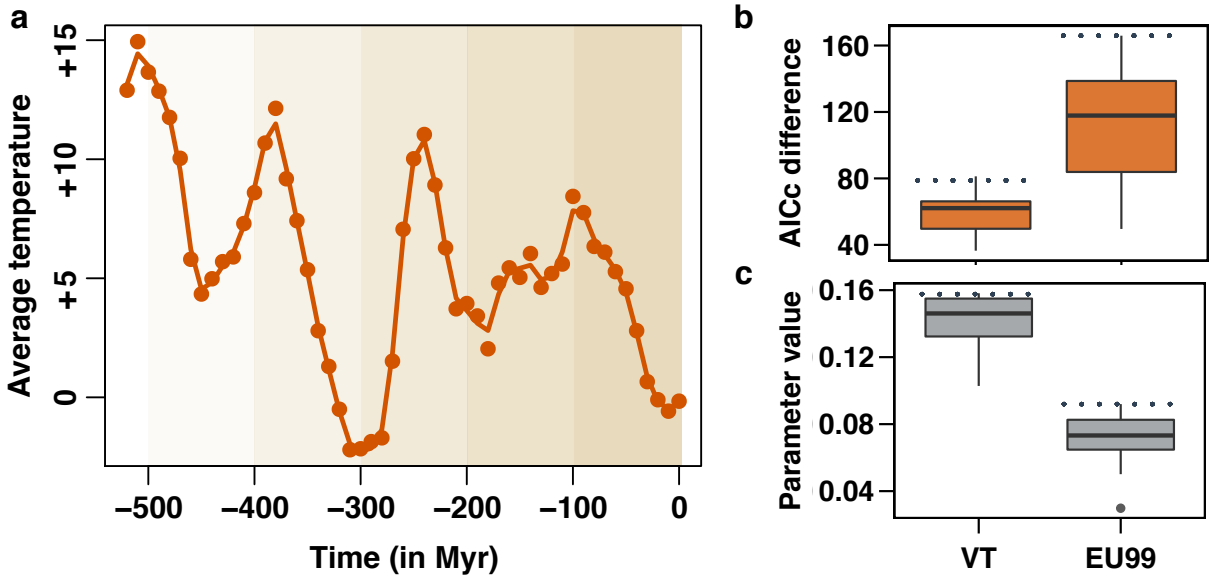


Figure II.5.3: Temperature-dependent diversification models reveal that global temperature positively associates with the speciation rates of arbuscular mycorrhizal fungi (AMF) in the last 500 million years. (a) Average global temperature in the last 500 million years (Myr) relative to the average temperature of the period 1960-1990. The smoothed orange line represents cubic splines with 33 degrees of freedom used to fit temperature-dependent models of AMF diversification with RPANDA. This default smoothing was estimated using the R function smooth.spline. (b) AICc difference between the best-supported time-dependent model and the temperature-dependent model in RPANDA, for the VT (left) and EU99 (right) delineations, using the BDES estimated sampling fraction. An AICc difference greater than 2 indicates that there is significant support for the temperature-dependent model. (c) Parameter estimations of the temperature-dependent models (speciation rate $\sim$ exp(parameter * temperature)). A positive parameter value indicates a positive effect of temperature on speciation rates. For both delineations, the boxplots represent the results obtained for the consensus tree and the 12 independent replicate trees. Boxplots indicate the median surrounded by the first and third quartiles, and whiskers extend to the extreme values but no further than 1.5 of the inter-quartile range. The horizontal dotted lines highlighted the values estimated for the consensus trees. Compared to the replicate trees, the consensus trees tend to present extreme values (stronger support for temperature-dependent model), which reinforces the idea that consensus trees can be a misleading representation (Janzen & Etienne, 2017).

EU99 trees (Figure II.5.4c-d; Supplementary Figure 34a). However, these results were no longer significant when controlling for phylogenetic non-independence between AMF units (Supplementary Figure 34b), likely because a single Glomeraceae clade, including the abundant and widespread morphospecies *Rhizophagus irregularis* and *R. clarus*, had both the highest speciation rates and the largest niche widths among AMF (Supplementary Figure 35).

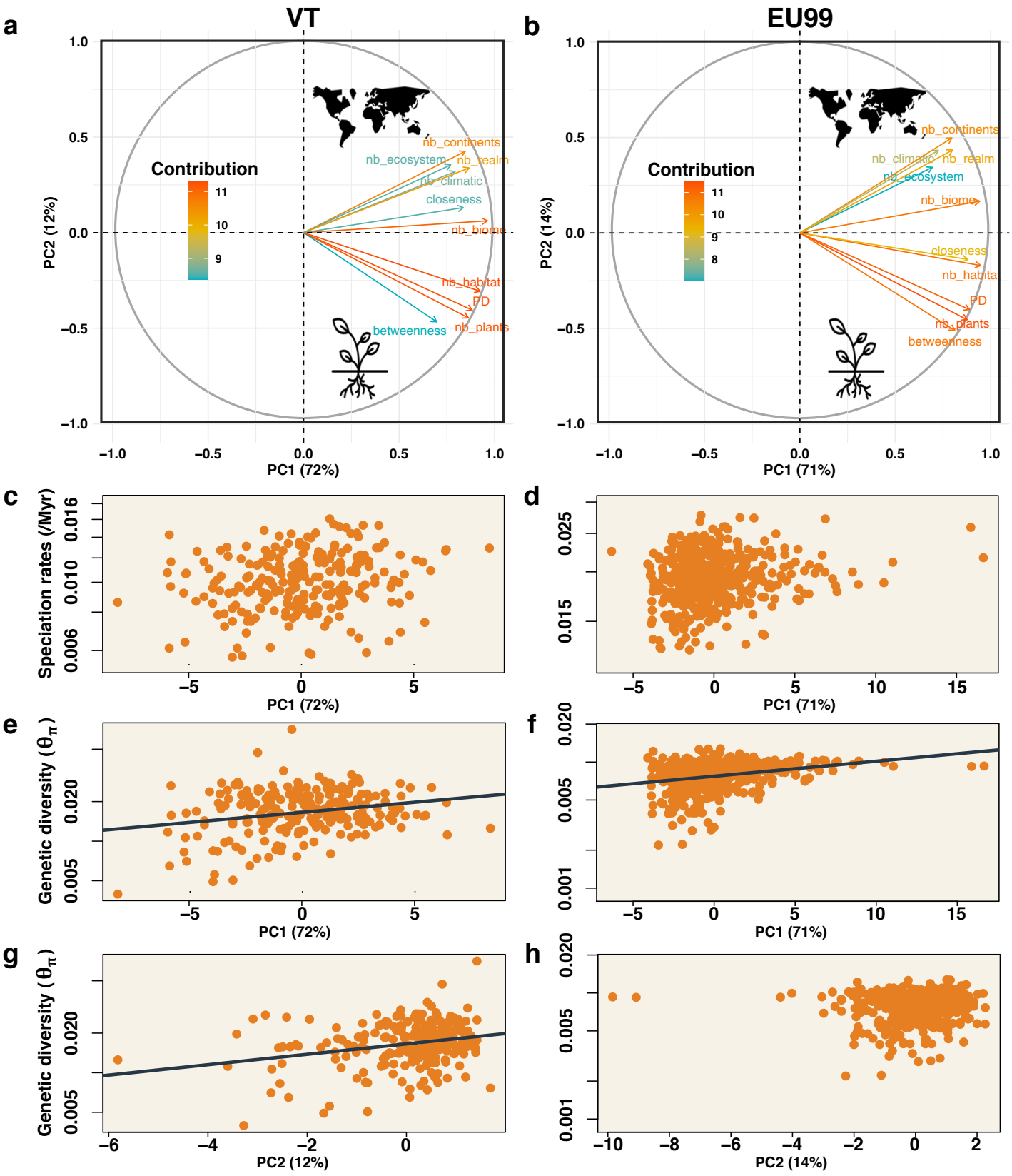

a
VT
Contribution
11
10
9
nb_continents
nb_ecosystem
nb_realm
nb_climatic
closeness
nb_biome
nb_habitat
PD
nb_plants
betweenness
PC2 (12%)
PC1 (72%)
b
EU99
Contribution
11
10
9
8
nb_continents
nb_climatic
nb_realm
nb_ecosystem
nb_biome
closeness
nb_habitat
PD
nb_plants
betweenness
PC2 (14%)
PC1 (71%)
c
Speciation rates (/Myr)
PC1 (72%)
d
PC1 (71%)
e
Genetic diversity (θπ)
PC1 (72%)
f
PC1 (71%)
g
Genetic diversity (θπ)
PC2 (12%)
h
PC2 (14%)

Figure II.5.4: Abiotic and biotic drivers of the species diversification and differentiation of arbuscular mycorrhizal fungi (AMF). (a-b) Projection of 10 abiotic and biotic variables on the two principal coordinates according to the VT (a) or EU99 (b) delineations. Principal coordinate analysis (PCA) was performed for the AMF units represented by at least 10 sequences. Colors represent the contribution of the variable to the principal coordinates. The percentage for each principal coordinate (PC) indicates its amount of explained variance. Tested variables were: the numbers of continents on which the AMF unit occurs (nb_continent), of realms (nb_realm), of ecosystems (nb_ecosystems), of habitats (nb_habitats), of biomes (nb_biomes), and climatic zones (nb_climatic) (Öpik *et al.*, 2010), as well as information about the associated plant species of each unit, such as the number of plant partners (nb_plants), the phylogenetic diversity of these plants (PD), and the betweenness and closeness measurement of each fungal unit in the plant-fungus interaction network (see Methods). (c-d) Speciation rates as a function of the PC1 coordinates for each VT (c) or EU99 (d) unit. Only the AMF consensus tree is represented here (other replicate trees are presented in Supplementary Figure 34). (e-h) Genetic diversity (Tajima's θ_π estimator) as a function of the PC1 (e-f) or PC2 (g-h) coordinates for each VT (e-g) or EU99 (f-h) unit. Only the AMF consensus tree is represented here (other replicate trees are presented in Supplementary Figure 34). The grey lines indicate the statistically significant linear regression between the two variables inferred using MCMCglmm.

Although the current AMF diversity is higher in the tropics (Supplementary Figure 36), we found no effect of latitude on speciation rates, regardless of the AMF delineation or the minimum number of sequences per AMF unit (MCMCglmm: P>0.05), and no effect of habitat or climatic zone either (Supplementary Figure 37). Similarly, we recovered no significant correlation between spore size and speciation rate (Supplementary Figure 38), nor between spore size and level of endemism (Supplementary Figure 39).

Finally, Tajima's estimator of AMF genetic diversity was significantly and positively correlated with niche width (PC1) for all AMF delineations and minimal number of sequences per AMF unit considered, and in particular with abiotic aspects of the niche (PC2) in many cases (Figure II.5.4e-h; Supplementary Figure 34). Genetic diversity was not correlated with speciation rate (Supplementary Figure 34), latitude, habitat, climatic zone (MCMCglmm: P>0.05), or spore size (PGLS: P>0.05).

Assessing the robustness of our findings:

When simulating the evolution of a short DNA gene and using it to infer its diversification, we were overall able to recover the simulated scenarios (Supplementary Figure 40): for clades simulated with constant speciation and explication rates, we mostly inferred the correct simulated constant rates, whereas when simulating declining speciation rates, we either estimated constant rates (for small clades) or decreasing speciation rates (Supplementary Figure 40). Despite the fact that we simulated slowly evolving DNA sequences and that many extant species had identical haplotypes (Supplementary Figure 1b), we did not observe an artefactual diversification slowdown in the recent past (Supplementary Figure 40). Using RPANDA, we mainly found a support for constant speciation rates with no extinction and did not tend to select temperature-dependent

models (Supplementary Figure 41).

Finally, when replicating our analyses on the LSU rRNA gene, we delineated 181 GMYC units. We also reported a period of high speciation rates between 200 and 100 Myr ago that was followed by a diversification slowdown toward the present, for all sampling fraction down to 50% (Supplementary Figure 42). For a sampling fraction of 50% or below, we nevertheless observed a plateau of the speciation rates in the last 50 Myr (Supplementary Figure 42). Using RPANDA, we also found a strong and significant signal of temperature dependency (Supplementary Figure 43), meaning that both SSU and LSU rRNA genes lead to the same conclusions.

Discussion:

AMF species delineations, diversity, and phylogeny:

Species delineations are difficult to apply in AMF, which are poorly differentiated morphologically and mainly characterized by environmental sequences (Bruns *et al.*, 2018). In addition, their reproduction mode is not well known and they have unique nuclear dynamics in their spores and hyphae (Kokkoris *et al.*, 2020). Our GMYC analyses suggest that biologically relevant AMF species-like units correspond to SSU rRNA haplotypes with a sequence similarity between 98.5 and 99%. With this criterion of species delineation, we estimate that there are between 1,300 and 2,900 AMF 'species'. These estimates are largely above the number of currently described morphospecies or VT (Supplementary Note 1) but remain low in comparison with other fungal groups, like the Agaricomycetes that include taxa forming ectomycorrhiza (Varga *et al.*, 2019).

Species delineations and phylogenies constructed from a single gene and short sequences are limited, but in the current state of data acquisition, relatively short metabarcoding sequences provide for most microbial groups, including AMF, the only current possibility to analyze their diversification dynamics (Davison *et al.*, 2015; Lewitus *et al.*, 2018; Louca *et al.*, 2018). Here, our phylogenies did not resolve the branching of the AMF orders, with node supports similar to those of previous studies (Krüger *et al.*, 2012; Davison *et al.*, 2015; Rimington *et al.*, 2018; Supplementary Note 2), confirming that additional genomic evidence is required to reach consensus. We considered this uncertainty in the phylogenetic reconstruction by repeating our analyses on a set of trees spanning the likely tree space. We hope that our study based on the SSU (or LSU) rRNA region alone will foster efforts to obtain more genetic data, including additional genomic information, with the aim of reconstructing better supported, comprehensive phylogenies.

AMF diversify slowly:

We found speciation rates for AMF an order of magnitude lower than rates typically found for macro-eukaryotes (Maliet *et al.*, 2019; Upham *et al.*, 2019), like plants (Zanne

et al., 2014), or Agaricomycetes (Varga *et al.*, 2019). Low speciation rates in AMF may be linked to their particular reproduction (Yildirir *et al.*, 2020), to their occasional long-distance dispersal that homogenizes populations globally over evolutionary timescales (Savary *et al.*, 2018), or to the fact that they are generalist obligate symbionts (Morlon *et al.*, 2012). Regardless of the proximal cause, and contrary to Agaricomycetes for example, which present a large diversity of species, morphologies, and ecologies, the niche space exploited by AMF is limited to plant roots and the surrounding soil because of their obligate dependence on plants for more than 400 Myr (Tisserant *et al.*, 2013; Rich *et al.*, 2017). Thus, although AMF species delineation based on the SSU rRNA gene can be a poor predictor of their functional diversity, our analyses based on this gene has revealed that AMF, despite their ubiquity, have poorly diversified in the last 500 Myr compared with other groups.

We found little evidence for species extinction in AMF, including at mass extinction events. Although AMF are relatively widespread and generalists, and low extinction rates have been predicted before based on their ecology (Morton, 1990), these low extinction rate estimates could also come from the difficulty of estimating extinction from molecular phylogenies (Rabosky, 2016), one of the limitations of phylogeny-based diversification analyses (Supplementary Note 3).

AMF diversification through time:

The observed peak of AMF diversification detected between 200 and 100 Myr (or 150-50 Myr depending on the models) was mainly linked to the fast diversification of the largest family Glomeraceae. This peak was concomitant with the radiation of flowering plants (Sauquet & Magallón, 2018), but also with a major continental reconfiguration, including the breakdown of Pangea and the formation of climatically contrasted landmasses (Davison *et al.*, 2015). This period was also characterized by a warm climate potentially favorable to AMF diversification, such that disentangling the impact of these various factors on AMF diversification is not straightforward. Interestingly, a peak of diversification at this period was also found in the Agaricomycetes forming ectomycorrhiza (Varga *et al.*, 2019).

This peak of diversification has been followed by a slowdown. Signals of diversification slowdowns sometimes result from methodological artifacts, including incorrect species delineation, biased phylogenetic reconstruction, and under-sampling (Moen & Morlon, 2014). We carefully considered uncertainty in species delineation, phylogenetic reconstruction, and under-sampling down to 50%. In addition, our GMYC analyses confirmed that the SSU rRNA gene evolves fast enough to delineate AMF species-like units; although some cryptic AMF species can have the same SSU sequence (Krüger *et al.*, 2012), our analyses support the overall existence of several SSU haplotypes per AMF unit. We found that the observed diversification slowdown was robust to all these potential artifacts, amplified under scenarios of high extinction, and also present when using the

LSU rRNA gene. In addition, we showed using simulations that the fact we were using a single slowly-evolving DNA gene is unlikely to artefactually generate such a strong slowdown. Slowdowns in diversification rates close to the present have often been interpreted as a progressive reduction of the number of available niches as species diversify and accumulate (Rabosky, 2009; Moen & Morlon, 2014). In AMF, this potential effect of niche saturation could be exacerbated by a reduction of their niches linked to both repetitive breakdowns of their symbiosis with plants and climatic changes. Indeed, since the Cretaceous, many plant lineages evolved alternative root symbioses or became non-symbiotic (Selosse & Le Tacon, 1998; Maherali *et al.*, 2016; Werner *et al.*, 2018; Brundrett & Tedersoo, 2018): approximately 20% of extant plants do not interact with AMF anymore (van der Heijden *et al.*, 2015). Additionally, the cooling of the Earth during the Cenozoic reduced the surface of tropical regions (Ziegler *et al.*, 2003; Meseguer & Condamine, 2020), which tend to be a reservoir of ecological niches for AMF (Read, 1991; Davison *et al.*, 2015; Brundrett & Tedersoo, 2018).

The difficulty of reconstructing past symbiotic associations prevents direct testing of the hypothesis that the emergence of new root symbioses in plants led to a diversification slowdown in AMF. However, we tested the hypothesis that global temperature changes affected diversification rates and found a strong relationship. Such associations between temperature and diversification rates have been observed before in eukaryotes and have several potential causes (Condamine *et al.*, 2019). Two prevailing hypotheses are the evolutionary speed hypothesis, stipulating that high temperatures entail higher mutation rates and faster speciation (Rohde, 1992), and the productivity hypothesis, stating that resources and associated ecological niches are more numerous in warm and productive environments, especially when the tropics are large (Clarke & Gaston, 2006). The latter hypothesis is particularly relevant for AMF, which have many host plant niches in the tropics and potentially less in temperate regions (Toussaint *et al.*, 2020), where a higher proportion of plants are non-mycorrhizal (Bueno *et al.*, 2017) or ectomycorrhizal (Brundrett & Tedersoo, 2018; Varga *et al.*, 2019). Hence, the observed effect of past global temperatures could reflect the shrinkage of tropical areas and the associated decrease of the relative proportion of arbuscular mycorrhizal plants.

A few AMF clades displayed a significant support for diversification models with a positive dependency on CO_2 concentrations, which reinforces the idea that for the corresponding AMF benefits retrieved from plants could have been amplified by high CO_2 concentrations and fostered diversification (Humphreys *et al.*, 2010; Field *et al.*, 2016). Conversely, we found a limited effect of land plant fossil diversity, which indicates that variations in the tempo of AMF diversification did not systematically follow those of land plants. Still, the possible concordance of the peak of AMF diversification with the radiation of the Angiosperms is noteworthy, in particular in Glomeraceae that frequently interact with present-day Angiosperms (Rimington *et al.*, 2018). The co-diversification with the plants might have been an important driver from the emergence of land plants until

the Mesozoic (Morton, 1990; Lutzoni *et al.*, 2018), but less so thereafter, when AMF diversification declined while some flowering plants radiated, including AMF-free groups such as the species-rich Orchidaceae, blurring co-diversification patterns (Supplementary Figure 44; Cleal & Cascales-Miñana, 2014; Ramírez-Barahona *et al.*, 2020).

AMF recent diversification:

Looking at the correlates of AMF present-day diversification rates, we found no effect of habitat or climatic zone, even though AMF are more frequent and diverse in the tropics (Davison *et al.*, 2015; Pärtel *et al.*, 2017; Toussaint *et al.*, 2020) and their speciation rates are positively correlated with global temperature. Further work, including a more thorough sampling of the distribution of AMF species across latitudes and habitats, would be required to confirm these patterns and to distinguish whether speciation events are indeed no more frequent in the tropics or, if they are, whether long-distance dispersal redistributes the new lineages at different latitudes over evolutionary time scales (Pärtel *et al.*, 2017). Similarly, although the temporal changes in the availability of AMF niches likely influenced the diversification of the group, we found little support for AMF species with larger niche width having higher lineage-specific speciation rates. We also note that there are important aspects of the niche that we do not (and yet cannot) account for in our characterization of AMF niche width: it is thought that some AMF species may mainly provide mineral nutrients extracted from the soil, whereas others may be more specialized in protecting plants from biotic or abiotic stresses (Chagnon *et al.*, 2013) and such (inter- or intra-specific) functional variations may have evolutionary significance. Finally, although spore size is often inversely related to dispersal capacity (Nathan *et al.*, 2008), which can either promote diversification by favoring founder speciation events, or limit diversification by increasing gene flow, we found no significant correlation between spore size and diversification rates, which may be explained either by a weak or absent effect or by the low number of species for which this data is available. In addition, the absence of correlation between spore size and level of endemism suggests that even AMF with large spores experience long-distance dispersal (Davison *et al.*, 2018; Kivlin, 2020). Thus, if large spores might limit dispersal at smaller (*e.g.* intra-continental) scales in AMF (Bueno & Moora, 2019; Chaudhary *et al.*, 2020), this does not seem to affect diversification.

In AMF, intraspecific variability is an important source of functional diversity (Munkvold *et al.*, 2004; Savary *et al.*, 2018) and their genetic diversity may indicate the intraspecific variability on which selection can act, potentially leading to species diversification. Here, geographically widespread AMF species appear to be more genetically diverse, as previously suggested by population genomics (Savary *et al.*, 2018), but do not necessarily speciate faster. Along with a decoupling between genetic diversity and lineage-specific speciation rate, this suggests that the accumulation of genetic diversity among distant subpopulations is not enough to spur AMF speciation.

Conclusion:

Our findings that AMF have low speciation rates, likely constrained by the availability of suitable niches, reinforce the vision of AMF as an "evolutionary cul-de-sac" (Malloch, 1987). We interpret the significant diversification slowdown toward the present as the conjunction of the emergence of plant lineages not associated with AMF and the reduction of tropical areas induced by climate cooling, in the context of obligate dependence of AMF on plants. Diversification slowdowns have often been interpreted as the signal of adaptive radiations (Harmon *et al.*, 2003; Moen & Morlon, 2014), that is clades that experienced a rapid accumulation of morphological, ecological, and species diversity (Simpson, 1953). Conversely, AMF provide here a striking example of a clade with slow morphological, ecological, and species diversification that features a pattern of diversification slowdown, likely reflecting the reduction of the global availability of their mycorrhizal niches.

References:

Alroy J. 2010. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology 53: 1211–1235.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis (A Prlic, Ed.). PLoS Computational Biology 10: e1003537.

Bredenkamp GJ, Spada F, Kazmierczak E. 2002. On the origin of northern and southern hemisphere grasslands. Plant Ecology 163: 209–229.

Brundrett MC, Tedersoo L. 2018. Evolutionary history of mycorrhizal symbioses and global host plant diversity. New Phytologist 220: 1108–1115.

Bruns TD, Corradi N, Redecker D, Taylor JW, Öpik M. 2018. Glomeromycotina: what is a species and why should we care? New Phytologist 220: 963–967.

Bueno CG, Moora M. 2019. How do arbuscular mycorrhizal fungi travel? New Phytologist 222: 645–647.

Bueno CG, Moora M, Gerz M, Davison J, Öpik M, Pärtel M, Helm A, Ronk A, Kühn I, Zobel M. 2017. Plant mycorrhizal status, but not type, shifts with latitude and elevation in Europe. Global Ecology and Biogeography 26: 690–699.

Chagnon P-L, Bradley RL, Maherali H, Klironomos JN. 2013. A trait-based framework to understand life history of mycorrhizal fungi. Trends in Plant Science 18: 484–491.

Chaudhary VB, Nolimal S, Sosa-Hernández MA, Egan C, Kastens J. 2020. Trait-based aerial dispersal of arbuscular mycorrhizal fungi. New Phytologist 228: 238–252.

Clarke A, Gaston KJ. 2006. Climate, energy and diversity. Proceedings of the Royal Society B: Biological Sciences 273: 2257–2266.

Cleal CJ, Cascales-Miñana B. 2014. Composition and dynamics of the great Phanerozoic Evolutionary Floras. Lethaia 47: 469–484.

Condamine FL, Rolland J, Morlon H. 2013. Macroevolutionary perspectives to environmental change (H Maherali, Ed.). Ecology Letters 16: 72–85.

Condamine FL, Rolland J, Morlon H. 2019. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support (R Etienne, Ed.). Ecology Letters 22: 1900–1912.

Correia M, Heleno R, da Silva LP, Costa JM, Rodríguez-Echeverría S. 2019. First evidence for the joint dispersal of mycorrhizal fungi and plant diaspores by birds. New Phytologist 222: 1054–1060.

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, *et al.* 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349: 970–973.

Davison J, Moora M, Öpik M, Ainsaar L, Ducousso M, Hiiesalu I, Jairus T, Johnson N, Jourand P, Kalamees R, *et al.* 2018. Microbial island biogeography: isolation shapes the life history characteristics but not diversity of root-symbiotic fungal communities. The ISME Journal 12: 2211–2224.

Delavaux CS, Sturmer SL, Wagner MR, Schütte U, Morton JB, Bever JD. 2020. Utility of large subunit for environmental sequencing of arbuscular mycorrhizal fungi: a new reference database and pipeline. New Phytologist: 1–5.

Egan C, Li D-W, Klironomos J. 2014. Detection of arbuscular mycorrhizal fungal spores in the air across different biomes and ecoregions. Fungal Ecology 12: 26–31.

Ezard T, Fujisawa T, Barraclough TG. 2009. SPLITS: SPecies' LImits by Threshold Statistics. : R-package.

Feijen FA, Vos RA, Nuytinck J, Merckx VSFT. 2018. Evolutionary dynamics of mycorrhizal symbiosis in land plant diversification. Scientific Reports 8: 10698.

Field KJ, Pressel S, Duckett JG, Rimington WR, Bidartondo MI. 2015. Symbiotic options for the conquest of land. Trends in Ecology & Evolution 30: 477–486.

Field KJ, Rimington WR, Bidartondo MI, Allinson KE, Beerling DJ, Cameron DD, Duckett JG, Leake JR, Pressel S. 2016. Functional analysis of liverworts in dual symbiosis with Glomeromycota and Mucoromycotina fungi under a simulated Palaeozoic CO_2 decline. ISME Journal 10: 1514–1526.

Foster GL, Royer DL, Lunt DJ. 2017. Future climate forcing potentially without precedent in the last 420 million years. Nature Communications 8: 14845.

Fujisawa T, Barraclough TG. 2013. Delimiting species using single-locus data and the generalized mixed yule coalescent approach: A revised method and evaluation on simulated data sets. Systematic Biology 62: 707–724.

Gelman A, Rubin DB. 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7: 457–472.

Hadfield JD. 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R-package. Journal of Statistical Software 33: 1–22.

Harmon LJ, Schulte JA, Larson A, Losos JB. 2003. Tempo and mode of evolutionary radiation in iguanian lizards. Science 301: 961–964.

van der Heijden MGA, Martin FM, Selosse MA, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytologist 205: 1406–1423.

Höhna S, May MR, Moore BR. 2016. TESS: An R-package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. Bioinformatics 32: 789–791.

Humphreys CP, Franks PJ, Rees M, Bidartondo MI, Leake JR, Beerling DJ. 2010. Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. Nature Communications 1: 103.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, *et al.* 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. Nature 443: 818–822.

Janzen T, Etienne RS. 2017. Inferring the role of habitat dynamics in driving diversification: evidence for a species pump in Lake Tanganyika cichlids. bioRxiv 11: 1–18.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Kivlin SN. 2020. Global mycorrhizal fungal range sizes vary within and among mycorrhizal guilds but are not correlated with dispersal traits. Journal of Biogeography 47: 1994–2001.

Kokkoris V, Stefani F, Dalpé Y, Dettman J, Corradi N. 2020. Nuclear dynamics in the arbuscular mycorrhizal fungi. Trends in Plant Science 25: 765–778.

Krüger M, Krüger C, Walker C, Stockinger H, Schüßler A. 2012. Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. New Phytologist 193: 970–984.

Lee J, Lee S, Young JPW. 2008. Improved PCR primers for the detection and identification of arbuscular mycorrhizal fungi. FEMS Microbiology Ecology 65: 339–349.

Lekberg Y, Vasar M, Bullington LS, Sepp S-KK, Antunes PM, Bunn R, Larkin BG, Öpik M. 2018. More bang for the buck? Can arbuscular mycorrhizal fungal communities be characterized adequately alongside other fungi using general fungal primers? New Phytologist 220: 971–976.

Lewitus E, Bittner L, Malviya S, Bowler C, Morlon H. 2018. Clade-specific diversification dynamics of marine diatoms since the Jurassic. Nature Ecology and Evolution 2: 1715–1723.

Louca S, Shih PM, Pennell MW, Fischer WW, Parfrey LW, Doebeli M. 2018. Bacterial diversification through geological time. Nature Ecology and Evolution 2: 1458–1467.

Lutzoni F, Nowak MD, Alfaro ME, Reeb V, Miadlikowska J, Krug M, Arnold AE, Lewis LA, Swofford DL, Hibbett D, *et al.* 2018. Contemporaneous radiations of fungi and plants linked to symbiosis. Nature Communications 9: 1–11.

Maherali H, Oberle B, Stevens PF, Cornwell WK, McGlinn DJ. 2016. Mutualism persistence and abandonment during the evolution of the mycorrhizal symbiosis. American Naturalist 188: E113–E125.

Maliet O, Hartig F, Morlon H. 2019. A model with many small shifts for estimating species-specific diversification rates. Nature Ecology & Evolution 3: 1086–1092.

Maliet O, Morlon H. 2020. Fast and accurate estimation of species-specific diversification rates using data augmentation. bioRxiv: 2020.11.03.365155.

Malloch DM. 1987. The evolution of mycorrhizae. Can. J. Plant. Path. 9: 398–402.

May MR, Höhna S, Moore BR. 2016. A Bayesian approach for detecting the impact of mass-extinction events on molecular phylogenies when rates of lineage diversification may vary (N Cooper, Ed.). Methods in Ecology and Evolution 7: 947–959.

Meseguer AS, Condamine FL. 2020. Ancient tropical extinctions at high latitudes contributed to the latitudinal diversity gradient. Evolution 74: 1966–1987.

Moen D, Morlon H. 2014. Why does diversification slow down? Trends in Ecology and Evolution 29: 190–197.

Morlon H, Kemps BD, Plotkin JB, Brisson D. 2012. Explosive radiation of a bacterial species group. Evolution 66: 2577–2586.

Morlon H, Lewitus E, Condamine FL, Manceau M, Clavel J, Drury J. 2016. RPANDA: An R-package for macroevolutionary analyses on phylogenetic trees. Methods in Ecology and Evolution 7: 589–597.

Morlon H, Parsons TL, Plotkin JB. 2011. Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences 108: 16327–16332.

Morton JB. 1990. Species and clones of arbuscular mycorrhizal fungi (Glomales, Zygomycetes): their role in macro- and microevolutionary processes. Mycotaxon (USA) 37: 493–515.

Munkvold L, Kjøller R, Vestberg M, Rosendahl S, Jakobsen I. 2004. High functional diversity within species of arbuscular mycorrhizal fungi. New Phytologist 164: 357–364.

Nathan R, Schurr FM, Spiegel O, Steinitz O, Trakhtenbrot A, Tsoar A. 2008. Mechanisms of long-distance seed dispersal. Trends in Ecology & Evolution 23: 638–647.

Öpik M, Davison J, Moora M, Zobel M. 2014. DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. Botany 92: 135–147.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist 188: 223–241.

Pärtel M, Öpik M, Moora M, Tedersoo L, Szava-Kovats R, Rosendahl S, Rillig MC, Lekberg Y, Kreft H, Helgason T, *et al.* 2017. Historical biome distribution and recent human disturbance

shape the diversity of arbuscular mycorrhizal fungi. New Phytologist 216: 227–238.

Perez-Lamarque B, Morlon H. 2019. Characterizing symbiont inheritance during host-microbiota evolution: Application to the great apes gut microbiota. Molecular Ecology Resources 19: 1659–1671.

Perez-Lamarque B, Selosse MA, Öpik M, Morlon H, Martos F. 2020. Cheating in arbuscular mycorrhizal mutualism: a network and phylogenetic analysis of mycoheterotrophy. New Phytologist 226: 1822–1835.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects (M Hedin, Ed.). Systematic Biology 55: 595–609.

Powell JR, Monaghan MT, Öpik M, Rillig MC. 2011. Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. Molecular Ecology 20: 655–666.

Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. ISME Journal 2: 997–1006.

R Core Team. 2020. R: A language and environment for statistical computing.

Rabosky DL. 2009. Ecological limits and diversification rate: Alternative paradigms to explain the variation in species richness among clades and regions. Ecology Letters 12: 735–743.

Rabosky DL. 2016. Challenges in the estimation of extinction from molecular phylogenies: A response to Beaulieu and O'Meara. Evolution 70: 218–228.

Ramírez-Barahona S, Sauquet H, Magallón S. 2020. The delayed and geographically heterogeneous diversification of flowering plant families. Nature Ecology and Evolution 4: 1232–1238.

Read DJ. 1991. Mycorrhizas in ecosystems. Experientia 47: 376–391.

Rich MK, Nouri E, Courty P-E, Reinhardt D. 2017. Diet of arbuscular mycorrhizal fungi: Bread and butter? Trends in Plant Science 22: 652–660.

Rimington WR, Pressel S, Duckett JG, Field KJ, Read DJ, Bidartondo MI. 2018. Ancient plants with ancient fungi: liverworts associate with early-diverging arbuscular mycorrhizal fungi. Proceedings of the Royal Society B: Biological Sciences 285: 20181600.

Rohde K. 1992. Latitudinal gradients in species diversity: The search for the primary cause. Oikos 65: 514.

Royer DL, Berner RA, Montañez IP, Tabor NJ, Beerling DJ. 2004. CO_2 as a primary driver of Phanerozoic climate. GSA Today 14: 4.

Sanders IR. 2003. Preference, specificity and cheating in the arbuscular mycorrhizal symbiosis. Trends in Plant Science 8: 143–145.

Sauquet H, Magallón S. 2018. Key questions and challenges in angiosperm macroevolution. New Phytologist 219: 1170–1187.

Savary R, Masclaux FG, Wyss T, Droh G, Cruz Corella J, Machado AP, Morton JB, Sanders IR. 2018. A population genomics approach shows widespread geographical distribution of cryptic genomic forms of the symbiotic fungus *Rhizophagus irregularis*. ISME Journal 12: 17–30.

Selosse MA, Le Tacon F. 1998. The land flora: a phototroph-fungus partnership? Trends in Ecology & Evolution 13: 15–20.

Sepp SK, Davison J, Jairus T, Vasar M, Moora M, Zobel M, Öpik M. 2019. Non-random association patterns in a plant–mycorrhizal fungal network reveal host–symbiont specificity. Molecular Ecology 28: 365–378.

Simon L, Bousquet J, Lévesque RC, Lalonde M. 1993. Origin and diversification of endomycorrhizal fungi and coincidence with vascular land plants. Nature 363: 67–69.

Simon L, Lalonde M, Bruns TD. 1992. Specific amplification of 18S fungal ribosomal genes from vesicular-arbuscular endomycorrhizal fungi colonizing roots. Applied and environmental microbiology 58: 291–5.

Simpson GG. 1953. The major features of evolution. New York: Columbia University Press.

Smith SE, Read DJ. 2008. Mycorrhizal Symbiosis. Elsevier.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences 108: 6187–6192.

Strullu-Derrien C, Selosse MA, Kenrick P, Martin FM. 2018. The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. New Phytologist 220: 1012–1030.

Stürmer SL. 2012. A history of the taxonomy and systematics of arbuscular mycorrhizal fungi belonging to the phylum Glomeromycota. Mycorrhiza 22: 247–258.

Tajima F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437–460.

Templeton AR. 2008. The reality and importance of founder speciation in evolution. BioEssays 30: 470–479.

Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P, Duensing N, Frei Dit Frey N, Gianinazzi-Pearson V, *et al.* 2013. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proceedings of the National Academy of Sciences of the United States of America 110: 20117–20122.

Toussaint A, Bueno G, Davison J, Moora M, Tedersoo L, Zobel M, Öpik M, Pärtel M. 2020. Asymmetric patterns of global diversity among plants and mycorrhizal fungi (F Pugnaire, Ed.). Journal of Vegetation Science 31: 355–366.

Upham NS, Esselstyn JA, Jetz W. 2019. Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation (AJ Tanentzap, Ed.). PLoS Biology 17: e3000494.

Varga T, Krizsán K, Földi C, Dima B, Sánchez-García M, Sánchez-Ramírez S, Szöllősi GJ, Szarkándi JG, Papp V, Albert L, *et al.* 2019. Megaphylogeny resolves global patterns of mushroom evolution. Nature Ecology & Evolution 3: 668–678.

Venice F, Ghignone S, Salvioli di Fossalunga A, Amselem J, Novero M, Xianan X, Sędzielewska Toro K, Morin E, Lipzen A, Grigoriev I V., *et al.* 2020. At the nexus of three kingdoms: the genome of the mycorrhizal fungus *Gigaspora margarita* provides insights into plant, endobacterial and fungal interactions. Environmental Microbiology 22: 122–141.

Werner GDA, Cornelissen JHC, Cornwell WK, Soudzilovskaia NA, Kattge J, West SA, Kiers ET. 2018. Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown. Proceedings of the National Academy of Sciences 115: 5229–5234.

Werner GDA, Cornwell WK, Sprent JI, Kattge J, Kiers ET. 2014. A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. Nature Communications 5: 4087.

Yildirir G, Malar C M, Kokkoris V, Corradi N. 2020. Parasexual and sexual reproduction in arbuscular mycorrhizal fungi: Room for both. Trends in Microbiology 28: 517–519.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506: 89–92.

Ziegler AM, Eshel G, McAllister Rees P, Rothfus TA, Rowley DB, Sunderlin D. 2003. Tracing the tropics across land and sea: Permian to present. Lethaia 36: 227–254.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3mNB6nc` or by scanning:

Chapter III.

Analyzing the evolution of cheating in host-microbiota mutualisms:

Cheating can represent a threat to the stability of species-rich host-microbiota mutualism, which must therefore rely on mechanisms to constrain and limit cheaters. In this chapter, we investigated the evolution of cheating in the mycorrhizal symbioses. We focused on mycoheterotrophy, *i.e.* the achlorophyllous plants that rely on their mycorrhizal fungi to get both their mineral and organic matter. In Article 6, we first explored the constraints upon the evolutionary emergence of mycoheterotrophic cheating in the arbuscular mycorrhizal symbiosis at the global scale. We combined network and phylogenetic analyses and developed a framework to investigate the presence of constraints upon cheating. Using the MaarjAM database, we studied a global interaction network (>25,000 interactions) between land plants and arbuscular mycorrhizal fungi informed with the phylogenies of both plants and fungi. Unlike mutualistic autotrophic plants, cheating plants appeared narrowly specialized towards some closely-related specialist fungi. Thus, cheaters tend to be specifically isolated into modules and the different mycoheterotrophic lineages convergently interact with 'cheating-susceptible' fungal partners. These results raised new hypotheses about the mechanisms (*e.g.* sanctions and/or habitat filtering) that actually constraint the interaction of mycoheterotrophic plants and their associated fungi with the rest of the autotrophic plants. In addition, we found a strict reciprocal specialization in the initially mycoheterotrophic lycopods (Lycopodiaceae), which could suggest that parental nurture is happening between green sporophytes and achlorophyllous gametophytes.

Then, we investigated whether similar patterns were found in local mycorrhizal networks including initially mycoheterotrophic plants (Lycopodiaceae) that we sampled in La Réunion island (Article 7). We sampled mycorrhizal networks across three communities in La Réunion and investigate the fungal sharing between the different co-occurring plants. We characterized root-associated fungal communities using two pairs of primers amplifying the 18S SSU rRNA gene and the ITS2 respectively. We found that there is a lot of fungal sharing between plant lineages, including between lycopods and other plant

lineages (*e.g.* ferns or flowering plants). Although adult lycopods are well connected to their surrounding plants by mycorrhizal fungi (*i.e.* no reciprocal specialization), we also found lycopod-restricted fungi, which might ensure parental nurture between gametophytes and sporophytes. Even if the initial focus of the project was the lycopods, the following Article 7 is exploring a more general question not directly related to the emergence of cheaters. Indeed, we investigated fungal sharing between the different distantly related plant lineages (and not only between lycopods and other groups) and looked at the structures of the resulting plant-fungus networks for the main fungal lineages (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, and Cantharellales). We found striking differences between the different fungal lineages in terms of specialization with plants and network structures, which highlights the ecological and evolutionary distinctiveness of these different mycorrhizal fungal lineages.

Contents of Chapter III

Chapitre III : Analyser l'évolution de la tricherie dans les mutualismes hôtes-microbiotes

La tricherie peut représenter une menace envers la stabilité des mutualismes hôtes-microbiotes riches en espèces, qui doivent donc reposer sur des mécanismes afin de contraindre et limiter les tricheurs. Dans ce chapitre, nous avons étudié l'évolution de la tricherie au sein des symbioses mycorhizennes. Nous nous sommes intéressés à la mycohétérotrophie, *i.e.* aux plantes non-chlorophylliennes qui reposent sur les champignons mycorhiziens pour obtenir à la fois leur matière minérale et organique. Dans l'Article 6, nous avons tout d'abord exploré les contraintes limitant l'émergence de tricheurs mycohétérotrophes dans la symbiose mycorhizenne arbusculaire. Nous avons combiné des analyses de réseaux et des analyses phylogénétiques pour déterminer la présence de contraintes envers la tricherie. Grâce à la base de données MaarjAM, nous avons étudié un réseau d'interactions à l'échelle mondiale, contenant plus de 25 000 interactions, entre plantes terrestres et champignons endomycorhiziens à arbuscules. Contrairement aux plantes autotrophes mutualistes, les plantes mycohétérotrophes tricheuses apparaissent étroitement spécialisées envers des champignons phylogénétiquement proches et eux-mêmes spécialistes. Ainsi, les tricheurs tendent à être spécifiquement isolés au sein de modules et les différentes lignées de mycohétérotrophes interagissent de façon convergente avec les mêmes partenaires fongiques « susceptible à la tricherie ». Ces résultats amènent à de nouvelles hypothèses quant aux mécanismes (sanction et/ou filtre d'habitat) qui limitent les interactions entre les tricheurs et leurs partenaires et le reste des plantes autotrophes. De plus, nous avons trouvé une spécialisation réciproque stricte chez certaines espèces initialement mycohétérotrophes de lycopodes (Lycopodiaceae), ce qui pourrait suggérer que des échanges nutritionnels auraient lieu entre les sporophytes adultes autotrophes et les gamétophytes mycohétérotrophes.

Pour finir, nous avons examiné si des patrons similaires étaient présents dans des réseaux mycorhiziens à l'échelle locale qui incluent des plantes initialement mycohétérotrophes (Lycopodiaceae) que nous avons échantillonnés sur l'île de la Réunion (Article 7). Nous avons étudié les réseaux mycorhiziens dans trois communautés de la Réunion et examiné le partage de champignons entre les différentes plantes qui co-occurrent. Nous avons caractérisé les communautés fongiques associées aux racines des plantes grâce à deux paires d'amorces amplifiant respectivement le gène de l'ARNr 18S et la région ITS2. Nous avons trouvé beaucoup de partage de champignons entre lignées de plantes, y compris entre les lycopodes et les plantes à fleurs ou fougères. Bien que les lycopodes adultes soient vraisemblablement connectés aux plantes voisines par des champignons mycorhiziens (*i.e.* pas de spécialisation réciproque), nous avons aussi trouvé des champignons spécifiques aux lycopodes, lesquels pourraient éventuellement mettre directement en réseau les sporophytes adultes autotrophes et les gamétophytes mycohétérotrophes. Même si la question des lycopodes est centrale à notre projet, l'Article 7 suivant est présenté de manière à aborder une question plus générale, qui n'est pas directement liée à la question des tricheurs. En effet, nous avons étudié le partage fongique entre différentes lignées de plantes phylogénétiquement éloignées et regardé les structures des réseaux plantes-champignons qui en résultent, pour les principales lignées de champignons présents (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, et Cantharellales). Nous avons trouvé des disparités marquantes entre les différentes lignées fongiques en termes de spécialisations de leurs interactions avec les plantes et de structure des réseaux, ce qui souligne les spécificités écologiques et évolutives de ces différentes lignées mycorhizennes.

Article 6: Cheating in arbuscular mycorrhizal mutualism: a network and phylogenetic analysis of mycoheterotrophy

Authors: Benoît Perez-Lamarque[1,2], Marc-André Selosse[1,3], Maarja Öpik[4], Hélène Morlon[2], Florent Martos[1]

[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

[2] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[3] Department of Plant Taxonomy and Nature Conservation, University of Gdansk, Wita Stwosza 59, 80-308 Gdansk, Poland

[4] University of Tartu, 40 Lai Street, 51 005 Tartu, Estonia

Abstract

While mutualistic interactions are widespread and essential in ecosystem functioning, the emergence of uncooperative cheaters threatens their stability, unless there are some physiological or ecological mechanisms limiting interactions with cheaters.

In this framework, we investigated the patterns of specialization and phylogenetic distribution of mycoheterotrophic cheaters *versus* non-cheating autotrophic plants and their respective fungi in a global arbuscular mycorrhizal network with >25,000 interactions.

We show that mycoheterotrophy repeatedly evolved among vascular plants, suggesting low phylogenetic constraints for plants. However, mycoheterotrophic plants are significantly more specialized than autotrophic plants, and they tend to be associated with specialized and closely related fungi. These results raise new hypotheses about the mechanisms (*e.g.* sanctions, or habitat filtering) that actually limit the interaction of mycoheterotrophic plants and their associated fungi with the rest of the autotrophic plants.

Beyond mycorrhizal symbiosis, this unprecedented comparison of mycoheterotrophic *versus* autotrophic plants provides a network and phylogenetic framework to assess the presence of constraints upon cheating emergences in mutualisms.

Keywords: arbuscular mycorrhiza, mutualism, cheating, mycoheterotrophy, ecological networks, reciprocal specialization, phylogenetic constraint.

Author contributions: BPL, MAS, MÖ, HM, and FM designed the study. MÖ gathered the data, BPL performed the analyses and wrote the first draft of the manuscript, and all

authors contributed substantially to the writing and revisions.

Acknowledgments: The authors thank members of the INEVEF team at ISYEB and the BIODIV team at IBENS for helpful comments on the article. They also thank Colin Fontaine, Marianne Elias, and Damien de Vienne for helpful discussions, David Marsh for English editing, the Editor Björn Lindahl, and four anonymous reviewers for improvements of earlier versions of this manuscript. This work was supported by a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL and the École Doctorale FIRE - Programme Bettencourt. MÖ was supported by the European Regional Development Fund (Centre of Excellence EcolChange) and the Estonian Research Council (IUT20-28). HM acknowledges support from the European Research Council (grant CoG-PANDA).

Data availability: All the data used in this work are available in the MaarjAM database (`https://maarjam.botany.ut.ee`; Öpik *et al.*, 2010).

Citation: Perez-Lamarque B, Selosse MA, Öpik M, Morlon H, Martos F. 2020. Cheating in arbuscular mycorrhizal mutualism: a network and phylogenetic analysis of myco-heterotrophy. New Phytologist 226: 1822–1835.

Introduction:

Mutualistic interactions are ubiquitous in nature and largely help to generate and maintain biodiversity (Bronstein, 2015). Since benefits in mutualism often come at a cost for cooperators (Douglas, 2008), some species, referred to as cheaters, have evolved an adaptive uncooperative strategy by retrieving benefits from an interaction without paying the associated cost (Sachs *et al.*, 2010). Although cheating compromises the evolutionary stability of mutualistic interactions (Ferriere *et al.*, 2002), its evolutionary origin and persistence until present (hereafter referred to as cheating emergence) is often limited by factors securing the persistence of mutualism (Bronstein *et al.*, 2003; Frederickson, 2013; Jones *et al.*, 2015). For instance, species often favor the most cooperative partners (*e.g.* conditional investment; Roberts & Sherratt, 1998), stop interactions with cheaters (Pellmyr & Huth, 1994), or even sanction them (Kiers *et al.* 2003). Cheating emergence can thus be constrained through physiological or biochemical mechanisms of the interaction and its regulation. In addition, cheating can be restricted to particular habitats or to partners with specific niches. Therefore, cheaters might be constrained to specialize on susceptible partners and/or particular habitats. Moreover, these different constraints (hereafter referred to as functional constraints) can be evolutionarily conserved or not (Gómez *et al.* 2010). If they are conserved, there will be phylogenetic constraints on the emergence of cheaters, as some species will have evolutionarily conserved traits that make them more or less likely to cheat or to be cheated upon (Lallemand *et al.*, 2016).

The framework of bipartite interaction networks, combined with the phylogeny of partners, is useful for analyzing the patterns susceptible to arise from constraints limiting the emergence of cheaters in mutualisms (Figure III.6.1). Analyses of bipartite networks have been extensively used to showcase the properties of mutualistic interactions (Bascompte *et al.*, 2003; Rezende *et al.*, 2007; Martos *et al.*, 2012), such as their level of specialization (number of partners), nestedness (do specialists establish asymmetric specialization with partners that are themselves generalists?), and modularity (existence of distinct sub-networks; Bascompte & Jordano 2014). These studies, most of them describing species interactions at a local scale, have shown that mutualistic networks are generally nested with specialists establishing asymmetric specialization with more generalist partners, unlike antagonistic networks, which tend to be modular, with partners establishing reciprocal specialization (Thebault & Fontaine, 2010). However, few analyses of bipartite networks have focused on the specialization of cheaters and how they influence nestedness and modularity (Fontaine *et al.*, 2011). By assembling networks at a regional scale, Joffard *et al.* (2018) showed that specialization of orchids toward pollinators was higher in deceptive cheaters (both sexual and food deceits) than in cooperative nectar-producing species, and Genini *et al.* (2010) showed that a network dominated by cooperative pollinators was nested, whereas another network dominated by nectar thieving insects was more modular. If cheaters specialize and form modules, this would suggest the presence of functional constraints limiting the set of species that they can exploit (Figure III.6.1b-v). Additionally, if cheaters emerged only once in a phylogeny (versus repeatedly), and/or if 'cheating-susceptible' partners are phylogenetically related (Merckx *et al.*, 2012), this would suggest that cheating involves some rare evolutionary innovations (Pellmyr et al,. 1996) and/or that cheating susceptibility is limited to few clades, meaning that cheating is phylogenetically constrained (Figure III.6.1a-i).

Here we study cheating emergences in arbuscular mycorrhizal mutualism between plant roots and soil Glomeromycotina fungi (Selosse & Rousset, 2011; Jacquemyn & Merckx, 2019). This symbiosis is at least 407 Myr-old (Strullu-Derrien *et al.*, 2018) and concerns ca. 80% of extant land plants and several hundred fungal taxa (Davison *et al.*, 2015; van der Heijden *et al.*, 2015). Arbuscular mycorrhizal fungi colonize plant roots and provide host plants with water and mineral nutrients, in return for organic carbon compounds (Rich *et al.*, 2017). Although obligate for both partners, this symbiosis is generally diffuse and not very specific (van der Heijden *et al.*, 2015), since multiple fungi colonize most plants, while fungi are usually shared among surrounding plant species (Verbruggen *et al.* 2012). Thus, fungi interconnect plant individuals of different species and allow resource movement between plants (Selosse *et al.*, 2006; Merckx, 2013). This allowed the emergence of achlorophyllous cheating plants, called mycoheterotrophs, which obtain carbon from their mycorrhizal fungi that are themselves fed by surrounding autotrophic plants (Merckx, 2013) - these plants are thus permanent cheaters, whatever the conditions or partners. Some of these plant species are entirely mycoheterotrophic over their lifecycle, while others are mycoheterotrophic only at early

stages before turning autotrophic (initially mycoheterotrophic), therefore shifting from being cheaters to becoming potentially cooperative partners (Merckx, 2013). Unlike other systems where cheaters are costly (they receive the benefits without paying the cost of the interaction) mostly for direct partners (*e.g.* in plant pollination), mycoheterotrophs are costly for both their direct fungal partners and the interconnected autotrophic plants, whose photosynthesis supplies the carbon (it represents a projected cost, transmitted through the network). Although uncooperative strategies between autotrophic plants and arbuscular mycorrhizal fungi may exist under certain conditions (Klironomos, 2003; Jacquemyn & Merckx, 2019; but discussed in Frederickson, 2017), autotrophs can supply photosynthetic carbon and are mostly cooperative, while mycoheterotrophs never supply photosynthetic carbon and are therefore necessarily uncooperative.

We evaluate the presence of functional constraints upon cheating by measuring specialization, nestedness and modularity in a composite plant-mycorrhizal fungal interaction network built from associations between species at multiple sites across the entire globe (Öpik *et al.*, 2010). Mycoheterotrophic plants are thought to be specialists interacting with few fungal species (Leake, 1994; Merckx, 2013), but whether or not these plant species are unusually specialized compared to autotrophic plants is still debated (Merckx *et al.*, 2012). Mycoheterotrophs could specialize on few fungal species if some functional constraints limit the set of fungi or habitats they can exploit, and if they have evolved particular strategies to obtain nutrients from their specific fungal partners (Blüthgen *et al.*, 2007). In terms of nestedness and modularity, arbuscular mycorrhizal networks are generally nested (Chagnon *et al.*, 2012; Sepp *et al.*, 2019); this pattern of asymmetrical specialization is generally thought to confer greater stability in relation to disturbance and resistance to species extinction (Thébault & Fontaine, 2010). How mycoheterotrophic plants affect nestedness has yet to be investigated. On the one hand, in the absence of functional constraints upon cheating, we would expect that mycoheterotrophs interact with generalist fungi to increase their indirect access to carbon via surrounding autotrophic plants, therefore increasing nestedness (Figure III.6.1b–v,viii). On the other hand, if autotrophic plants are able to avoid costly interactions with fungi associated with mycoheterotrophs (physiological constraints), or if mycoheterotrophs are only tolerated in particular habitats (ecological constraints), we expect a reciprocal specialization between mycoheterotrophs and their fungi and thus an increase of modularity and a decrease of nestedness (Figure III.6.1b–v,vii). Establishment of an extreme reciprocal specialization between entirely mycoheterotrophs and fungi exclusively associated with such plants seems unlikely though, since an autotrophic carbon source is required.

With regards to phylogenetic constraints on mycoheterotrophy, we already know that mycoheterotrophic strategies evolved multiple times (Merckx, 2013), generating monophyletic groups of mycoheterotrophic plants, which suggests weak phylogenetic constraints on the emergence of mycoheterotrophy in plants. However, the fungi interacting with independent mycoheterotrophic lineages might be phylogenetically closely related

(Merckx *et al.*, 2012), which would indicate phylogenetic constraints on fungi (Figure III.6.1a–iii). The presence of such phylogenetic constraints has yet to be confirmed in a large phylogenetic context including the fungi of autotrophic plants. Moreover, if as we expect, only a set of phylogenetically close fungi interact with all mycoheterotrophic plant lineages, an important follow-up question is whether these fungi were acquired independently by autotrophic ancestors, or whether they were acquired by symbiont shift from other mycoheterotrophic plants.

Methods:

MaarjAM database and interaction matrix:

The MaarjAM database is a web-based database (`http://maarjam.botany.ut.ee`; accessed in June 2019 after a very recent update) of publicly available sequences of Glomeromycotina fungi, with information on the host plants, geographical location and biomes for the recorded interactions (Öpik *et al.*, 2010). We used an approach with a compiled network, where all locally described physical mycelial interactions between species are merged and studied at larger scales (as in Joffard *et al.*, 2018). Although such a compiled network can be sensitive to several biases (see Discussion), it offers unique opportunities to study the emergence of mycoheterotrophy in a large evolutionary and ecological perspective (*e.g.* Werner *et al.*, 2018). Among the 41,989 interactions between plants and Glomeromycotina, we filtered out the data from MaarjAM for the fungi to satisfy the following criteria (Supplementary Table S1a): (i) amplification of the 18S rRNA gene, (ii) fungus identified from plant roots (*i.e.* excluding soil samples), (iii) interaction in a natural ecosystem (*i.e.* excluding anthropogenic or highly disturbed ecosystems), (iv) host plant identified at the species level, and (v) a *virtual taxon* (VT) assignation available in MaarjAM. The VTs are a classification (=species proxy) of arbuscular mycorrhizal fungi designed by applying a $\geq$97% sequence similarity threshold to the 18S rRNA gene sequences, and by running phylogenetic analysis to ensure VT monophyly (Öpik *et al.*, 2013, 2014). In the following, we assumed that we have a full representation of all fungal partners associated with each plant species in the dataset. The filtered dataset yielded a binary interaction matrix of 490 plant species (hereafter 'plants'), 351 VTs (hereafter 'fungi'), and 26,350 interactions (Figure III.6.2), resulting from the compilation of 112 publications from worldwide ecosystems (Supplementary Figure 1; Supplementary Table 1b). In order to estimate the sampling fraction of Glomeromycotina fungi in our dataset, we plotted rarefaction curves of the number of fungal species as a function of the sampling fraction (for the observed number of interactions or for the number of sampled plant species) and we estimated the total number of species using the *specpool* function ('vegan' R-package, based on Chao index; Oksanen *et al.*, 2019). We separately performed rarefaction analyses for mycoheterotrophic species only. Moreover, in order to check the robustness of our results, we repeated all the analyses on a subsampled version of the MaarjAM database accessed in October 2017 (Supplementary Figure 2).

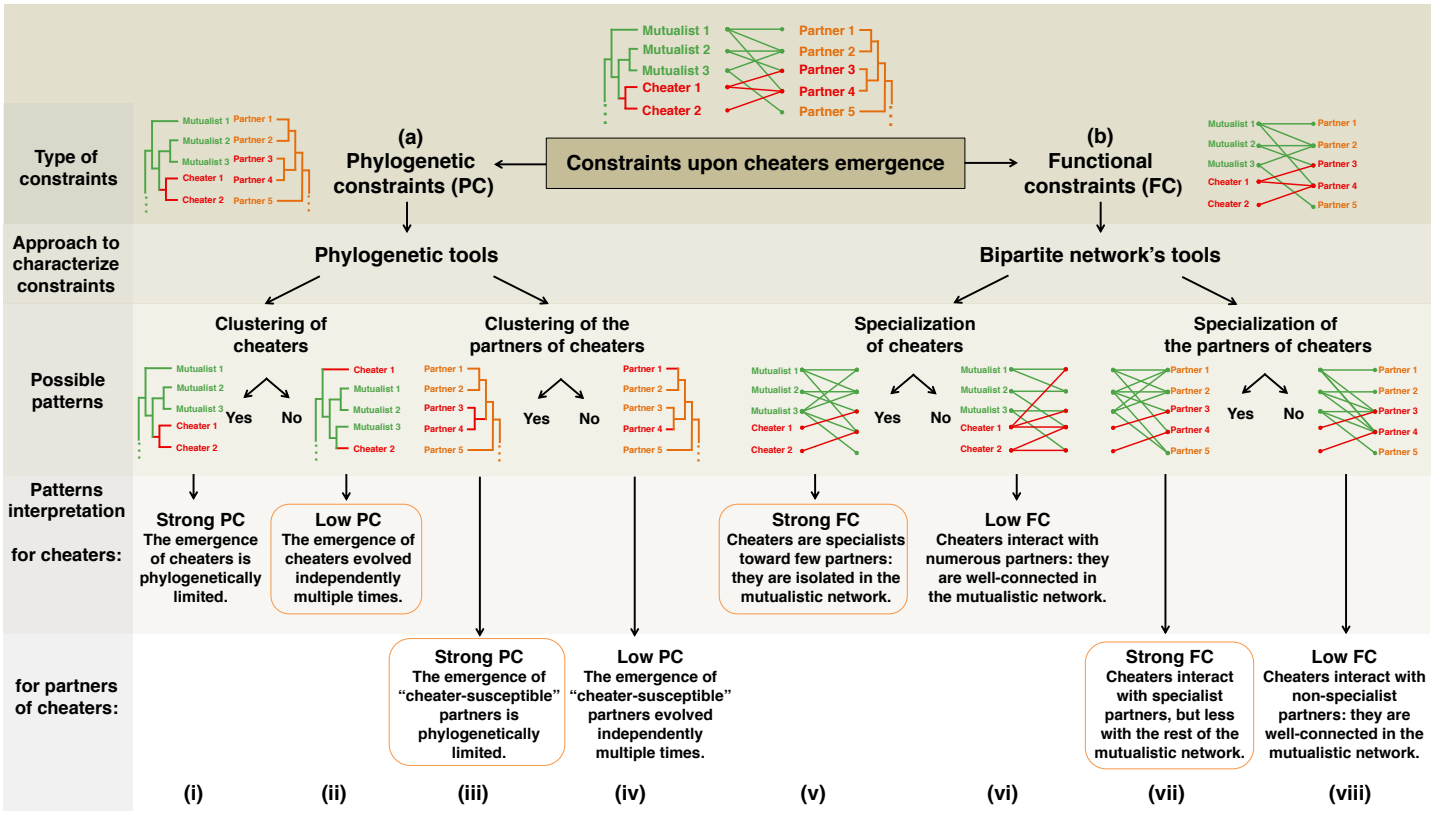


Figure III.6.1: Conceptual framework used in this study to evaluate the constraints upon the emergence of mycoheterotrophic cheater plants in arbuscular mycorrhizal symbiosis. (a) Strong phylogenetic constraints (PC) should affect the phylogenetic distributions of mycoheterotrophic cheater plants and/or their fungal partners; whereas (b) functional constraints (FC; *e.g.* physiological or ecological constraints) should affect the network structure *i.e.* level of specialization of mycoheterotrophic cheater plants and/or their partners. Therefore, by investigating specialization and phylogenetic clustering of mycoheterotrophic cheaters and of their fungal partners, we evaluated functional and phylogenetic constraints. This can be done by using and interpreting bipartite network tools (a – *e.g.* computation of nestedness, measures of partner degree, and partner specialization) or phylogenetic tools (b – *e.g.* measure of phylogenetic dispersion), respectively. Interpreting the observed patterns of phylogenetic clustering and network structure directly indicates the strength of the constraints. For instance, strong phylogenetic clustering of the cheaters and their partners (i-iii) suggests that the emergence of cheaters and their susceptible partners is rare and limited, whereas low phylogenetic clustering (ii-iv) suggests that cheating evolved multiple times. Regarding functional constraints, generalist cheaters (vi) might indicate that their partners do not have any mechanisms preventing uncooperative interactions (low constraints). Conversely, specialist cheaters (v) might indicate that cheaters cannot interact with most partners (high constraints). Moreover, if the partners of cheaters are generalists (viii - low constraints), asymmetrical specialization ensures that cheaters are well connected in the interaction network (high nestedness), whereas if they are specialists (vii - high constraints), reciprocal specialization on both sides drives the isolation of mycoheterotrophic plants into modules, thus decreasing nestedness. Mutualistic species are represented in green and their partners are in orange, whereas cheaters and their partners are represented in red. Mutualistic interactions are thus represented in green, whereas antagonistic interactions (cheating) are in red. The patterns and interpretations from the present study on mycoheterotrophic cheaters are shown in the orange frames.

Phylogenetic reconstructions:

We aligned consensus sequences of the 351 fungi with MUSCLE (Edgar, 2004) and ran a Bayesian analysis using BEAST2 to reconstruct the fungal phylogeny (Bouckaert *et al.* 2014, Supplementary Methods S1). We obtained the phylogenetic relationships between the 490 host plants by pruning the time-calibrated supertree from Zanne *et al.* (2013) using Phylomatic (`http://phylodiversity.net/phylomatic/`). We also used the Open Tree of Life website (`http://opentreeoflife.org`) and the 'rotl' R-package (Michonneau *et al.* 2016; R Core Team, 2019) for grafting of 41 plant taxa missing from the pruned supertree (as polytomies at the lowest taxonomy level possible; Supplementary Methods S1). We set tree root calibrations at 505 million years (Myr) for the fungi (Davison *et al.*, 2015) and 440 Myr for the plants (Zanne *et al.*, 2014).

Nature of the interaction:

We assigned to each plant its 'nature of the interaction' with fungi according to its carbon nutrition mode according to an on-line database (`http://mhp.myspecies.info/category/myco-heterotrophic-plants/`) and individual publications (Boullard, 1979; Winther & Friedman, 2008; Field *et al.* 2015): autotroph (n=434, 88.6%), entirely mycoheterotroph (n=41, 8.4%), or initially mycoheterotroph (n=15, 3.1%). We assigned each fungus to three categories: 'associated with autotrophs' if the fungus interacts with autotrophic plants only (n=280, 79.8%), 'associated with entirely mycoheterotrophs' if the fungus interacts with at least one entirely mycoheterotroph (n=54, 15.4%), 'associated with initially mycoheterotrophs' if the fungus interacts with at least one initially mycoheterotroph (n=23, 6.6%), or 'associated with mycoheterotrophs' if the fungus interacts with at least one entirely or initially mycoheterotroph (n=71, 20.2%; Supplementary Table 2). Only five fungi are associated with both entirely and initially mycoheterotrophic plants. Our dataset included mycoheterotrophs from 18 publications. While only 41 entirely mycoheterotrophic species were included out of 267 described species (Jacquemyn & Merckx, 2019), all known entirely mycoheterotrophic families were represented by at least one plant species, except the families Aneuraceae (liverwort, one mycoheterotrophic species), Iridaceae (monocotyledons, three species), and Podocarpaceae (gymnosperm, one controversial species). Similarly, our dataset missed only a few initially mycoheterotrophic families, such as Schizaeaceae (Boullard, 1979).

Network nestedness, modularity, and specialization of cheaters:

In order to assess the functional constraints upon cheating, we tested the effect of mycoheterotrophy on network structure (Figure III.6.1b). First, we measured nestedness in: (i) the overall network (490 plants, 351 fungi, and 26,350 interactions), (ii) the network restricted to autotrophic plants (434, 344, and 26,087) and (iii) the network restricted to entirely and initially mycoheterotrophic plants (56, 71, and 263), using the function $NODF2$ in the R-package bipartite (Dormann *et al.* 2008). We tested the significance of NODF values (nestedness metric based on overlap and decreasing fill; Supplementary Methods S2

- List of abbreviations) by using two types of null models (N=100 for each type): the first model (*r2dtable* from the stats R-package - null model 3) maintains the marginal sums of the network (the sums of each row and each column), whereas the less stringent second model (*vaznull* from the bipartite R-package - null model 2) produces slightly different marginal sums (interactions are randomized with species marginal sums as weights, and each species must have at least one interaction), while maintaining the connectance (proportion of observed interactions). We calculated the Z-score, which is the difference between the observed value and the mean of the of null-models values divided by their standard deviation (Z-scores greater than 1.96 validate a significant nestedness with an alpha-risk of 5%). Positive z-scored NODF values indicate nested networks.

Second, to further evaluate the specialization of mycoheterotrophic plants, we computed several network indices for each plant. The degree (k) is the number of partners with which a given plant or fungus interacts in the bipartite network. The degree is high (*vice versa* low) when the species is generalist (*vice versa* specialist). The partner specialization (Psp) is the mean degree (k) averaged for all the fungal partners for a given plant species (Taudiere *et al.*, 2015): a high (*vice versa* low) Psp characterizes a species interacting mainly with generalist (*vice versa* specialist) partners. Simultaneously low k and Psp values feature a reciprocal specialization (Figure III.6.1b-v,vii). We tested whether k and Psp were statistically different among autotrophic, entirely mycoheterotrophic and initially mycoheterotrophic plants using non-parametric Kruskal-Wallis tests and pairwise Mann-Whitney U tests. To assess the significance of k and Psp values, we built null-model networks (N=1,000) using the function $permatfull$ in the vegan R-package (null model 1), keeping the connectance constant but allowing different marginal sums. Then, in order to detect specialization at the clade scale toward partners, for any given clade of every node in the plant or fungus phylogenies, we calculated the partner fidelity (Fx) as the ratio of partners exclusively interacting with this particular clade divided by the total number of partners interacting with it. We consider the clade as 'faithful' and the corresponding set of partners as 'clade-specific' when Fx>0.5 (i.e., more than 50% exclusive partners). We used analysis of covariance (ANCOVA) to test the effect of the nature of the interaction on partner fidelity Fx accounting for clade size, which corrects the bias of having high partner fidelity Fx in older clades including many plants. To confirm that the patterns of specialization at the global scale held at a more local scale, we reproduced the analyses of specialization (k and Psp) in two continental networks in South America and Africa, which represented a high number of interactions and mycoheterotrophic species.

Third, we investigated signatures of reciprocal specialization in the overall network structure. We used the DIRTLPAwb+ algorithm (Beckett, 2016) to infer modules and assess their significance (a module is significant if it encompasses a subset of species interacting more with each other than with the rest of the species) and used the function *components* of the R-package igraph (Csardi & Nepusz, 2006) to detect cases of extreme

reciprocal specializations leading to independent modules (two species belong to two distinct independent modules if there is no path in the network going from one to the other, *i.e.* an independent module is the smallest subset of species exclusively interacting with each other).

We replicated these statistical tests without the initially mycoheterotrophic Lycopodiaceae forming different network patterns (see Results).

Phylogenetic distribution of cheating:

In order to assess phylogenetic constraints, we explored the phylogenetic distribution of mycoheterotrophic plants and their associated fungi (Figure III.6.1a). First, we investigated the phylogenetic distribution of mycoheterotrophy, *i.e.* if mycoheterotrophic plants and their fungal partners were more or less phylogenetically related than expected by chance (patterns of clustering *versus* overdispersion). We computed the net relatedness index (NRI) and the nearest taxon index (NTI) using the PICANTE R-package (Kembel *et al.*, 2010). NRI quantifies the phylogenetic structure of a species set based on the mean pairwise distances, whereas NTI quantifies the terminal structure of the species set by computing the mean phylogenetic distance to the nearest taxon of every species (Gotelli & Rohde, 2002). To standardize the indices, we generated 999 null models with the option 'taxa.labels' (shuffles the taxa labels). Significant positive (resp. negative) NRI and NTI values indicate phylogenetic clustering (resp. overdispersion). We computed these indices (i) on the plant phylogeny to evaluate the phylogenetic structure of entirely mycoheterotrophic and initially mycoheterotrophic plant distribution, and (ii) on the fungal phylogeny to investigate if fungi associated with mycoheterotrophs were phylogenetically structured (we successively tested the distribution of the fungi associated with mycoheterotrophs, entirely mycoheterotrophs, or initially mycoheterotrophs, and then of the fungi associated with each specific mycoheterotrophic family). Similarly, for each plant, we computed the partners' mean phylogenetic pairwise distance (MPD), that is the average phylogenetic distance across pairs of fungal partners (Kembel *et al.*, 2010): a low value of MPD indicates that the set of partners is constituted of closely related species. The effect of mycoheterotrophy on MPD values and its significance were evaluated as for k and Psp values above.

Second, in order to assess whether fungal partners of a given mycoheterotrophic family were derived from fungal partners of autotrophic ancestors or were secondarily acquired from other mycoheterotrophic lineages, we compared in an evolutionary framework the sets of fungi associated with plants with different natures of the interaction. To do so, we computed the unweighted UniFrac distance (Lozupone & Knight, 2005) between sets of fungi interacting with each pair of plants in the network. For each of the seven mycoheterotrophic families, we compared the UniFrac distances across (i) every pair of plant species of this family, (ii) every pair comprising one plant of this family and one plant of the most closely related autotrophic family (see Table III.6.2), (iii) every pair

composed of one plant of this family and one plant belonging to other mycoheterotrophic families, and (iv) every pair comprising one plant of this family and one more distant autotrophic plant (*i.e.* all autotrophic plants except those of the most closely related autotrophic family). This analysis was not performed on mycoheterotrophic Petrosaviaceae, which were represented by only one species and were too divergent to define a reliable autotrophic sister clade.

We tested differences between groups of distances using Mann-Whitney U tests. We also performed a principal coordinates analysis (PCoA) from all the UniFrac dissimilarities of sets of fungal partners, and tested the effect of the nature of the interaction on the two principal coordinates, using Kruskal-Wallis tests. Finally, to examine the extent to which the nature of the interaction affects fungal partners, we used permutational analysis of variance (PERMANOVA, *adonis* function in the vegan R-package), with 10,000 permutations.

Results:

Completeness of the dataset:

We estimated a total number of 373±9 fungal species (Chao index), which indicated that the 351 fungi in the dataset included most of the arbuscular mycorrhizal fungal diversity (94% ± 2%; Supplementary Figure 3). Concerning mycoheterotrophic species, we estimated a total of 117 ± 19 fungi associated with all mycoheterotrophs, 110 ± 27 fungi associated with entirely mycoheterotrophs, and 54 ± 24 fungi associated with initially mycoheterotrophs. Our dataset thus encompassed sampling fractions of 60% ± 10% for fungi associated with mycoheterotrophs, 49% ± 10% for fungi associated with entirely mycoheterotrophs, and 40% ± 28% for fungi associated with initially mycoheterotrophs. Although our dataset did not include all the fungi associated with mycoheterotrophic species, the following results were not sensitive to the sampling fractions of mycoheterotrophs and their fungal partners (Supplementary Figure 2).

Network nestedness, modularity, and specialization of mycoheterotrophs:

The overall network had a significant positive nestedness value (Z-score=9.2, P=1.10^{-20}, Supplementary Table 3). Nestedness increased when only autotrophic plants were considered (Z-score=16.6, P=8.10^{-62}), whereas it was not significant in the network of only mycoheterotrophs (Z-score=1.44, P=0.075): mycoheterotrophic plants reduced nestedness, signifying that they displayed higher reciprocal specializations. Reciprocal specializations were confirmed by the analyses of modularity, which found no significant large modules (*i.e.* the inferred large modules presented more inter-modules than intra-module interactions), suggesting that the overall structure was not modular, but detected few significant small independent modules (Supplementary Table 4). In addition to

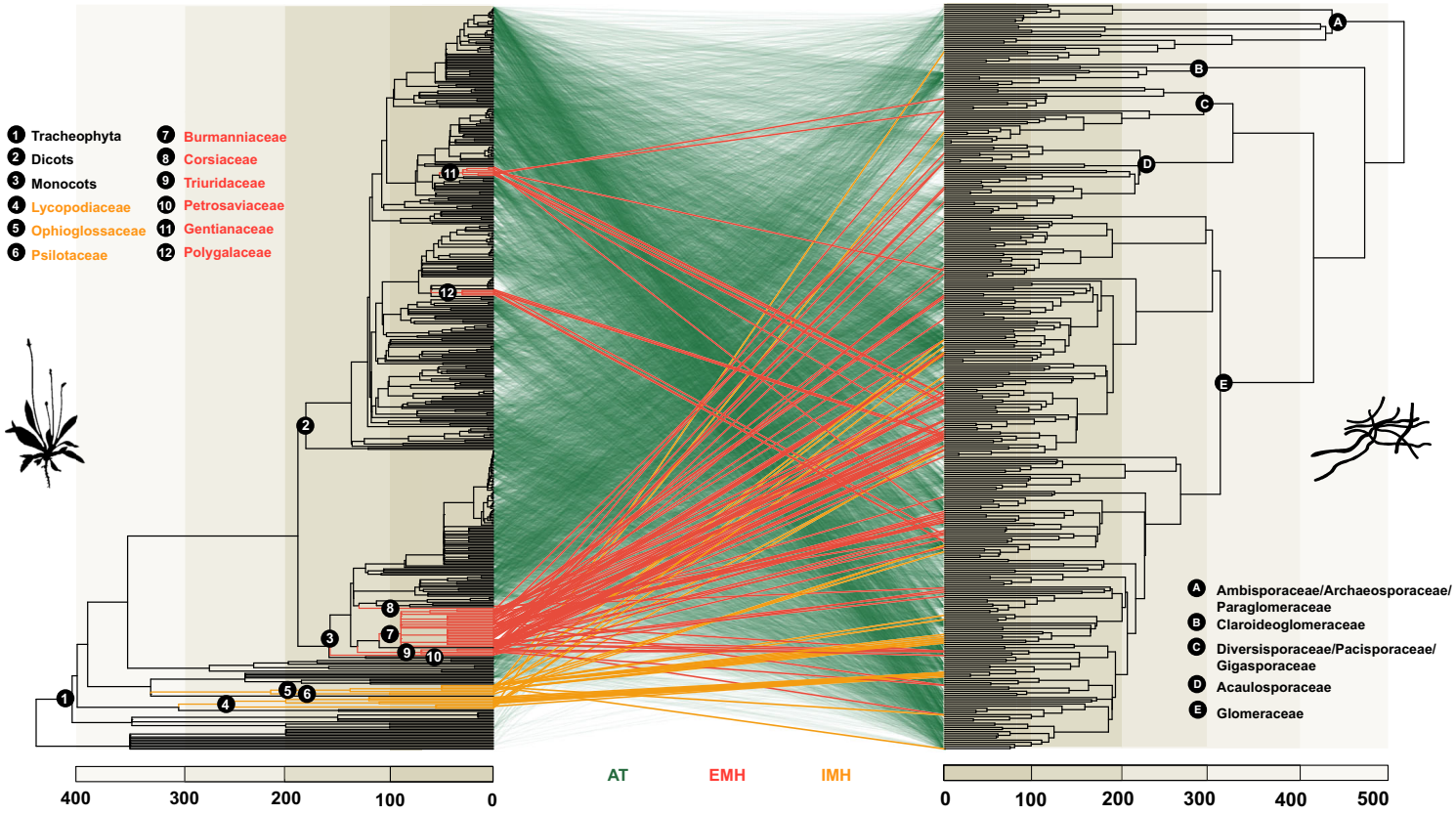


Figure III.6.2: Phylogenetic distribution of mycoheterotrophy in global arbuscular mycorrhizal mutualism. (Categories are defined according to the plant carbon nutrition modes, *i.e.* AT: autotrophic; EMH: entirely mycoheterotrophic throughout the life cycle of the individual plant; and IMH: initially mycoheterotrophic in the life cycle). Phylogenetic trees of 390 plants (left side) and 351 fungi (right side) forming 26,350 interactions (links) in the MaarjAM database. Links are colored according to the autotrophic (green), entirely mycoheterotrophic (red), or initially mycoheterotrophic (orange) nature of the plant. Major plant and fungal clades are named. Mycoheterotrophy encompasses 41 entirely mycoheterotrophic species in 6 monophyletic families [Burmanniaceae (25 spp.), Gentianaceae (6 spp.), Triuridaceae (4 spp.), Polygalaceae (4 spp.), Corsiaceae (1 sp.), and Petrosaviaceae (1 sp.)], and 15 initially mycoheterotrophic species in 3 families [Ophioglossaceae (ferns; 5 spp.), Psilotaceae (ferns; 2 spp.), and Lycopodiaceae (clubmoss; 8 spp.)]. Scales of the phylogenetic trees are in million years (Myr).

a main module encompassing most species (481 out of 490 plants and 346 out of 351 fungi), we found three small independent modules: (i) 6 initially mycoheterotrophic Lycopodiaceae plants and three exclusive fungi (*Glomus* VT127, VT158, VT394); (ii) two autotrophic plants from salt marshes (*Salicornia europaea* and *Limonium vulgare*) with one *Glomus* (VT296); and (iii) the entirely mycoheterotrophic *Kupea martinetugei* with a unique *Glomus* (VT204).

From the degrees (k), we found that entirely and initially mycoheterotrophic plants were significantly more specialized than autotrophic plants and interacted with on average more than five times fewer fungi (Kruskal-Wallis H=87.2; P=1.2.10^{-19}; Figure III.6.3a; Table III.6.1). Partner specializations (Psp) indicated that mycoheterotrophs interacted with more specialized fungi (fungi associated with mycoheterotrophs interact on average with two times fewer plants; Kruskal-Wallis H=47.2; P=5.6.10^{-11}; Figure III.6.3a). We found similar evidence for mycoheterotrophic reciprocal specializations by reanalyzing the network excluding the family Lycopodiaceae (Table III.6.1; significance assessments using null models are shown in Supplementary Table 5). This pattern of reciprocal specialization of mycoheterotrophic plants and their associated fungi held at a smaller geographical scale in the African and South American networks (Supplementary Figure 4; Supplementary Table 6; yet the difference was not significant for Psp in the South American network, probably due to the small number of species and the low power of the statistical tests).

Index	Kruskal-Wallis test	Whitney U tests		
		AT vs. EMH	AT vs. IMH	IMH vs. EMH
Plant degree (*k*)	**1.2e-19 (1.4e-17)**	**5.3e-16**	**4.2e-7**	0.97
Fungal partner specialization (*Psp*)	**5.6e-11 (1.3e-8)**	**1.1e-4**	**4.5e-9**	0.054
Mean phylogenetic pairwise distance of fungal partners (*MPD*)	**1.2e-4 (2.0e-3)**	**8.0e-4**	**6.8e-3**	0.11

Table III.6.1: Effect of the nature of the interaction (*i.e.* plant carbon nutrition modes) on indices of network structure and phylogenetic distributions. (Categories are defined according to the plant carbon nutrition modes, *i.e.* AT: autotrophic; EMH: entirely mycoheterotrophic over development; and IMH: initially mycoheterotrophic in development). The second column corresponds to P-values of Kruskal-Wallis tests for the overall network with or without (in brackets) the Lycopodiaceae. The last three columns correspond to P-values of Whitney U tests (pairwise tests) for the overall network including the Lycopodiaceae. P-values lower than 5% (significance level) are shown in bold.

The partner fidelity index (Fx) showed that very few plant and fungi clades interacted with 'clade-specific' partners (*i.e.* Fx>0.5), and most fungi were shared between different plant clades (Figure III.6.3c). Among exceptions, however, the clade of initially mycoheterotrophic Lycopodiaceae was characterized by a high partner fidelity index (Fx>0.8), reflecting a strong association with a clade of three Lycopodiaceae-associated fungi (Sup-

plementary Figure 5). Thus, not only did these 6 Lycopodiaceae species and their fungal partners form an independent module, but the Lycopodiaceae-associated fungi also formed a monophyletic clade within Glomeromycotina. The estimated clade age was 250 Myr for the Lycopodiaceae and 49 Myr for the Lycopodiaceae-associated fungi (Figure III.6.3d), which diverged 78 Myr ago from the other *Glomus* fungi.

Phylogenetic distribution of cheating :

The partners' mean phylogenetic pairwise distance (MPD) indicated that fungi associated with entirely or initially mycoheterotrophs (or even with all mycoheterotrophs) were phylogenetically more closely related than fungi associated with autotrophs (Kruskal-Wallis H=18.0; P=1.2.10^{-4}; Table III.6.1; Figure III.6.3b). NRI and NTI values (Supplementary Table 7) also confirmed significant clustering on the fungal phylogeny on fungi associated with mycoheterotrophs, entirely mycoheterotrophs, or initially mycoheterotrophs; this clustering held at the family level for fungi associated with each of four main mycoheterotrophic families (namely Burmanniaceae, Triuridaceae, Polygalaceae, and Ophioglossaceae). In terms of the plants, only the entirely mycoheterotrophs were significantly clustered, mainly because they all were angiosperms and mostly monocotyledons, but this did not apply to mycoheterotrophs in general, nor to initially mycoheterotrophs (Supplementary Table 7). These phylogenetic clusters were visually noticeable on fungal and plant phylogenetic trees (Supplementary Figures 6 & 7). This suggests that although mycoheterotrophy evolved several times independently in plants, mycoheterotrophic plants interact mainly with closely related fungi (see also Figure III.6.2).

Looking specifically at the fungi shared among mycoheterotrophic plants highlighted differences between entirely and initially mycoheterotrophs (Table III.6.2). While the initially mycoheterotrophic Lycopodiaceae family formed an independent module with three specific *Glomus* VTs, another initially mycoheterotrophic family Ophioglossaceae also had 2 exclusive fungi (*Glomus* VT134 and VT173) among a total of 15 fungi. When comparing the fungi shared between mycoheterotrophic families (Table III.6.2), mainly two closely related families, Burmanniaceae and Triuridaceae, tended to share some fungi with other mycoheterotrophic families.

The decomposition of UniFrac dissimilarities between sets of fungal partners using a PCoA, showed a clear pattern of clustering of mycoheterotrophic species, indicating that the set of fungal partners associated with mycoheterotrophs were more similar than expected by chance (P<1.10^{-16} for PCoA1; P=9.10^{-3} for PCoA2; Figure III.6.4a). Similarly, the PERMANOVA analysis indicated that the nature of the interaction (initially mycoheterotrophic, entirely mycoheterotrophic, or autotrophic) predicted 6.5% of the variance (P=0.0001). By comparing the UniFrac dissimilarities between sets of fungal partners according to the nature of the interaction and plant family relatedness, we observed that all mycoheterotrophic families had fungal partners more similar to each other than those of other autotrophic families (Figure III.6.4b; Supplementary Table 8). Some families

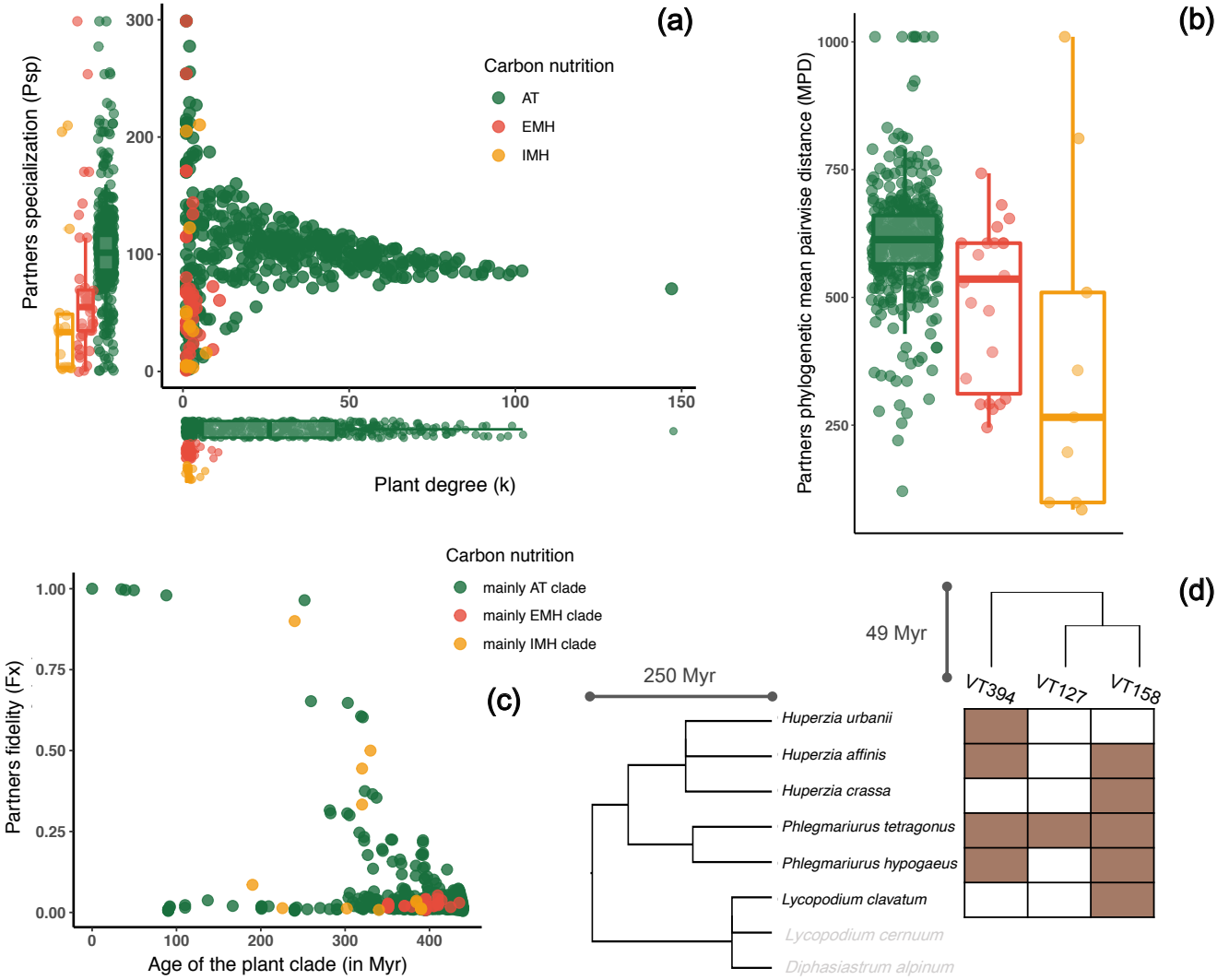


Figure III.6.3: Effect of the nature of the interaction on specialization (*k* and *Psp*), the partner's mean phylogenetic distance (MPD), and partner fidelity (*Fx* - Supplementary Methods S2 - List of abbreviations): (Categories are defined according to the plant carbon nutrition modes, *i.e.* AT: autotrophic; EMH: entirely mycoheterotrophic over development; and IMH: initially mycoheterotrophic in development). (a): Plant degree (k) against fungal partner specialization (Psp) (*i.e.* the average degree of fungal partners); dots in the bottom left corner indicate reciprocal specialization. For each axis, boxplots represent the one-dimensional projection of k and Psp. (b): Mean phylogenetic pairwise distance (MPD) of the sets of fungal partners. Boxplots present the median surrounded by the first and third quartiles, and whiskers extend to the extreme values but no further than 1.5 of the inter-quartile range. (c): Fidelity (Fx) toward fungal partners in relation to the age of the plant clade. Clades are defined according to their main carbon nutrition mode of their plants (over 50%). The yellow dots departing from other mycoheterotrophic clades (high Fx values) correspond to clades of Lycopodiaceae. (d): Independent network between the clubmoss family Lycopodiaceae (rows) and their three arbuscular mycorrhizal fungi (columns), with their respective phylogenetic relationships.

		Most closely related autotrophic sister-clade in our dataset (and divergence time in million years)	Number of plant species	Number of fungal partners	Burmanniaceae	Corsiaceae	Gentianaceae	Petrosaviaceae	Polygalaceae	Triuridaceae	Lycopodiaceae	Ophioglossaceae	Psilotaceae	Total number of shared fungi	Total number of exclusive fungi
EMH	**Burmanniaceae**	Dioscoreaceae **(110 Myr)**	25	38		0%	13%	3%	8%	16%	2%	6%	0%	16	0
	Corsiaceae	Melanthiaceae Liliaceae Smilacaceae **(129 Myr)**	1	1	0		0%	0%	0%	0%	0%	0%	0%	0	0
	Gentianaceae	Apocynaceae **(52 Myr)**	6	8	5	0		0%	9%	8%	0%	0%	0%	6	0
	Petrosaviaceae		1	2	1	0	0		0%	5%	0%	0%	0%	1	0
	Polygalaceae	Polygalaceae **(60 Myr)**	4	4	3	0	1	0		10%	0%	0%	0%	3	1
	Triuridaceae	Dioscoreaceae **(131 Myr)**	4	19	8	0	2	1	2		8%	3%	0%	11	1
IMH	**Lycopodiaceae**	Selaginellaceae **(303 Myr)**	8	7	1	0	0	0	0	2		0%	0%	3	3
	Ophioglossaceae	Aspleniaceae Dryopteridaceae Gleicheniaceae Lygodiaceae Osmundaceae Pteridaceae **(330 Myr)**	5	15	3	0	0	0	0	1	0		6%	4	2
	Psilotaceae		2	2	0	0	0	0	0	0	0	1		1	0

Table III.6.2: Fungal sharing between nine entirely (EMH) or initially (IMH) mycoheterotrophic plant families. Number (lower part of the matrix) and percentage (upper part) of fungi shared between family pairs. The last two columns represent (i) the total number of fungi shared with other entirely or initially mycoheterotrophic families, and (ii) the number of fungi exclusive to this family (*i.e.* not shared with any other mycoheterotrophic or autotrophic family). The second column indicates the most closely related autotrophic sister clade of each family; it can be one family, a higher clade, the family itself if autotrophic species were compiled in the MaarjAM database (*e.g.* Polygalaceae), or none in the case of Petrosaviaceae (which forms a too divergent distinct branch). Boxes are shaded according to the number of shared fungi (white: no shared fungus, black: many shared fungi).

(Burmanniaceae, Polygalaceae, Triuridaceae, Lycopodiaceae, and Ophioglossaceae) had fungal partners significantly more similar to partners interacting with their closest autotrophic relatives (P>0.05) than to partners interacting with other autotrophic families (P<10^{-16}). This suggests phylogenetic conservatism of fungal partners during the evolution of mycoheterotrophic nutrition in these families. For other mycoheterotrophic families (Corsiaceae, Gentianaceae, and Psilotaceae), fungal partners were significantly more similar to partners interacting with other mycoheterotrophic families than to partners interacting with their closest autotrophic relatives, the latter being as distant as other autotrophic families (Supplementary Table 8). This points to a shift to new fungal partners correlated with the evolution of mycoheterotrophic nutrition in these three families.

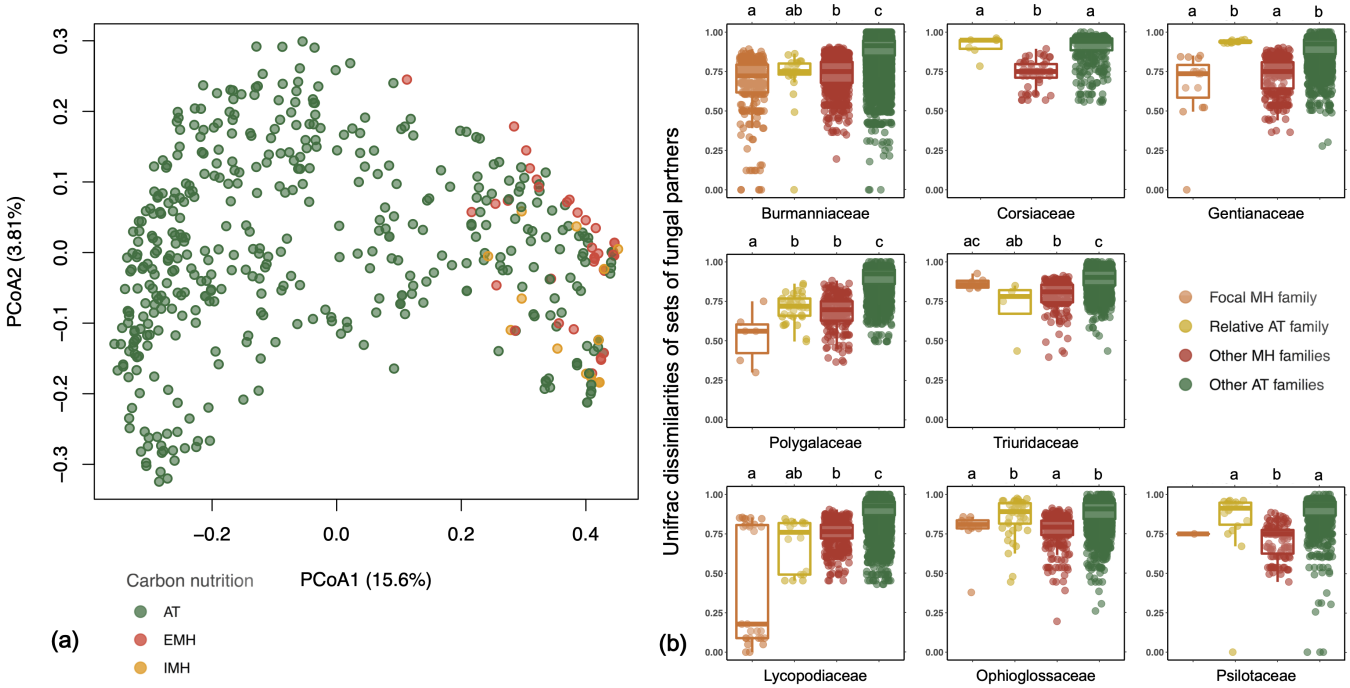


Figure III.6.4: Dissimilarities between sets of fungal partners associated according to the nature of the interaction. (a): Principal coordinates analysis (PCoA) from UniFrac dissimilarities of sets of fungal partners. Every dot corresponds to a plant species and is colored according to its autotrophic (green), entirely mycoheterotrophic (red), or initially mycoheterotrophic (orange) nature. Only the first two principal axes explaining, respectively, 15.6% and 3.8% of the variation were kept. (b): Dissimilarities between sets of fungal partners associated with different mycoheterotrophic plant families. For each mycoheterotrophic family, UniFrac dissimilarities of sets of fungal partners are calculated between one particular mycoheterotrophic species belonging to the focal mycoheterotrophic family and another plant species (from the same family, from the closest related autotrophic family, from other mycoheterotrophic families, or from other autotrophic plant families). All the groups cannot be calculated for every mycoheterotrophic family, due to the low number of species within families Corsiaceae and Psilotaceae. Lowercase letters above each panel represent significant differences between categories (Mann-Whitney U tests). Boxplots present the median surrounded by the first and third quartiles, and whiskers extend to the extreme values but no further than 1.5 of the inter-quartile range.

Discussion:

By combining network and phylogenetic analyses, we assessed constraints upon the emergence of mycoheterotrophic cheating in arbuscular mycorrhizal mutualism. Although the network was nested, we found evidence for reciprocal specialization in the case of mycoheterotrophic plants (specialists) and their fungal partners (also specialists). We even observed unexpected, extreme reciprocal specialization for some initially mycoheterotrophic lineages associating with fungi exclusively interacting with these plant lineages. Finally, we found that independently emerged mycoheterotrophic plant lineages share many closely related fungi, and that in some of these lineages fungal partners were likely acquired from autotrophic ancestors, while in others they were likely acquired by symbiont shift, suggesting different evolutionary pathways leading to mycoheterotrophy.

Cheaters are isolated by reciprocal specialization:

We confirmed that mycoheterotrophic plants are more specialized toward few mycorrhizal fungal partners than autotrophic plants (Merckx *et al.*, 2012) and showed for the first time that their fungal partners are overall more specialized than fungi associated with autotrophic plants. This reciprocal specialization is not strict (with the exception of Lycopodiaceae, see below), since mycoheterotrophs and their fungal partners need some connection to autotrophic plants, yet sufficient to lower nestedness in the arbuscular mycorrhizal network. The observed trend toward reciprocal specialization and reduced nestedness suggests that mycoheterotrophic cheating is an unstable ecological and evolutionary strategy, which could explain the relatively recent origin of mycoheterotrophic clades (Figure III.6.2). Indeed, reciprocal specialization confers high extinction risks for both interacting partners, which is one of the main hypotheses explaining why mutualistic networks tend to be nested, with asymmetrical specialization (*i.e.* specialists interact with generalist partners; Thébault & Fontaine, 2010). Whatever its origin, the reciprocal specialization of cheaters and their partners has also been suggested in other mutualisms (Genini *et al.*, 2010). A parasitic nature of entirely mycoheterotrophic plants has often been mooted (Bidartondo, 2005; Merckx, 2013), albeit without direct support, in the absence of data on fitness of fungal partners and autotrophic plants providing carbon to mycoheterotrophs (van der Heijden *et al.*, 2015). Our analysis *a posteriori* supports the view of entirely mycoheterotrophic plants as parasitic cheaters. However, we cannot exclude the possibility that mycoheterotrophs might provide some advantages to their mycorrhizal fungi (*e.g.* shelter or vitamins; Brundrett, 2002; Selosse & Rousset, 2011), making them useful partners for some specific fungal species, despite their carbon cost. Further empirical evidence is needed to clarify this.

There are several not mutually exclusive explanations for this reciprocal mycoheterotrophic specialization. First, physiological constraints may act if conditional investment and partner choice occur in the mycorrhizal symbiosis (Kiers *et al.*, 2011), meaning that

each partner would preferentially interact with the most mutualistic of the many partners they encounter in soil. Mycoheterotrophic cheaters might have been able to successfully avoid these constraints by specifically targeting a few specific fungi susceptible to mycoheterotrophy, with which they now interact in specialized parasitism (Selosse & Rousset, 2011). Regarding the fungi, we can speculate that 'cheated' fungi that provide mycoheterotrophs with carbon entail a greater carbon cost for autotrophic plants than other fungi, and that autotrophic plants therefore tend to avoid interactions with these fungi. This would result in a trend to reciprocal specialization, and the partial isolation of mycoheterotrophic cheaters and their fungal partners from the mutualistic network. Second, the pattern of reciprocal specialization could result from physiological traits of the fungal species, as yet unknown to us, which make them more likely to be avoided by autotrophic plants and to associate with mycoheterotrophic plants. Third, such a pattern of reciprocal specializations could also come from ecological constraints limiting the niches and habitats of mycoheterotrophic plants. Indeed, mycoheterotrophic plants often tend to occur specifically in patches of low soil fertility (Gomes *et al.*, 2019). It is important to acknowledge that although the global pattern of reciprocal specialization observed in the present work is likely to be linked to cheating, it might also be influenced by the specific local environmental conditions where cheating is promoted. For instance, because mycoheterotrophs primarily persist in these low fertility habitats where access to essential mineral nutrients for autotrophic plants is limiting, we can speculate that it might still be advantageous for autotrophic plants to interact with poorly cooperative fungal partners associated with mycoheterotrophs, which provide less mineral nutrient in relation to their carbon cost. Additionally, low nutrient availability in the environments of mycoheterotrophs might also limit the available pool of mycorrhizal fungi: the relative specialization of mycoheterotrophic plants could be the consequence of low availability of fungal partners in these specific habitats. Yet, there is ample evidence that mycoheterotrophic species are specialized on one or few fungi in various environments from all over the world, where several to many suitable fungi should also be available. For instance, in a similar symbiosis, mycoheterotrophic orchids specialize on few saprotrophic fungi in tropical forests where many saprotrophic fungi occur (Martos *et al.*, 2009).

An in-depth sampling of mycorrhizal networks (particularly weighted networks) in various local communities containing mycoheterotrophs would be required to test whether reciprocal specialization occurs at the local scale and will shed more light on the mechanisms regulating the interaction. Indeed, we observed a trend to reciprocal specialization in a large-scale interaction network compiled from mycorrhizal interactions described in different ecosystems around the world, not in locally described physical mycelial networks. This allowed us to analyze a global ecological pattern, representing the complete evolutionary history of the partners, and is justified by the very low endemism of arbuscular mycorrhizal fungi and thus the absence of strong geographic structure (Davison *et al.*, 2015; Savary *et al.*, 2018). It is noteworthy that similar patterns of specialization were found in the African and South American networks (Supplemen-

tary Figure 4). On the other hand, a species may appear to be relatively more specialized in a global network than it actually is in local communities.

Our rarefaction analyses indicated that including more mycoheterotrophic species in this dataset should reveal more fungal species associated with mycoheterotrophs. Yet, given that our dataset covers almost all mycoheterotrophic families and that our results are robust to the sampling fraction of mycoheterotrophs and their associated fungi (Supplementary Figure 2), we expect the unsampled fungi associated with unsampled mycoheterotrophs to be phylogenetically related and specialists to the same degree as the sampled fungi associated with sampled mycoheterotrophs. A low sampling fraction of fungi associated with mycoheterotrophic plants is even expected given the trend of reciprocal specialization: as mycoheterotrophic species tend to be specialists interacting with specialist fungi, we would need to sample most of the mycoheterotrophic species to obtain most of their specialist associated fungi.

In this study, we used a simple dichotomy of plants considered either as mutualistic autotrophs or as (either entirely or initially) mycoheterotrophic cheaters. However, mycoheterotrophy is not the only uncooperative strategy in this symbiosis: mycorrhizal interactions rather represent a continuum between mutualism and parasitism, both in terms of plants (Jacquemyn & Merckx, 2019) and fungi (Johnson *et al.*, 1997; Klironomos, 2003). Physiological constraints are thus thought to constitutively maintain the stability of the mycorrhizal symbiosis (Kiers *et al.*, 2003, 2011) against many forms of cheating, including the specific case of mycoheterotrophy. Moreover, we did not consider context dependency, which has a non-negligible impact on the functioning of mycorrhizal interactions (Chaudhary *et al.*, 2016). Although the mutualism-parasitism continuum or the context dependency could have hidden the observed patterns, the fact that we observed significant differences in the specialization between autotrophic and mycoheterotrophic plants and high similarities between sets of fungal partners associated with different mycoheterotrophic plant lineages suggests that the observed patterns are likely robust to our simplifications.

Independent emergences of entirely mycoheterotrophic cheating converge on closely related susceptible fungi :

Mycoheterotrophic cheating emerged multiple times in different clades of the phylogeny of vascular land plants, indicating weak phylogenetic constraints. This likely results from the low specificity in arbuscular mycorrhizal symbiosis, which allows convergent interactions (Bittleston *et al.*, 2016) in different plant clades. Such convergences would have happened during the evolution of mycoheterotrophic plants with similar fungi susceptible to cheating. Thus, physiological or ecological constraints leading to reciprocal specialization appear to be the main barrier to the emergence of cheating in arbuscular mycorrhizal mutualism.

There are, however, phylogenetic constraints on the fungal side. We found few fungal clades that interacted with independent mycoheterotrophic plant lineages, and these clades were phylogenetically related, as already reported by Merckx *et al.* (2012); accordingly, fungal partners associated with mycoheterotrophs seem to be less phylogenetically diverse than those associated with autotrophic plants. The physiological traits that underlie variation in susceptibility of fungi to mycoheterotrophy remain unclear (Chagnon *et al.* 2013; van der Heijden & Scheublin 2007) and obtaining more information on fungal functional traits would greatly improve our understanding of mycoheterotrophic systems, the habitat distribution of mycoheterotrophs and their associated fungi, and what make fungi susceptible to mycoheterotrophy or not. Studying the functional traits of susceptible fungi, which are exceptions to the widespread avoidance of non-cooperative partners (Selosse & Rousset, 2011), will be particularly useful for understanding how fungi avoid cheating.

The acquisition of susceptible fungi depends on the mycoheterotrophic plant lineage. In some mycoheterotrophic lineages, such as Burmanniaceae, fungal partners were closely related to the fungal partners of autotrophic relatives, suggesting that the fungi associated with mycoheterotrophs are derived from the fungal partners of cooperative autotrophic ancestors. In other mycoheterotrophic lineages, such as Gentianaceae or Corsiaceae, fungal partners were more closely related to fungal partners of other mycoheterotrophic lineages than to autotrophic relatives, suggesting that the fungi associated with mycoheterotrophs were acquired secondarily rather than derived from the partners of autotrophic ancestors. A few mycoheterotrophic plant lineages lacked closest autotrophic relatives in our analysis (*e.g.* mycoheterotrophic Gentianaceae should be compared to autotrophic Gentianaceae, not represented in the MaarjAM database), which may bias our analyses towards supporting secondary transfer from other mycoheterotrophic plants rather than acquisition from autotrophic ancestors. Still, similar fungi were found in mycoheterotrophic Burmanniaceae and their closest autotrophic relative after a 110-Myr-old divergence, while mycoheterotrophic Gentianaceae and their closest autotrophic relative have distinct fungal partners after a divergence of only 52 Myr.

Interestingly, all entirely mycoheterotrophic families are evolutionarily relatively recent: the oldest monocotyledonous entirely mycoheterotrophic families, such as Burmanniaceae and Triuridaceae, are only 110-130 Myr old, and the dicotyledonous entirely mycoheterotrophic families Gentianaceae and Polygalaceae are even more recent (around 50-60 Myr; Figure III.6.2). The oldest mycoheterotrophic families show conservatism for fungal partners, while the most recently evolved ones display secondary acquisition. We can speculate that mycoheterotrophy initially emerged in the monocotyledons thanks to suitable cheating-susceptible fungal partners; more recently evolved entirely mycoheterotrophic lineages (especially in dicotyledons) then convergently reutilized these fungal partners. Complementary analyses including more sampling of the mycoheterotrophic families and their closest autotrophic relatives would be needed to test this speculation.

Independent networks and parental nurture in initially mycoheterotrophs:

Our results serendipitously revealed that two initially mycoheterotrophic families, Ophioglossaceae and Lycopodiaceae, seem to have exclusive mycorrhizal associations, as they interacted with fungi that did not interact with any other plant family. In these families, the fungi are present during both mycoheterotrophic underground spore germination and in the roots of adult autotrophic individuals (Winther & Friedman, 2007, 2008). Autotrophic adults likely act as the carbon source (Field *et al.*, 2015), part of which is dedicated to the offspring. This further supports the hypothesis by Leake *et al.* (2008) proposing parental nurture where germinating spores would be indirectly nourished by surrounding conspecific sporophytes. Parental nurture is not universal to all initially mycoheterotrophic families though; in the initially mycoheterotrophic Psilotaceae, for example, fungal partners are shared with surrounding autotrophic plants (Winther & Friedman, 2009). In initially mycoheterotrophic independent networks, the overall outcome for the fungus over the plant lifespan may actually be positive: fungi invest in mycoheterotrophic germinations that represent future carbon sources (Field *et al.*, 2015). In other words, initially mycoheterotrophic plants do not cheat their exclusive fungi, but postpone the reward. We note, however, that the existence of independent networks for these families should be confirmed in studies of local communities.

We found an extreme reciprocal specialization between Lycopodiaceae and a single *Glomus* clade. More studies are required to confirm that this pattern does not result from undersampling of the fungi interacting with these Lycopodiaceae species. Unlike other early-diverging plant clades that tend to interact with early-diverging fungal clades, the Lycopodiaceae (250-Myr-old) associate with a 49-Myr-old clade that diverged 78 Myr ago from all other *Glomus* (Rimington *et al.*, 2018). Thus, this highly specific interaction results from a secondary acquisition: some species of Lycopodiaceae may have initially developed mycoheterotrophic interactions with a wider set of fungi, and later evolved into a specific mutualistic parental nurture with their exclusive fungi, raising the possibility of co-evolution between both clades.

Conclusion:

Our analysis of mycoheterotrophy in arbuscular mycorrhizal symbiosis illustrates a globally mutualistic system where cheaters tend to be limited by reciprocal specialization. Such reciprocal specialization between mycoheterotrophic cheaters and their 'cheating-susceptible' partners, potentially due to partner choice, sanctions, and/or habitat restrictions, reduces nestedness in the network. Phylogenetic constraints occur on the fungal but not the plant side, as independently emerged mycoheterotrophic families convergently interact with closely related fungi. In addition, our results challenge the general cheater status of mycoheterotrophy, highlighting a dichotomy between true mycoheterotrophic cheaters and possibly cooperative, initially mycoheterotrophic systems with parental nurture. Beyond mycorrhizal symbiosis, we invite the use of our combina-

tion of network and phylogenetic approaches to evaluate the nature of constraints upon cheating in other multiple-partner mutualisms (*e.g.* pollination or seed dispersal).

References:

Bascompte J, Jordano P. 2013. Mutualistic networks. Princeton: Princeton University Press.

Bascompte J, Jordano P, Melián CJ, Olesen JM. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences of the United States of America 100: 9383–9387.

Beckett SJ. 2016. Improved community detection in weighted bipartite networks. Royal Society Open Science 3: 140536.

Bidartondo MI. 2005. The evolutionary ecology of myco-heterotrophy. New Phytologist 167: 335–352.

Bittleston LS, Pierce NE, Ellison AM, Pringle A. 2016. Convergence in multispecies interactions. Trends in Ecology and Evolution 31: 269–280.

Blüthgen N, Menzel F, Hovestadt T, Fiala B, Blüthgen N. 2007. Specialization, constraints, and conflicting interests in mutualistic networks. Current Biology 17: 341–346.

Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis (A Prlic, Ed.). PLoS Computational Biology 10: e1003537.

Boullard B. 1979. Considérations sur la symbiose fongique chez les Ptéridophytes. Syllogeus 19 : 1-58.

Bronstein JL. 2015. Mutualism (JL Bronstein, Ed.). Oxford: Oxford University Press.

Bronstein JL, Wilson WG, Morris WF. 2003. Ecological dynamics of mutualist/antagonist communities. The American Naturalist 162: S24–S39.

Brundrett MC. 2002. Coevolution of roots and mycorrhizas of land plants. New Phytologist 154: 275–304.

Chagnon P-LL, Bradley RL, Klironomos JN. 2012. Using ecological network theory to evaluate the causes and consequences of arbuscular mycorrhizal community structure. New Phytologist 194: 307–312.

Chagnon P-L, Bradley RL, Maherali H, Klironomos JN. 2013. A trait-based framework to understand life history of mycorrhizal fungi. Trends in Plant Science 18: 484–491.

Chaudhary VB, Rúa MA, Antoninka A, Bever JD, Cannon J, Craig A, Duchicela J, Frame A, Gardes M, Gehring C, *et al.* 2016. MycoDB, a global database of plant response to mycorrhizal fungi. Scientific Data 3: 160028.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. InterJournal Complex Systems Complex Sy: 1695.

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, *et al.* 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349: 970–973.

Dormann CF, Gruber B, Fründ J. 2008. Introducing the bipartite package: analysing ecological networks. R News 8: 8–11.

Douglas AE. 2008. Conflict, cheats and the persistence of symbioses. New Phytologist 177: 849–858.

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792–1797.

Ferriere R, Bronstein JL, Rinaldi S, Law R, Gauduchon M. 2002. Cheating and the evolutionary stability of mutualisms. Proceedings of the Royal Society B: Biological Sciences 269: 773–780.

Field KJ, Leake JR, Tille S, Allinson KE, Rimington WR, Bidartondo MI, Beerling DJ, Cameron DD. 2015. From mycoheterotrophy to mutualism: Mycorrhizal specificity and functioning in *Ophioglossum vulgatum* sporophytes. New Phytologist 205: 1492–1502.

Fontaine C, Guimarães PR, Kéfi S, Loeuille N, Memmott J, van der Putten WH, van Veen FJF, Thébault E. 2011. The ecological and evolutionary implications of merging different types of networks. Ecology Letters 14: 1170–1181.

Frederickson ME. 2013. Rethinking mutualism stability: Cheaters and the evolution of sanctions. Quarterly Review of Biology 88: 269–295.

Frederickson ME. 2017. Mutualisms are not on the verge of breakdown. Trends in Ecology and Evolution 32: 727–734.

Genini J, Morellato LPC, Guimarães PR, Olesen JM. 2010. Cheaters in mutualism networks (I Bartomeus, Ed.). Biology Letters 6: 494–497.

Gomes SIF, van Bodegom PM, Merckx VSFT, Soudzilovskaia NA. 2019. Environmental drivers for cheaters of arbuscular mycorrhizal symbiosis in tropical rainforests. New Phytologist 223: 1575–1583.

Gómez JM, Verdú M, Perfectti F. 2010. Ecological interactions are evolutionarily conserved across the entire tree of life. Nature 465: 918–921.

Gotelli NJ, Rohde K. 2002. Co-occurrence of ectoparasites of marine fishes: a null model analysis. Ecology Letters 5: 86–94.

van der Heijden MGA, Martin FM, Selosse MA, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytologist 205: 1406–1423.

van der Heijden MGA, Scheublin TR. 2007. Functional traits in mycorrhizal ecology: their use for predicting the impact of arbuscular mycorrhizal fungal communities on plant growth and ecosystem functioning. New Phytologist 174: 244–250.

Jacquemyn H, Merckx VSFT. 2019. Mycorrhizal symbioses and the evolution of trophic modes in plants (R Shefferson, Ed.). Journal of Ecology 107: 1567–1581.

Joffard N, Massol F, Grenié M, Montgelard C, Schatz B. 2019. Effect of pollination strategy, phylogeny and distribution on pollination niches of Euro-Mediterranean orchids (I Bartomeus, Ed.). Journal of Ecology 107: 478–490.

Johnson NC, Graham JH, Smith FA. 1997. Functioning of mycorrhizal associations along the mutualism-parasitism continuum. New Phytologist 135: 575–586.

Jones EI, Afkhami ME, Akçay E, Bronstein JL, Bshary R, Frederickson ME, Heath KD, Hoeksema JD, Ness JH, Pankey MS, *et al.* 2015. Cheaters must prosper: Reconciling theoretical and empirical perspectives on cheating in mutualism (N van Dam, Ed.). Ecology Letters 18: 1270–1284.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26: 1463–1464.

Kiers ET, Duhamel M, Beesetty Y, Mensah JA, Franken O, Verbruggen E, Fellbaum CR, Kowalchuk GA, Hart MM, Bago A, *et al.* 2011. Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. Science 333: 880–882.

Kiers ET, Rousseau RA, West SA, Denlson RF. 2003. Host sanctions and the legume-rhizobium mutualism. Nature 425: 78–81.

Klironomos JN. 2003. Variation in plant response to native and exotic arbuscular mycorrhizal fungi. Ecology 84: 2292–2301.

Lallemand F, Gaudeul M, Lambourdière J, Matsuda Y, Hashimoto Y, Selosse MA. 2016. The elusive predisposition to mycoheterotrophy in Ericaceae. The New phytologist 212: 314–319.

Leake JR. 1994. The biology of myco-heterotrophic ('saprophytic') plants. New Phytologist 127: 171–216.

Leake JR, Cameron DD, Beerling DJ. 2008. Fungal fidelity in the myco-heterotroph-to-autotroph life cycle of Lycopodiaceae: A case of parental nurture? New Phytologist 177: 572–576.

Lozupone C, Knight R. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71: 8228–8235.

Martos F, Dulormne M, Pailler T, Bonfante P, Faccio A, Fournel J, Dubois MP, Selosse MA. 2009. Independent recruitment of saprotrophic fungi as mycorrhizal partners by tropical achlorophyllous orchids. New Phytologist 184: 668–681.

Martos F, Munoz FF, Pailler T, Kottke I, Gonneau C, Selosse MA. 2012. The role of epiphytism in architecture and evolutionary constraint within mycorrhizal networks of tropical orchids. Molecular Ecology 21: 5098–5109.

Merckx VSFT. 2013. Mycoheterotrophy: An Introduction. In: Merckx VSFT, ed. Mycoheterotrophy. New York, NY: Springer New York, 1–17.

Merckx VSFT, Janssens SB, Hynson NA, Specht CD, Bruns TD, Smets EF. 2012. Mycoheterotrophic interactions are not limited to a narrow phylogenetic range of arbuscular mycorrhizal fungi. Molecular Ecology 21: 1524–1532.

Michonneau F, Brown JW, Winter DJ. 2016. rotl: an R-package to interact with the Open Tree of Life data (R Fitzjohn, Ed.). Methods in Ecology and Evolution 7: 1476–1481.

Oksanen J, Kindt R, Pierre L, O'Hara B, Simpson GL, Solymos P, Stevens MH. HH, Wagner H, Blanchet FG, Kindt R, *et al.* 2016. vegan: Community Ecology Package, R-package version 2.4-0. R-package version 2.2-1.

Öpik M, Davison J, Moora M, Zobel M. 2014. DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. Botany 92: 135–147.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist 188: 223–241.

Öpik M, Zobel M, Cantero JJ, Davison J, Facelli JM, Hiiesalu I, Jairus T, Kalwij JM, Koorem K, Leal ME, *et al.* 2013. Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. Mycorrhiza 23: 411–430.

Pellmyr O, Huth CJ. 1994. Evolutionary stability of mutualism between yuccas and yucca moths. Nature 372: 257–260.

Pellmyr O, Leebens-Mack J, Huth CJ. 1996. Non-mutualistic yucca moths and their evolutionary consequences. Nature 380: 155–156.

Revell LJ. 2012. phytools: An R-package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rezende EL, Lavabre JE, Guimarães PR, Jordano P, Bascompte J. 2007. Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 448: 925–928.

Rich MK, Nouri E, Courty P-E, Reinhardt D. 2017. Diet of arbuscular mycorrhizal fungi: Bread and butter? Trends in Plant Science 22: 652–660.

Rimington WR, Pressel S, Duckett JG, Field KJ, Read DJ, Bidartondo MI. 2018. Ancient plants with ancient fungi: liverworts associate with early-diverging arbuscular mycorrhizal fungi. Proceedings of the Royal Society B: Biological Sciences 285: 20181600.

Roberts G, Sherratt TN. 1998. Development of cooperative relationships through increasing investment. Nature 394: 175–179.

Sachs JL, Russell JE, Lii YE, Black KC, Lopez G, Patil AS. 2010. Host control over infection and proliferation of a cheater symbiont. Journal of Evolutionary Biology 23: 1919–1927.

Savary R, Masclaux FG, Wyss T, Droh G, Cruz Corella J, Machado AP, Morton JB, Sanders IR. 2018. A population genomics approach shows widespread geographical distribution of cryptic genomic forms of the symbiotic fungus *Rhizophagus irregularis*. ISME Journal 12: 17–30.

Selosse MA, Richard F, He X, Simard SW. 2006. Mycorrhizal networks: des liaisons dangereuses? Trends in Ecology and Evolution 21: 621–628.

Selosse MA, Rousset F. 2011. The plant-fungal marketplace. Science 333: 828–829.

Sepp SK, Davison J, Jairus T, Vasar M, Moora M, Zobel M, Öpik M. 2019. Non-random association patterns in a plant–mycorrhizal fungal network reveal host–symbiont specificity. Molecular Ecology 28: 365–378.

Spatafora JW, Chang Y, Benny GL, Lazarus K, Smith ME, Berbee ML, Bonito G, Corradi N, Grigoriev I, Gryganskyi A, *et al.* 2016. A phylum-level phylogenetic classification of zygomycete

fungi based on genome-scale data. Mycologia 108: 1028–1046.

Strullu-Derrien C, Selosse MA, Kenrick P, Martin FM. 2018. The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. New Phytologist 220: 1012–1030.

Taudiere A, Munoz F, Lesne A, Monnet A-CC, Bellanger J-MM, Selosse MA, Moreau P-AA, Richard F. 2015. Beyond ectomycorrhizal bipartite networks: projected networks demonstrate contrasted patterns between early- and late-successional plants in Corsica. Frontiers in Plant Science 6: 681.

Thébault E, Fontaine C. 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329: 853–856.

Verbruggen E, van der Heijden MGA, Weedon JT, Kowalchuk GA, Rö-Ling WFM. 2012. Community assembly, species richness and nestedness of arbuscular mycorrhizal fungi in agricultural soils. Molecular Ecology 21: 2341–2353.

Werner GDA, Cornelissen JHC, Cornwell WK, Soudzilovskaia NA, Kattge J, West SA, Kiers ET. 2018. Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown. Proceedings of the National Academy of Sciences 115: 5229–5234.

Winther JL, Friedman WE. 2008. Arbuscular mycorrhizal associations in Lycopodiaceae. New Phytologist 177: 790–801.

Winther JL, Friedman WE. 2009. Phylogenetic affinity of arbuscular mycorrhizal symbionts in *Psilotum nudum*. Journal of Plant Research 122: 485–496.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506: 89–92.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/2QsCjEB` or by scanning:

Article 7: Fungal sharing, specialization, and structural distinctiveness in the plant root microbiomes of distantly related plant lineages

Authors: Benoît Perez-Lamarque[1,2], Rémi Petrolli[1], Christine Strullu-Derrien[1], Dominique Strasberg[3], Hélène Morlon[2], Marc-André Selosse[1,4], Florent Martos[1]

[1] Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum national d'histoire naturelle, CNRS, Sorbonne Université, EPHE, Université des Antilles; CP39, 57 rue Cuvier 75 005 Paris, France

[2] Institut de biologie de l'École normale supérieure (IBENS), École normale supérieure, CNRS, INSERM, Université PSL, 46 rue d'Ulm, 75 005 Paris, France

[3] Peuplements Végétaux et Bioagresseurs en Milieu Tropical, Université de La Réunion, UMR PVBMT, 97 400 Saint-Denis, La Réunion, France

[4] Department of Plant Taxonomy and Nature Conservation, University of Gdansk, Wita Stwosza 59, 80-308 Gdansk, Poland

Author contributions: All authors designed the study. BPL, HM, and FM performed the fieldwork; BPL and FM the molecular works; and BPL performed the analyses. BPL wrote a first draft of the manuscript.

Acknowledgments: The authors thank Jean-Marie Pausé, Benoît Lequette, Audrey Valery, and Claudine Ah-Peng for support during the field work. The Parc National de La Réunion authorized sampling in La Réunion (N° DIR-I-2019-149). Logistic support was provided by the field station of Marelongue, funded by the P.O.E., Reunion National Park and OSU Reunion. They also thank the Service de Systématique Moléculaire (UMS 2700) for technical support, as well as Chantal Griveau, Amélia Bourceret, Céline Bonillo, and Maarja Öpik. This work was supported by the Agence Nationale de la Recherche (ANR-19-CE02-0002). BPL was supported by a doctoral fellowship from the École Normale Supérieure de Paris attributed to BPL and the École Doctorale FIRE – Programme Bettencourt. HM acknowledges support from the European Research Council (grant CoG-PANDA).

Citation: Perez-Lamarque B, Petrolli R, Strullu-Derrien C, Strasberg D, Morlon H, Selosse MA, Martos F. Fungal sharing, specialization, and structural distinctiveness in the plant root microbiomes of distantly related plant lineages. ***in preparation***.

Abstract

Plant root microbiomes, and in particular the mycorrhizal and endophytic fungi they contain, play fundamental roles in plant nutrition and protection. The multiple fungal lineages involved in these symbiotic interactions are often shared between surrounding plant species, which form dense and complex networks of interactions in local communities. These plant-fungal networks are particularly important in the functioning of the entire community as they often allow resource movements and inter-plant communications. However, to what extent these fungi are shared *versus* specialized when the community includes very distantly related plant lineages remains unclear. Here, we used high-throughput sequencing of the 18S rRNA and ITS genes to identify the fungi colonizing plant roots. We therefore investigated the structure of the network of plant-fungus interactions in three local communities, including flowing plants, ferns, and clubmosses, in contrasted habitats across la Réunion island.

We found that the root endophytic microbiomes were dominated by five main mycorrhizal fungal lineages: Glomeromycotina, Mucoromycotina (Endogonales), Helotiales, Sebacinales, and Cantharellales. We noticed significant differences in the endophytic microbiota compositions across the three sampled communities, especially concerning the presence of Mucoromycotina, which appeared to be very environment dependent. In each local community, though the endophytic microbiota significantly cluster across plant taxonomic groups, we observed a lot of fungal sharing between surrounding plants, including between phylogenetically distant plants (*e.g.* clubmosses and flowering plants). We also showed that the level of specialization varies according to the fungal lineages, irrespectively of the environmental conditions: Plant-Glomeromycotina associations appeared to be the most recurrent but the least specialized interactions, resulting in networks with a nested structure, whereas Mucoromycotina and Cantharellales were more specialized and more sporadic in their interactions with plants, resulting in less connected and more modular networks, and Helotiales and Sebacinales present intermediate levels of specializations.

Our study looking exhaustively at endophytic interactions within local communities revealed the distinctiveness between the different plant-fungus symbioses, probably underpinned by their singular ecologies and evolutionary histories. It also showed that microbial sharing is widespread in local communities, even among distantly related plants, which questions the role these fungi can play in the communities and highlights the importance of considering networks of interactions rather than isolated macroorganisms and their associated microbes.

Keywords: plant microbiota, mycorrhiza, endophyte, ecological networks, specializations, mycoheterotrophy, .

Introduction:

Plant root microbiomes play fundamental roles in the functioning of plants: they contribute to their nutrition, improve their protection, and foster their development (Selosse *et al.*, 2004; Berendsen *et al.*, 2012; Philippot *et al.*, 2013; van der Heijden *et al.*, 2015). In particular, mycorrhizal fungi colonizing the roots of most plant species on Earth supply the plants with mineral matter gathered in the soil in exchange for plant-assimilated carbon (Smith & Read, 2008; Brundrett & Tedersoo, 2018). These fungi have also been a major driver of the emergence of land plants and their latter evolution in the past 400 million years (Selosse & Le Tacon, 1998; Field *et al.*, 2015; Strullu-Derrien *et al.*, 2018). During that time, several types of mycorrhizas evolved, which imply various fungal lineages (van der Heijden *et al.*, 2015), including the Glomeromycotina subphylum forming the widespread and ancestral arbuscular mycorrhiza, the Endogonales order (from the Mucoromycotina subphylum), and more recently several lineages among the Basidiomycota division (*e.g.* the orders Sebacinales, Cantharellales…) or the Ascomycota division (*e.g.* the orders Helotiales, Pezizales,…) forming the ectomycorrhizas, the ericoid or the orchid mycorrhizas (Brundrett & Tedersoo, 2018). While some fungal lineages became obligate plant-associated symbionts, like the Glomeromycotina, other mycorrhizal lineages, like the Mucoromycotina, the Sebacinales, or the Cantharellales, present a more diverse panel of ecologies, from saprophytes to obligate symbionts (Weiß *et al.*, 2016; Miyauchi *et al.*, 2020).

These main categories of mycorrhizas had been proposed more than a century ago thanks to pioneer microscopic observations (Frank, 1885; Smith & Read, 2008). However, they have been recently challenged by the advances of DNA sequencing technology that has revealed "out-of-the-textbooks" interactions (Hoysted *et al.*, 2018) and question the niches truly occupied by some fungi (Selosse *et al.*, 2018). Indeed, many fungi often colonize plant tissues in an endophytic niche, without forming a true (visible) mycorrhizal association; thus, it seems to exist an endophytic continuum between fully functional mycorrhizal interactions and saprophytic colonizations. For instance, herbaceous plants that typically associate with Glomeromycotina fungi had been found to be also colonized by "ectomycorrhizal" fungal lineages (Schneider-Maunoury *et al.*, 2020); and similarly, Mucoromycota fungi had been detected in epiphytic orchids (Novotná *et al.*, 2018). In other words, in local communities, the mycorrhizal fungi of some plants are frequently present as endophytes in the roots of neighboring plants (Selosse *et al.*, 2009). However, without proper experimental evidence (Hoysted *et al.*, 2020), such colonizations by mycorrhizal fungal lineages say nothing about the actual functionality of these interactions yet (are there any nutritional exchanges?), and we therefore simply refer to them as endophytic interactions.

In a local community, there is often no one-to-one interactions between a plant individual and its mycorrhizal fungus, but rather a complex and dense network linking dif-

ferent plant taxa and fungal lineages (Simard *et al.*, 1997; Verbruggen *et al.*, 2012). These multiple-partner plant-fungus networks, that result from partner selection on both sides (Kiers *et al.*, 2011; Werner & Kiers, 2015), allow the movement of carbohydrates between plants (Simard *et al.*, 1997) or even interindividual communication (Babikova *et al.*, 2013), which can be essential for the good functioning of the ecological communities. However, whether plant lineages that diverged hundreds of million years ago are homogenously sharing similar fungi or whether plant-associated fungi tend to segregate because of different nutrition strategies and/or evolutionary constraints has often been debated (Rimington *et al.*, 2018), but rarely investigated in the local communities, as many studies only sampled communities enriched in flowering plants (Sepp *et al.*, 2019), with few to no representing taxa of the "early-diverging" plant lineages, such as the bryophytes, lycopods, or ferns.

From the studied communities including mainly flowering plants, plant-fungus interactions appeared to be non-random (Vandenkoornhuyse *et al.*, 2003; Sepp *et al.*, 2019). Their specificity ranges from moderately specific interactions resulting in a lot of fungal sharing between plant species (Toju *et al.*, 2015; Sepp *et al.*, 2019) to very specific interactions (Toju *et al.*, 2016; Article 6). Patterns of plant-fungus interactions are often studied by using bipartite networks, and the structure of these resulting networks has been found to be nested (*i.e.* specialists interact with generalists and generalists form a core of interactions, (Jacquemyn *et al.*, 2011; Montesinos-Navarro *et al.*, 2012; Sepp *et al.*, 2019)) and/or modular (*i.e.* some subsets of species tend to form separated compartments, (Chagnon *et al.*, 2012; Martos *et al.*, 2012; Jacquemyn *et al.*, 2015)). Nestedness is a typical structure harbored by mutualistic networks (Bascompte *et al.*, 2003), providing greater stability (Thébault & Fontaine, 2010), while modularity observed in plant-fungus networks might instead be mediated by plant and fungal traits (Bahram *et al.*, 2014; Chagnon *et al.*, 2015). In addition, these plant-fungus interactions can be evolutionary conserved, and mycorrhizal networks then often exhibit significant phylogenetic signals, with distantly related species interacting with less similar partners than closely related ones (Jacquemyn *et al.*, 2011; Martos *et al.*, 2012; Tedersoo *et al.*, 2013; Montesinos-Navarro *et al.*, 2015). Furthermore, the structural properties of these flowering plant-fungus networks seem to significantly vary according to the fungal lineages and the environmental conditions (van der Heijden *et al.*, 2015; Põlme *et al.*, 2018; Rimington *et al.*, 2019).

In contrast, little is known about the mycorrhizal symbioses of early-diverging plants (Rimington *et al.*, 2018). Among these lineages, the Lycopodiaceae (or clubmosses), a vascular plant family that emerged more than 250 million years ago, have evolved particular mycorrhizal interactions specific to their alternation of generations: the diploid sporophyte is autotrophic, whereas the haploid gametophyte is usually achlorophyllous and underground (Boullard, 1979). Achlorophyllous gametophytes therefore rely on their associated mycorrhizal fungi for both their organic and mineral nutrition (Boullard, 1979; Winther & Friedman, 2008), a strategy referred to as mycoheterotrophy (Merckx, 2013).

Lycopods were mainly thought to interact with Glomeromycotina fungi (Schmid & Oberwinkler, 1993), but recent studies demonstrated that some species likely also associates with Mucoromycotina and Basidiomycota (Horn *et al.*, 2013; Rimington *et al.*, 2015). In a recent meta-analysis of plant-Glomeromycotina interactions at a global scale (Article 6), some lycopod species appeared to specifically interact with a distinct clade of Glomeromycotina, forming a separate module of interaction, which is very unusual in the arbuscular mycorrhizal symbiosis and could be due to parental nurture (Leake *et al.*, 2008). However, whether this strong specialization in lycopod-fungus interactions also exists in the local communities remains yet unknown.

Here, we studied the plant-fungus associations in three local communities with contrasted environmental conditions across La Réunion island. We investigated (i) what are the endophytic fungi colonizing the phylogenetically distant plant taxa in these contrasted habitats and (ii) how these fungi are shared between surrounding plants in each local community. We sampled roots of the main plant species in communities including bryophytes, lycopods, ferns, and flowering plants, and characterized their fungal partners by using metabarcoding technics targeting the 18S rRNA and ITS fungal marker genes. We reconstructed the endophytic networks of interactions between plants and fungi at the local scale, analyzed the structure of the interaction networks, and evaluated the degree of specialization of these plant-fungus interactions. We expected to find diverse fungal colonizations across the different plant taxa and predicted to see less fungal sharing between the phylogenetically distant plant lineages, resulting in lineage-specific fungus, phylogenetic signals in plant-fungus interactions, and potentially modular network structures.

Methods:

Study sites and sampling:

The study was conducted in La Réunion island in July 2019. In order to maximize the phylogenetic distances between the vascular plant species co-occurring in a sampling site, we chose three plant communities containing lycopod sporophytes (clubmosses), ferns and/or bryophytes, and flowering plants across contrasted habitats (Strasberg *et al.*, 2005): Grand brûlé (young lava flows close to the ocean with abundant non-indigenous plant species, S21°16'39'', E55°47'29''), Plaine-des-Palmistes (*Pandanus* wet thicket on old lava flows in the central valley, S21°07'08'', E55°38'36'', altitude 900 meters), and Dimitile (leeward rainforest on old lava flows of the crests of Cilaos circus, S21°16'39'', E55°47'29'', altitude 2,000 meters). These three communities thus represent diverse habitats with contrasted environmental conditions, especially in terms of altitude, disturbance, and humidity (Figure III.7.1).

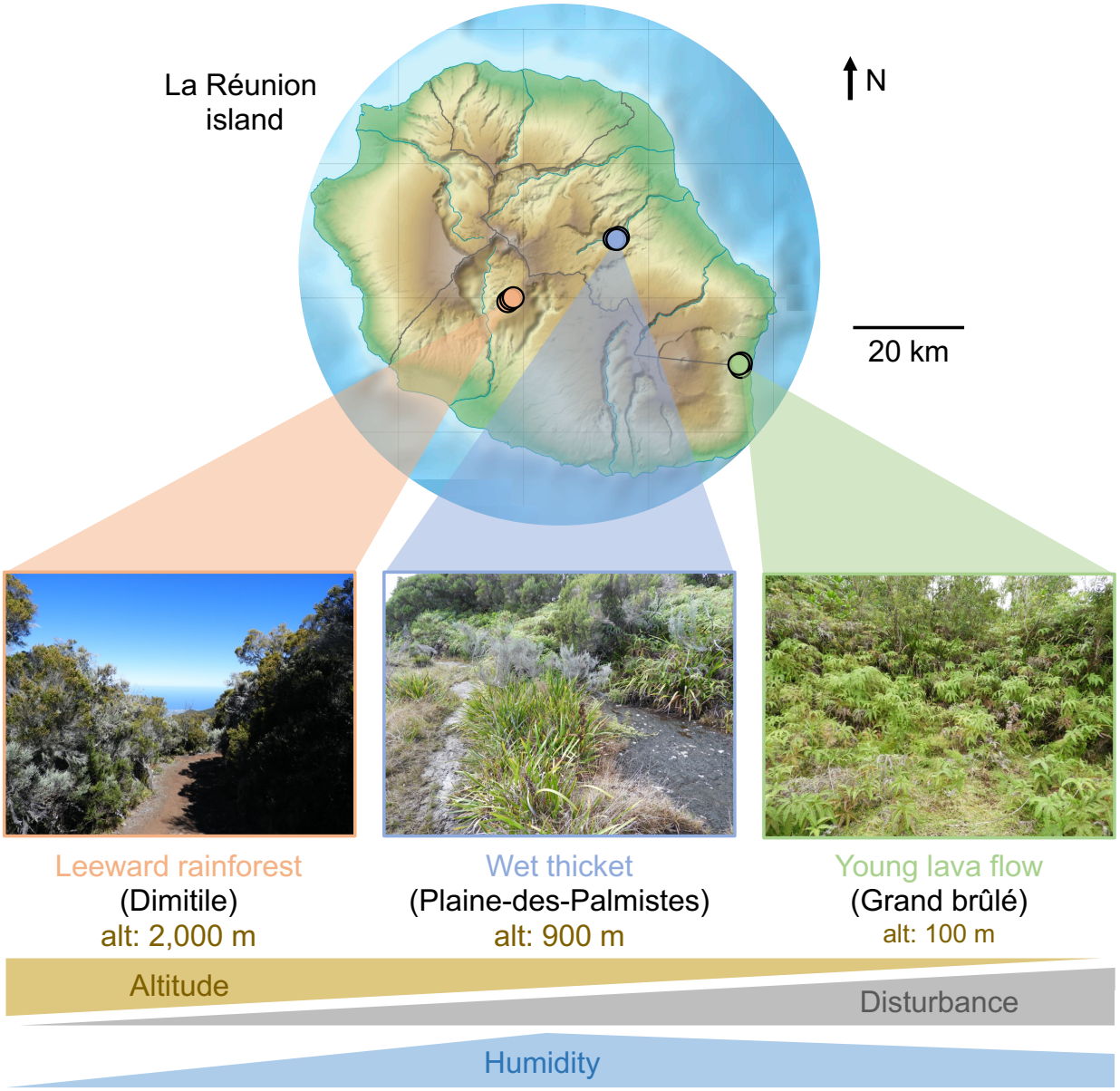


Figure III.7.1: The three sampled communities correspond to habitats with contrasted environment conditions. A map of La Réunion island indicating the three sampled communities in this study. The sampling sites were characterized by different vegetations and abiotic conditions with different altitudes, level of disturbance, humidity, and soil conditions: (right) Grand brûlé (young lava flows close to the ocean on the wet East coast, (middle) Plaine-des-Palmistes (*Pandanus* wet thicket on old lava flows in the central valley, altitude 900 meters), and (left) Dimitile (leeward rainforest on old lava flows in the dry crests of Cilaos circus on the West side dominated by ericoid vegetation, altitude 2,000 meters). In each sampled community, three replicates distant from 50 meters to 500 meters were sampled. The photos illustrate the overall vegetations in each sampled community and the gradients at the bottom resume the main variations in the environments.

In each community, we sampled 3 replicates distant from 50 meters to 500 meters; as lycopods tend to have patchy distributions in the communities, we specifically targeted zones where they were present. For each replicate, we harvested the roots of up to 3 individuals for each plant species in a radius of 1.5 meters. When possible, the whole root

systems were carefully removed from the soil, cleaned with water, and dried in silica gel. In Dimitile, some plant species (*Erica reunionensis*, *Phylica nitida*, and *Stoebe passerinoides* in particular) were too large to dig up the whole root systems: several soil cores were then collected and up to 23 individual roots per replicate were amassed without a direct species identification on the field (they were identified using DNA sequencing - see next section). Among the sampled plant species, we classified them as indigenous, exotic, or invasive species based on the Mascarine Cadetiana database (`https://mascarine.cbnm.org`).

Molecular analyses:

Dried roots were crushed using autoclaved tungsten beads in the TissueLyser II (Qiagen) and approximately 30 mg of root powder were kept for the downstream analyses. Plant and fungal DNA were extracted using the Genomic DNA Plant kit (NucleoSpin 96 Plant II, Macherey-Nagel) and following the manufacturer's instructions. Negative controls were carefully added during DNA extraction.

In order to better characterize the various fungi susceptible to colonize these different plant species, we amplified two nuclear regions of the rDNA operon, the 18S rDNA gene and the ITS2 region, using two sets of tagged primers: the AMADf-AMDGr pair (Berruti *et al.*, 2017) for 18S rRNA and the ITS86F-ITS4 pair (White *et al.*, 1990; Turenne *et al.*, 1999) for ITS2, that respectively amplify a fragment of 380 and 280 base pairs on average. The former marker gene rather detects Mucoromycta (including Glomeromycotina and Mucoromycotina), whereas the latter marker is more specific to Ascomycota and Basidiomycota. PCRs were performed using the AmpliTaq Gold™ 360 kit (Thermo Fisher Scientific) following the manufacturer's instructions and ran for a total of 35 cycles. Each sample was characterized by a unique combination of tagged primers (following Taberlet *et al.*, 2018; Petrolli *et al.*, 2021), and some combinations of primer pairs were left empty to evaluate the amount of tagged primer jumping between samples. In addition, negative PCR controls were also conducted (i) to identify the eventual external contaminants and (ii) to evaluate the baseline cross-contamination between samples during the library preparation. For each sample, the PCR step was replicated at least 2 times to ensure better characterization of the fungal microbiota.

For both libraries (constituted of either the 18S rRNA or the ITS amplicons), PCR products were purified using magnetic bead-based clean-up (NucleoMag™ NGS Clean-up and Size Select, Macherey-Nagel), quantified using Qubit (Qubit dsDNA High Sensitivity Assay Kit, Invitrogen), and 3 ng of DNA amplicons per sample and all the negative controls were sequenced using Illumina 2x250 bp MiSeq technology (v3 chemistry, Fasteris, Geneva).

Bioinformatics:

We obtained a total of 26,914,809 reads for 18S rRNA and 25,250,698 reads for ITS that were processed using a pipeline based on VSEARCH (Rognes *et al.*, 2016) available on GitHub (`https://github.com/BPerezLamarque/HOME/`). In short, pair reads were assembled, quality checked (removing reads that had on average more than 2 errors), demultiplexed (using cutadapt; Martin, 2011), and clustered into operational taxonomic units (OTUs) using two different methods. We first used Swarm v2 (Mahé *et al.*, 2015) with the fastidious option, a clustering approach that does not rely on a global threshold of similarity but instead uses local thresholds and amplicon abundances. Secondly, we performed a classical 97% OTU clustering using VSEARCH. We removed the chimera in both sets of OTUs using *uchime_denovo* in VSEARCH and assigned taxonomy to each OTU using *usearch_global* (BLAST algorithm). For the latter step, we used the Silva database r138 (Quast *et al.*, 2013) for the taxonomic assignations of the 18S rRNA OTUs and the UNITE database v8.2 (Nilsson *et al.*, 2019) for those of the ITS OTUs. Because the Silva database contains only a few Agaricomycetes sequences, we also assigned taxonomy at the order level to these Agaricomycetes 18S rRNA OTUs using the Fungal 18S RefSeq Targeted Loci Project (NCBI BioProject - PRJNA39195). We finally built an OTU table for each marker and removed OTUs present in less than 10 reads. Given that both Swarm and 97% OTU clustering gave qualitatively similar results, only the results obtained with Swarm OTUs are presented in the main text (those of 97% OTUs are in Supplementary information). Glomeromycotina characterized using the 18S rRNA marker are often assigned to a virtual taxa (VT) based on the MaarjAM database (Öpik *et al.*, 2010), however, the AMADf-AMDGr primer pair used in this study led to 18S rRNA amplicons that only partially overlap with virtual taxa and prevented us to do the assignations (Davison *et al.*, 2015).

The decontam pipeline in R (R Core Team, 2020) was applied to filter out the contaminants of our OTU tables, using the read abundances within the negative controls as well as the abundance profile of each OTU (Davis *et al.*, 2018). We evaluated the amount of primer jumping and cross-contaminations thanks to our range of negative controls (Supplementary Figure 1). We also verified that samples that were physically close in a plate (the organization of the plates was kept identical from the DNA extraction to the final pooling) did not tend to present a composition more similar because of cross-contaminations. To do so, the similarities of the composition were evaluated using Bray-Curtis dissimilarities computed using the *vegdist* function of the vegan R-package (Oksanen *et al.*, 2016) and were correlated with the Euclidian physical distances in each PCR plates using Mantel tests (p-value>0.05 indicated non-significant correlations). Finally, the 18S rRNA and ITS OTUs assigned to plant species were used to identify the roots directly collected in the soil in Dimitile. We only kept the fungal OTUs for the following analyses.

In order to identify endophytic fungi, we used FUNGuild (Nguyen *et al.*, 2016), a

program that automatically assigns the possible niches of a fungal OTU based on its taxonomic assignation. Thanks to the output of FUNGuild and manual filtering, we only retained the mycorrhizal OTUs and the endophytic OTUs that were likely mycorrhizal (*i.e.* the fungi that are mycorrhizal in some plant lineages and only endophytic in others). Note that an important proportion of these selected OTUs were potential saprotrophs or commensal endophytes. Thus, for simplicity, we thereafter simply referred to all these OTUs as "endophytic OTUs". Samples having less than 20 endophytic reads were discarded. We used bar plots to represent the relative abundances of each endophytic fungi in the different root samples. Next, we reconstructed the phylogenetic tree of all the endophytic OTUs: we aligned the OTU sequences using MAFFT (Katoh & Standley, 2013), trimmed them with trimAl (Capella-Gutierrez *et al.*, 2009), selected the best substitution model using ModelFinder (Kalyaanamoorthy *et al.*, 2017), and reconstructed the maximum-likelihood tree using IQ-TREE (Nguyen *et al.*, 2015) with 1,000 SH-aLRT and ultrafast bootstraps.

Measuring the influence of the sampling on plant-fungus interactions:

We first performed rarefaction analyses to see how the fungal diversity with each plant species increases as a function of the number of sampled plant individuals.

Second, we investigated whether we could merge the three sampling replicates within each community. To do so, we first compared the alpha diversity of the samples within each sampling replicate using the total OTU richness, Shannon index, or Faith's phylogenetic diversity. Second, we performed principal coordinate analyses (PCoA) and permutational analyses of variance (PERMANOVA; *adonis* function from the R-package vegan; Oksanen *et al.*, 2016) to investigate whether samples from the same species but different replicates tend to host similar endophytic fungi.

Third, to investigate the effect of the sampled community on the endophytic compositions, we used a PERMANOVA based on Bray-Curtis dissimilarities to test whether endophytic compositions were significantly different across the three sampled community when comparing (i) all the plant species or (ii) only the plant species simultaneously present in several sampled communities. We also visualized the similarities of the endophytic microbiota between pairs of samples by performing hierarchical clustering: We built the dendrogram between samples using neighbor joining (*nj* function in the R-package ape; Paradis *et al.*, 2004) based on the endophytic beta diversities.

Measuring the influence of the main plant taxonomic groups on plant-fungus interactions:

In each sampled community, we examined whether root samples belonging to the same plant taxonomic group (bryophytes, lycopods, ferns, monocots (excluding orchids), dicots (excluding ericaceous species), orchids, or ericaceous species) were colonized by

similar fungal OTUs, using both PERMANOVA and hierarchical clustering. Based on the endophytic composition of the different plant species (see Results), we also specifically replicated our PERMANOVA analyses on the 5 most abundant endophytic fungal groups: the Glomeromycotina phylum, the Endogonales order (hereafter referred to as Mucoromycotina), the Sebacinales order, the Helotiales order, and the Cantharellales order. The two former groups were characterized using the 18S rRNA marker, whereas the three latter groups were characterized using the ITS marker.
All the diversity analyses were replicated by using generalized UniFrac distances, a phylogenetically-informed diversity index computed using the R-package GUniFrac (Chen *et al.*, 2012). Given that this did not qualitatively changed our results, we only reported the results obtained with Bray-Curtis dissimilarities in the main text.

Finally, in each sampled community and for each fungal group, we evaluated whether closely related plant species tend to interact with similar fungi by measuring the phylogenetic signals in species interactions. To do so, we first reconstructed the phylogenetic trees of the plant species: the plant mega-phylogeny from (Zanne *et al.*, 2014) was pruned using Phylomatic (`http://phylodiversity.net/phylomatic/`) to obtain the phylogenetic tree of the plant species sampled in our three sampled communities. Plant taxa that were not identified at the species levels were added as polytomies at the origin of the clade. Next, to measure phylogenetic signals, following Article 4, we used Mantel tests to assess the Pearson correlation between plant phylogenetic distances and the weighted UniFrac distances measuring the dissimilarity of their sets of fungal partners. The significance of the correlation was evaluated using 10,000 permutations.

Reconstructing endophytic networks:

To reconstruct plant-fungus interaction networks, we needed first to decide what can be considered as 'a likely interaction' between a plant and an endophytic fungus. Following Toju *et al.* (2014), given that we had a heterogenous number of endophytic reads per sample and in order to avoid counting spurious interactions in samples with high coverage, we converted the read abundances into relative abundances and only considered that there was an interaction between a plant and a fungal OTU if the OTU was represented by at least 1% of the total endophytic reads of the root sample. In addition, based on our estimates of cross contaminations (Supplementary Figure 1), we considered that having less than 5 reads of an OTU within a sample was likely not a colonization but rather came from contamination. Preliminary analyses (not shown) using other cutoffs (*e.g.* 10 reads and 0.1%) did not qualitatively affect our results. Importantly, we chose to treat the problem of heterogenous numbers of reads using relative abundances over rarefactions, as recent studies showed that rarefactions can be misleading (McMurdie & Holmes, 2014) and that using relative abundances is less biased (McKnight *et al.*, 2019).

In each sampled community, we reconstructed the plant-fungus network for the 5 main endophytic fungal lineages (Glomeromycotina, Mucoromycotina, Sebacinales, Helo-

tiales, or Cantharellales) and considered 3 types of species-level networks: binary networks that do not consider interaction strengths and two types of weighted networks that differently account for interaction strengths. First, binary networks correspond to presence/absence (1/0) networks that indicate whether an interaction between one plant species and one fungal OTU has been found in at least one sample. Second, we considered abundance networks that are based on OTU read abundances within root samples: for a given plant-OTU interaction, we reported its relative abundance as the number of reads belonging to this OTU per thousand of endophytic reads colonizing the corresponding plant species (note that relative abundances were computed by sample before merging the sample per species, such that each sample has the same contribution to the species total abundances). However, relative read abundances are subject to PCR amplification biases or variations in the rDNA copy number and can therefore be a bad proxy for the true fungal abundances colonizing the roots (Toju *et al.*, 2014). Thus, we considered a third type of network, the incidence networks, that reports weighted interactions without directly using read abundances: for each plant-fungus interaction, it indicates the number of root samples in which the interaction had been found. To check that abundance and incidence networks gave similarly quantify plant-fungus interaction strengths, we measured the relationship between both using linear models.

Analyzing the structure of endophytic networks:

We investigated the overall structure of the endophytic network according to the fungal group and the sampled community. We first computed the connectance of each network (the percentage of realized interactions) and the checkerboard score (Cscore) that measures the mean partner avoidance of pairs of species in a binary network, *i.e.* a high Cscore indicates that plant species tend to avoid interacting with the same fungal OTUs. Then, we investigated whether plant-fungus networks were significantly nested, by computing the NODF2 index for binary networks and the weighted NODF index for weighted networks (*nested* function in the bipartite R-package; Dormann *et al.*, 2008). Finally, we performed modularity analyses using Newman's algorithm for binary networks and Beckett's algorithm for weighted networks (*computeModules* function in the bipartite R-package). Modularity algorithms search for the most modular structure in the network and output the modularity value (M), which corresponds to the number of interactions within modules divided by the total number of interactions (within and between modules).

The significance of these structural properties was evaluated using two null models. A null model relies on a randomization strategy of the original network that excludes a particular process of interest and thus provides null interaction networks generated in the absence of this process (Gotelli, 2000). The first null model, generated using the quasiswap algorithm (implemented in the *permut* function of R-package vegan; Oksanen *et al.*, 2016) keeps constant the connectance and the marginal sums (*i.e.* the total number of interactions per plant species or fungal OTUs). Thus, the quasiswap null model investi-

gates whether the structural properties of the network are conserved when plant-fungus interactions are randomly attributed based on the total availabilities of each interactor, with the additional constraint of keeping a similar connectance. The second null model shuffles the sample names and therefore randomly attributes the fungi associated with each root sample to a plant species. Thus, the shuffle-sample null model tests whether the emerging patterns in the species-level network comes from plant species properties and not sample properties (Toju *et al.*, 2014). 10,000 null models were computed using either the quasiswap or the shuffle-sample algorithms from our different original networks (the binary networks, the abundance networks, and the incidence networks).
By computing each index (*e.g.* nestedness, modularity...) for each null model and comparing their values to the ones of the original networks (Manly, 2018), we get p-values indicating whether the observed networks have significant structural properties. For instance, for the nestedness, if p-value$\leq$2.5% (2.5% of the null models have a higher or equal NODF value than the original one), the network is significantly nested; alternatively, if p-value$\geq$97.5%, it is significantly anti-nested.
Both null models were used to investigate the significance of the NODF values, checkerboard scores, and modularity values. In addition, we used the shuffle-sample null models to evaluate the significance of the connectance.

Evaluating the specialization of plant-fungus interactions:

Next, we evaluated the specialization of plant-fungus interactions in each endophytic network. In the following, we only used the abundance network as a proxy for weighted interactions: indeed, the incidence networks contained generally too little weighted information (only up to few samples per species) to represent a useful measure of weighted interactions in the following index of specializations.

We started by measuring the specialization of each plant species toward its fungal partners in each sampled community and for each fungal group. We first computed the normalized degree of each plant species, as the number of associated fungal partners divided by the total number of available partners (*ND* function in the bipartite R-package): this indicates whether a species tends to be specialist (degree close to 0) or generalist (degree close to 1). Second, we computed d' which measures the plant preferences for fungal partners (*dfun* function in the bipartite R-package; Blüthgen *et al.*, 2006): a d' value close to 0 indicates that the plant species interacts with the most abundant fungal partners available with little specificity, whereas a d' value close to 1 indicates that the plant species specifically interacts with partners irrespectively of the abundance of other fungi. From the d' values, we computed H2', which is a network-level measure of specialization (Blüthgen *et al.*, 2006).

For each network, the significance of the indices of specialization was evaluated by generating, from the original abundance network, 10,000 null models using the Patefield algorithm (Patefield, 1981; Blüthgen *et al.*, 2006; Dormann, 2011) implemented in

the *r2table* function: this null model algorithm keeps constant the marginal sums of the original network, and therefore allow us to test whether the observed patterns of specialization are similar when interactions are only constrained by species abundances; we then referred to them as the marginal null models. In addition, we also used the shuffle-sample null models (see the previous section).

Next, we performed motif analyses to check for differences in the patterns of interactions at the species-level: for each endophytic network, we computed the frequencies of the motifs containing between 2 and 5 species using the *mcount* function from the BMOTIF R-package (Simmons *et al.*, 2019) and compared motif frequencies in the different networks using PCoA. Finally, we particularly focused on the Lycopodiaceae species in each sampled community and compared their motif frequencies with those of the surrounding plant species to investigate whether they tend to be more associated with Lycopodiaceae.

Results:

Identifying endophytic interactions:

A total of 233 root samples successfully amplified fungi with an average coverage of 60,537 fungal reads per sample ($\pm$29,044) for ITS and 19,414 fungal reads per sample ($\pm$14,974) for 18S (after removal of the contaminants and plant reads; Supplementary Table 1). These reads respectively were clustered into 5,236 Swarm OTUs for ITS and 4,371 Swarms OTUs for 18S. When filtering the endophytic OTUs susceptible to be mycorrhizal with FUNGuild, we obtained 622 OTUs for ITS and 1,177 OTUs for 18S, with a coverage larger than 1,000 reads in most root samples (Supplementary Figure 2).

The two markers characterized different aspects of these 'endophytic' fungal microbiota (Figure III.7.2; Supplementary Figure 3). Indeed, the 18S rRNA marker successfully detected colonization by Glomeromycotina and Mucoromycotina (Endogonales) fungi but failed at precisely characterizing Basidiomycota (they were at best identified at the order levels) and Helotiales (Ascomycota) were not detected at all with this marker. Conversely, the ITS marker failed at detecting Mucoromycotina fungi (they were only detected when their abundances were very high in the root samples according to the 18S rRNA marker), but successfully amplified and identified Basidiomycota and Ascomycota, including the abundant Sebacinales, Helotiales, Cantharellales, and Agaricales orders. Colonization status by Glomeromycotina (presence or absence of at least one Glomeromycotina OTU in a sample) was generally very consistent across samples of the same plant species (Supplementary Table 2). To a lesser extent, the colonization status by Sebacinales and Helotiales was also quite regular (*e.g.* for *Lycopodiella cernua*), but the colonization seemed more facultative for other plant species (*e.g.* for *Dicranopteris linearis*; Supplementary Table 2). Conversely, Mucoromycotina and Cantharellales colonizations

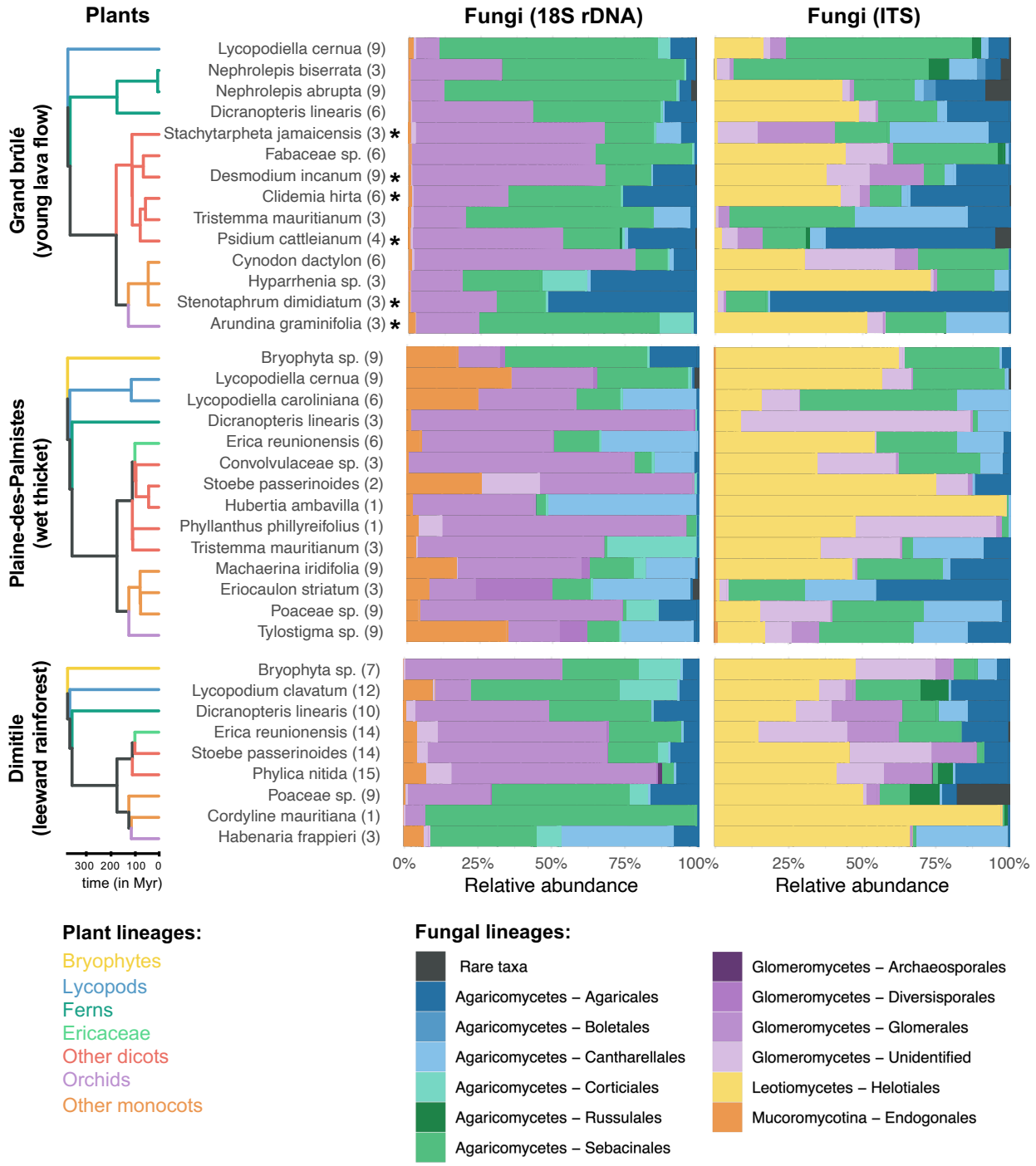


Figure III.7.2: Endophytic compositions vary according to the plant species and the habitats. For each plant species, the different root samples were merged and the relative abundances of the endophytic fungi are indicated according to the 18S rRNA (left) or ITS (right) markers. Plant species are separated according to the sampled community (Grand brûlé, Plaine-des-Palmistes, or Dimitile) covering contrasted habitats. For each species, the number of individual root systems sampled is indicated in brackets. In each sampled community, a phylogenetic tree of the plants is represented on the left, with branch colors indicating the main taxonomic groups we considered in our study. The bar plots represent in colors the class and the order of each endophytic fungus. Rare taxa (representing less than 0.5% of the data are represented in dark grey). Asterisks indicates the exotic species. Only results for the Swarm OTUs are represented (but 97% OTUs gave very similar results).

were recurrent in only a few plant species like *Lycopodiella caroliniana* or *Tylostigma sp.*, but were very sporadic in others (Supplementary Table 2). We noticed that both measures of interaction strength, using relative read abundance or interaction incidence, were significantly correlated (Supplementary Figure 4), suggesting that abundant interactions also tend to correspond to frequent ones.

We found that the root samples were very heterogeneous in term of alpha diversity, with some samples being very diverse, and other not (Supplementary Figure 5). The total numbers of endophytic OTUs in each sampled community varied between 48 and 92 (Supplementary Table 3): we noticed that the fungal diversity tend to be lower in Plaine-des-Palmistes, but it was not significantly different. In addition, rarefaction analyses indicated that the number of sampled individuals per species was not sufficient to get the entire diversity of fungi associated with most plant species, especially because of the important variability in composition between samples (Supplementary Figures 3 & 6).

The sampled communities influence patterns of endophytic interactions:

We found an important effect of the sampled communities on the endophytic microbiota compositions, revealed by the hierarchical clustering and the PCoA of all the samples, that both showed a clear clustering of the samples across the three sampled communities (Figure III.7.3; Supplementary Figure 9a-b; PERMANOVA: p-value<0.05). These shifts where also found when comparing the endophytic compositions of the plant species that were simultaneously sampled in different communities (Supplementary Figure 9c-d), suggesting that this different across sampled communities were not due to differences in plant species present, but likely to the different environmental conditions. For instance, Mucoromycotina were much more abundant in Plaine-des-Palmistes and rarer in other sampled communities (Figure III.7.2). When comparing the endophytic fungi present in the root samples of *Lycopodiella cernua* present in both Grand brûlé and Plaine-des-Palmistes, we found a significant shift in their composition, with enrichment in both Mucoromycotina (and Helotiales) in Plaine-des-Palmistes (Supplementary Figure 8).

Moreover, very local effects seemed also important in the assembly of the endophytic interactions as we also found an effect of the sampling replicates per community: samples from the same replicate tend to cluster together (Supplementary Figure 7b; PERMANOVA: p-value<0.05), which can slightly erase the signal of the plant taxonomic groups (Supplementary Figure 7a). However, given that this clustering per sampling replicate was moderate (especially when using UniFrac distances), we still merged the different replicate to performed the following analyses at the community level.

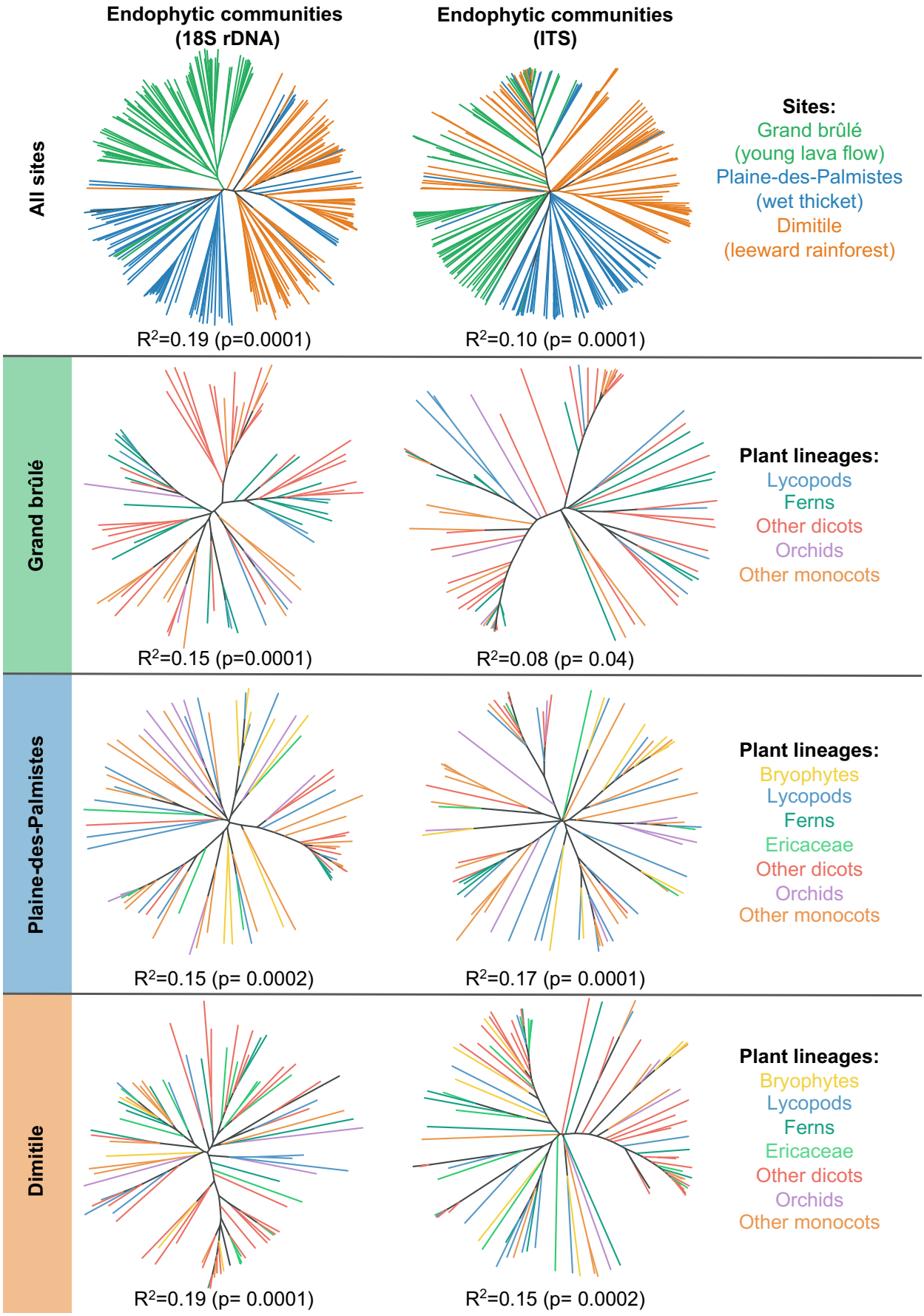

Endophytic communities (18S rDNA)
Endophytic communities (ITS)
All sites
Sites:
Grand brûlé (young lava flow)
Plaine-des-Palmistes (wet thicket)
Dimitile (leeward rainforest)
R²=0.19 (p=0.0001)
R²=0.10 (p= 0.0001)
Grand brûlé
Plant lineages:
Lycopods
Ferns
Other dicots
Orchids
Other monocots
R²=0.15 (p=0.0001)
R²=0.08 (p= 0.04)
Plaine-des-Palmistes
Plant lineages:
Bryophytes
Lycopods
Ferns
Ericaceae
Other dicots
Orchids
Other monocots
R²=0.15 (p= 0.0002)
R²=0.17 (p= 0.0001)
Dimitile
Plant lineages:
Bryophytes
Lycopods
Ferns
Ericaceae
Other dicots
Orchids
Other monocots
R²=0.19 (p= 0.0001)
R²=0.15 (p= 0.0002)

Figure III.7.3: Both the habitat and the plant taxonomic group influence the endophytic communities, despite a large heterogeneity due to fungal sharing across plant species. Dendrogram representations of the different endophytic communities across all the sampled communities (top row) or within each sampled community (bottom rows; Grand brûlé, Plaine-des-Palmistes, or Dimitile) based on the 18S rRNA (left) or ITS (right) markers. For each community, we computed the dissimilarity between pairs of samples (using Bray-Curtis distances) and reconstructed the dendrogram using neighbor joining: two plant root samples that are close in the dendrogram tend to have similar endophytic compositions. Branches are colored according to the sampled community (top row) or to the plant taxonomic group (bottom rows). Dendrograms were built using Swarm OTUs (but 97% OTUs gave very similar results). For each dendrogram, we also indicated the results of the PERMANOVA (R^2 and p-value based on 10,000 permutations) testing the effect of the sampled community (top row) or the plant taxonomic groups (bottom rows) on the endophytic Bray-Curtis beta diversity between root samples.

Plant taxonomic groups influence patterns of endophytic interactions:

Looking into details at the endophytic composition of the root samples (Figure III.7.2; Supplementary Figure 3), we found clear shifts in the endophytic associations according to the plant taxonomic groups. Interestingly, while ferns were mainly colonized by Glomeromycotina, Helotiales, and Sebacinales but not with Mucoromycotina, Mucoromycotina were regularly found across many plant species of others groups, including the two other early diverging lineages bryophytes and lycopods. In addition, lycopods were also frequently colonized by Sebacinales, Glomeromycotina, Helotiales, and even some Cantharellales (a few *Lycopodiella caroliniana* were abundantly colonized by a *Ceratobasidium*; Supplementary Figure 3). We noticed, that when Mucoromycotina were present in lycopods, Glomeromycotina were also generally present, resulting in frequent Mucoromycotina-Glomeromycotina dual symbioses (Hoysted *et al.*, 2018). Most samples from dicots and monocots were mainly associated with Glomeromycotina, and then with Helotiales, Sebacinales, and Cantharellales to a lesser extent. Besides their typical Helotiales partners, ericaceous species were also colonized in a significant proportion by Sebacinales, Glomeromycotina, and Cantharellales. Orchids were associated with Cantharellales and Sebacinales, but we also found some unexpected colonization by Mucoromycotina and Glomeromycotina. Such changes in the endophytic associations according to the plant taxonomic groups were confirmed using PERMANOVA (p-value<0.05; Figure 3; Supplementary Table 4) and visually detectable when using hierarchical clustering (Figure III.7.3) or PCoA (Supplementary Figure 7a) that both showed a trend of a clustering per plant taxonomic groups.

When studying separately each type of endophytic interaction, we also found that plant species from the same plant taxonomic group tend to interact with similar fungal OTUs (PERMANOVA; Supplementary Table 4), except for Cantharellales (as most of the OTUs are rather plant species-specific rather than shared at the plant taxonomic group level; see next section). The percentage of explained variance by plant taxonomic group tended to be lower for plant-Glomeromycotina interactions (R<0.15), than for other fun-

gal groups, probably because many Glomeromycotina OTUs were found across numerous samples irrespectively of their plant group (Supplementary Table 4).

Finally, within each sampled community, the presence of phylogenetic signals in plant-fungus interactions was scarce (Mantel tests: p-values>0.05; Supplementary Table 5). Indeed, closely related plants did not appear to significantly interact with similar Glomeromycotina, Mucoromycotina, Sebacinales, or Cantharellales OTUs. Conversely, closely-related plant species tend to associate with similar Helotiales OTUs and with a similar number of Sebacinales and Helotiales OTUs. This suggests that the changes in the plant-associated endophytic composition are better explained by the discrete shifts of the main plant taxonomic group, than by a continuous function of evolutionary time (as measured by Mantel tests).

Structure of the endophytic networks:

The reconstructed species-level networks for the five main endophytic fungal groups (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, or Cantharellales) resulted in networks of different sizes, reflecting their variability in terms of interactions with plants (Figure III.7.4; Supplementary Figure 10). Plant-Glomeromycotina networks visually presented species-rich, well-connected, typical nested structures with a core of abundant generalists surrounding by rare specialists, whereas plant-Mucoromycotina and plant-Cantharellales networks appeared to be species-poor, less connected, and much more modular, and plant-Sebacinales and plant-Helotiales networks had intermediate topologies (Figure III.7.4). When looking at the position of the different plant species in the networks, we noticed that plants from the same taxonomic group tend to be closer (Figure III.7.4), as previously indicated by the hierarchical clustering and PERMANOVA (Figure III.7.3; Supplementary Table 4). However, this clustering was limited and lycopods, ferns, and bryophytes appeared to be generally well connected by shared fungi to flowering plants (monocots and dicots; Figure III.7.4; Supplementary Figure 10). Conversely, orchids and ericaceous species tend to often form different modules, separated from other species forming the interaction core (see for instance the plant-Sebacinales network in Grand brûlé or the plant-Mucoromycotina network in Plaine-des-Palmistes). Details about the fungal genera involved in these particular interactions can be seen in Supplementary Figure 11: We noticed that Glomeromycotina OTUs were mainly composed of Glomeraceae (especially the widespread *Glomus* and *Rhizophagus* genera) and that Mucoromycotina were mainly represented by *Endogone*. In contrast, Sebacinales were represented by both *Sebacina* (Sebacinaceae) and *Serendipita* (Serendipitaceae), and many plants simultaneously interacted with both fungal genera (Supplementary Figure 11). Similarly, Helotiales and Cantharellales OTUs corresponded to several fungal families, but many of them appeared to be specifically restricted to certain plant species only (Supplementary Figure 11). Lastly, when looking at the exotic and invasive plants, we noticed that these species were relatively well integrated into the endophytic networks (a lot of fungal sharing), even if many of them were also associated with specific fungi,

therefore tending to be on the edges of the network representations (Supplementary Figure 12).

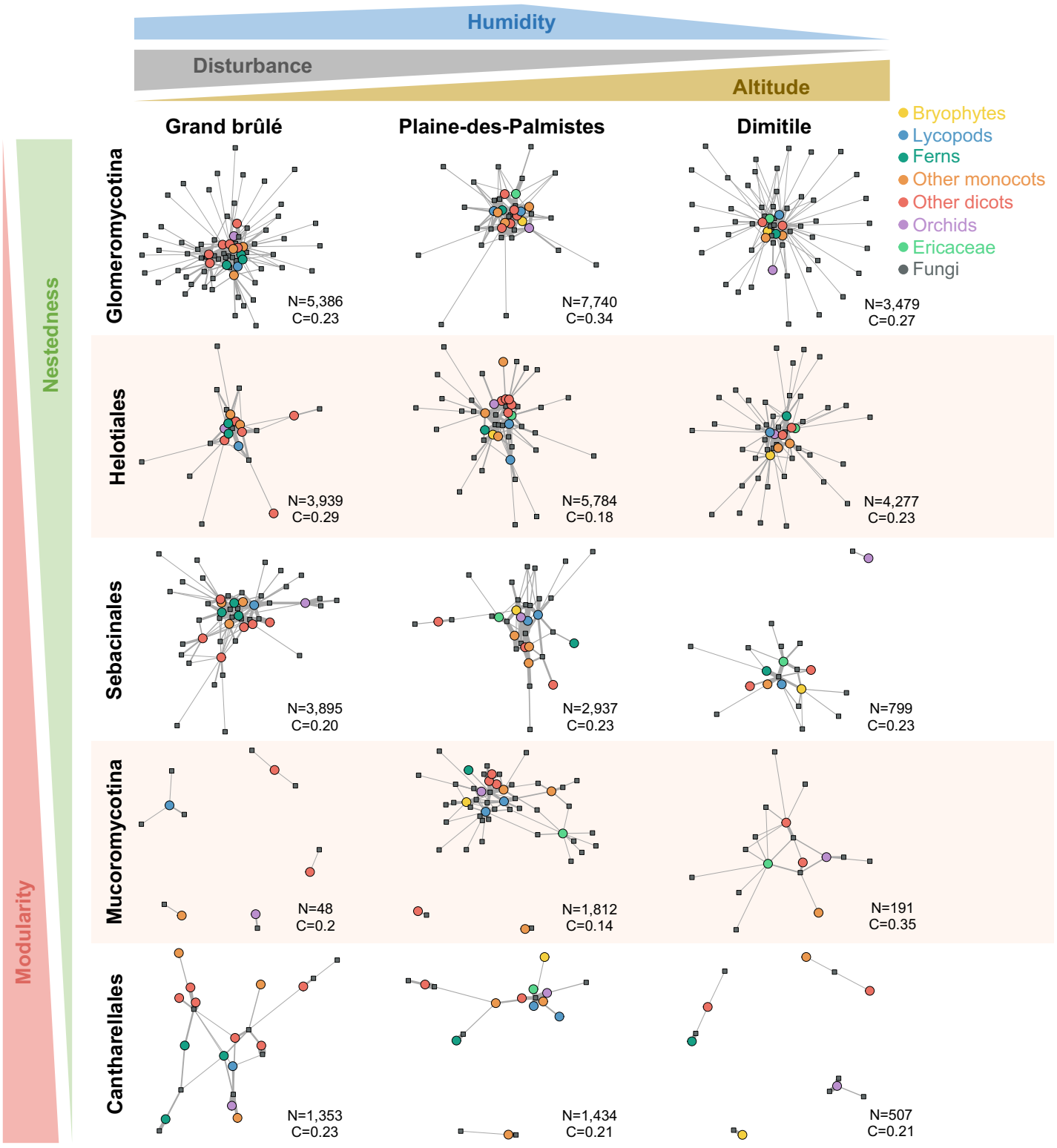


Figure III.7.4: The different fungal lineages influence more the plant-fungus network structures than the sampled communities, despite large habitat variations. Network representation at the species-level of the different endophytic networks (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, or Cantharellales) in each sampled community (Grand brûlé, Plaine-des-Palmistes, or Dimitile). Colored round nodes represent plant species (colors indicate the main plant taxonomic groups) and grey squared nodes correspond to fungal OTUs. Grey links represent plant-fungus interactions and their widths are proportional to interaction abundances. Fungal lineages (on rows) are ordered according to their network structures: networks that tend to be nested are on the top, whereas network that tend to be modular are on the bottom (based on Supplementary Tables 7-9). The sampled communities are on columns and we indicated the environmental gradients on the top. For each network, the total number of normalized reads (N) and the connectance (C) is indicated. Networks were visualized using the igraph R-package for the Swarm OTUs (but 97% OTUs gave very similar results). Details about the fungal taxonomy can be seen in Supplementary Figure 11.

When quantitatively investigating the network structures, we found that the endophytic networks presented an important range of connectance from 0.14 to 0.34, and tend to be less connected than the shuffle-sample null models (Figure III.7.4, Supplementary Table 6). This means that plant-fungus interactions are more alike between samples from the same plant species than between samples from different species. In terms of nestedness, compared to the quasiswap null models, and when considering weighted interactions, we found that large networks, like plant-Glomeromycotina and plant-Helotiales networks, tend to be significantly nested (Supplementary Table 7a). Conversely, smaller networks, plant-Sebacinales, plant-Cantharellales, and plant-Mucoromycotina networks were mostly non-significantly nested. When considering shuffle-sample null models, all networks were non-significantly nested, except for plant-Glomeromycotina incidence networks (Supplementary Table 7b). We also found similar trends when comparing the Cscore of the networks to null models (Supplementary Table 8): nested/anti-checkerboard structures (*i.e.* a strong asymmetrical specialization with an important overlap in shared partners) were only significant in large networks and when considering interactions strengths. Finally, we found contrasted evidence for modular structures in the endophytic networks: most of the networks were not significantly modular, and those that were significantly modular (in particular plant-Mucoromycotina, plant-Cantharellales, and plant-Helotiales networks) presented M values (the proportion of within-modules interactions) below 0.50 (Supplementary Table 9), suggesting that these inferred modules explained less than 50% of the endophytic interactions. However, many of these non-significances might arise from the fact that the network sizes are relatively small and might reduce the power of the comparisons to null models.

Specialization of endophytic interactions:

H2' values were lower in plant-Glomeromycotina networks than in the other endophytic networks (Figure III.7.5a; Supplementary Figure 13), suggesting that plant-Glomeromycotina interactions tend to be less specialized than other endophytic interac-

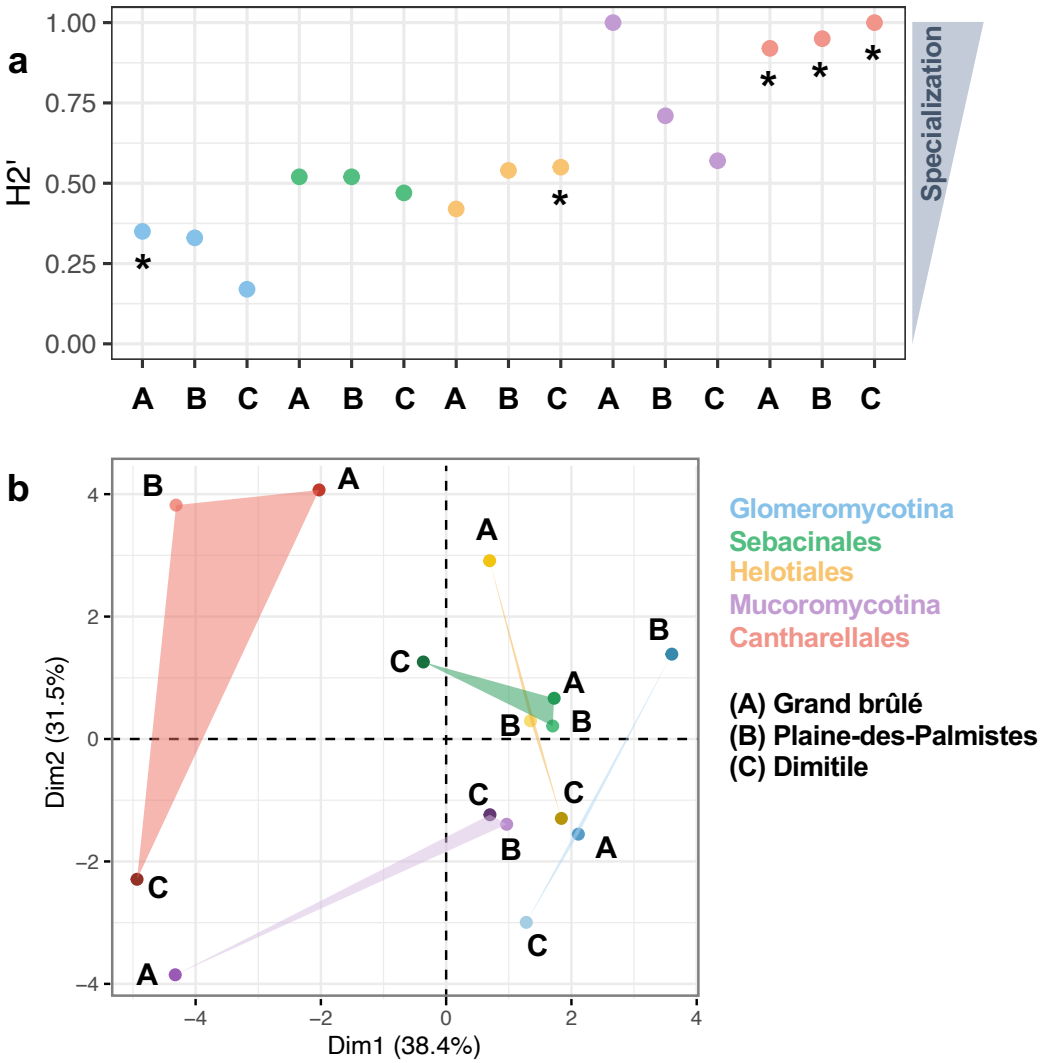


Figure III.7.5: Fungal sharing in the plant-fungus networks varies across the different fungal lineages. (a) Interaction specializations (H2′) are lower in plant-Glomeromycotina networks than in other plant-fungus networks For each endophytic network (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, or Cantharellales) in each sampled community (Grand brûlé (A), Plaine-des-Palmistes (B), or Dimitile (C)), a colored dot indicates the network-level interaction specialization (H2′) obtained with Swarm OTUs. The significance of the H2′ values was evaluated using null models maintaining marginal sums or shuffle-sample null models: all the H2′ values were significant for the marginal sums null models, and asterisks indicate when the H2′ values are significant based on the shuffle-sample null models (see Supplementary Figure 13 for details). (b) Motif frequencies significantly differ between the endophytic networks Principal coordinate analyses (PCoA) of the bipartite motif frequencies (from 2 to 5 species per motif) of each endophytic network (Glomeromycotina, Mucoromycotina, Sebacinales, Helotiales, or Cantharellales) in each sampled community (Grand brûlé (A), Plaine-des-Palmistes (B), or Dimitile (C)). The colored triangle areas represent the proximity within the sampled communities for the different groups of fungi. Motif analyses were performed on the unweighted networks of Swarm OTUs.

tions. While plant-Mucoromycotina and plant-Cantharellales interactions appeared to be highly specialized (high H2′ values), plant-Sebacinales and plant-Helotiales interactions presented an intermediate level of specialization (Figure III.7.5a). Looking at the spe-

cialization of the individual plant species, both normalized degree and d' indicated that most plant species were more specialized toward their endophytic fungi than expected if the interactions were randomly distributed based on species abundances (marginal null models; Supplementary Figure 14). However, these specialization values were generally no longer significant when compared with those of the shuffle-sample null models: many plant-fungus specializations therefore appeared to be driven by sample effects rather than species effects.

When comparing the motifs frequencies of the different endophytic networks using PCoA, we found that each type of endophytic network tends to cluster together (especially when using unweighted motifs; Figure III.7.5b; Supplementary Figure 15), suggesting that each of them had a particular motif signature. When looking at which motifs were differentially abundant (Supplementary Figure 16), we found that motifs where plants occupy a generalist position (*e.g.* motifs 8, 11, or 12) tended to be more abundant in plant-Glomeromycotina networks, but more networks would be needed to properly test this. To ensure that the low level of specialization observed in plant-Glomeromycotina interactions did not arise from the use of the 18S rRNA marker, we reproduced the analyses using the more variable ITS marker instead and still reported the same patterns of low specialization (Supplementary Figure 17). Similarly, replicating the plant-Sebacinales network analyses with 18S rRNA marker still showed that plant-Sebacinales interactions are on average more specialized than plant-Glomeromycotina ones (Supplementary Figure 18).

Finally, we investigated the motif frequencies of the Lycopodiaceae species and found that their motif frequencies were very variable according to the species and the sampled community (Supplementary Figure 19). Compared with the surrounding plant species, *Lycopodiella cernua* in Grand brûlé, *Lycopodiella caroliniana* in Plaine-des-Palmistes, or *Lycopodium clavatum* in Dimitile tend to be relatively more associated with Glomeromycotina partners that are well connected to other plant species in the networks (*e.g.* positions 19, 26, or 32). In contrast, for the other endophytic fungi, the Lycopodiaceae species tend to more frequently interact with specific fungi that are not connected with any other plant species in the network (*e.g.* positions 17, 20, 23, or 33).

Discussion:

In this study, we exhaustively characterized the plant-associated endophytic fungi within three contrasted local communities including distantly related plant species. We found that these plant communities across la Réunion island were mainly colonized by 5 main fungal lineages. Contrary to our expectations, we noticed a lot of fungal sharing between plant lineages that diverge >350 million years ago. When looking into detail at the different fungal lineages, we found striking differences in terms of specialization and

network structure (*i.e.* in the way the fungi connect the plant species), suggesting that they all establish singular interactions between the plant species. Interestingly, while the compositions of the endophytic microbiota vary according to the sampled communities, the plant-fungus network structures appeared to be resilient to the environmental variations.

Characterizing the composition of the root-associated fungal microbiomes:

First of all, we noticed that the ITS region (amplified with the ITS1-f and ITS4 primers) and the 18S rRNA gene (amplified with the AMAD-f and AMDG-r primers) primers) characterized different aspects of the root-associated fungal microbiota: While the 18S rDNA offers better visualization of all the fungal orders, the ITS enables better identification of the Dikarya (Figure III.7.2). These results therefore encourage systematically characterize the composition of plant-associated fungal microbiota by targeting both sets of regions. In particular, we noticed that Mucoromycotina (Endogonales) were almost always missed when using the common ITS1-f and ITS4 primer pairs, whereas they appear to be frequent endophytes and major mycorrhizal symbionts (Rimington *et al.*, 2015; Hoysted *et al.*, 2018). In addition, we did not detect any Tulasnellaceae in our study, while it is suspected to be a widespread saprotrophic group in the rhizosphere of plants including orchids, but is generally not successfully amplified when using regular ITS2 primers (Martos *et al.*, 2012; Vogt-Schilb *et al.*, 2020; Petrolli *et al.*, 2021). Thus, even with our dual marker strategy, we might still miss some root-associated fungal lineages.

We retrieved in our analyses the main types of mycorrhiza, which tend to be conserved across the main plant taxonomic groups. As expected, Glomeromycotina abundantly colonized most plant lineages, including bryophytes, lycopods, ferns, and many flowering plants (Brundrett & Tedersoo, 2018). Similarly, we observed that Mucoromycotina were frequently associated with a range of plants (Hoysted *et al.*, 2018), with the exceptions of ferns, as previously suggested by Rimington *et al.* (2015). We also confirmed that Sebacinales are major fungal endophytes (Weiß *et al.*, 2016), in particular frequently colonizing lycopods (Horn *et al.*, 2013). Similarly, Helotiales and Cantharellales were also retrieved as endophytes of many plant species. More surprisingly, we also detected abundant Mucoromycotina colonizations in orchids: Such associations had been previously detected in epiphytic orchids (Novotná *et al.*, 2018) and further works using the 18S rRNA marker should be pursued to investigate whether orchid-Mucoromycotina associations are indeed frequent.

Most plant individuals were generally colonized by several endophytic fungi susceptible to be mycorrhizal. For instance, we noticed that dual colonizations by Mucoromycotina and Glomeromycotina were particularly frequent, especially in lycopods. Experiments have demonstrated that such dual symbioses can both be functional and have complementary roles (Field *et al.*, 2016). Unfortunately, here, the molecular detection of a fungus in a plant root says nothing about the nature of the interaction (Toju *et al.*, 2016): it

can either correspond to a functioning mycorrhiza, to an opportunistic endophyte, or to a sporulating fungus that is transiently colonizing the root (Brundrett & Tedersoo, 2018). Testing which endophytic colonizations correspond to mycorrhizal ones, and whether colonizations by multiple fungal lineages are all functional and complementary would require further experimental works.

We also found a strong effect of the sampled communities on the endophytic microbiota compositions (Figure III.7.3). This suggests that environmental conditions, like altitude, disturbance, and humidity, influence the fungal distributions and the endophytic microbiota. Here, we cannot unravel the different effects of the environmental variables: Sampling more communities extensively covering the environment gradients would be necessary to do so. Among the strong community effects, we found that Mucoromycotina fungi were relatively abundant in the wet thickets (Plaine-des-Palmistes) and mostly absent in other sampled communities. Consequently, the endophytic compositions of *Lycopodiella cernua* present a clear shift according to its environment: while they mainly associate with Glomeromycotina and Sebacinales in Grand brûlé, Mucoromycotina represent up to 90% of the endophytic reads in Plaine-des-Palmistes. Further sampling should investigate whether Mucoromycotina only dominate the root endophytic microbiota in wet habitats.

When looking at the fungal composition of the root samples from the same species, we noticed an important heterogeneity. Although fungal microbiota from the same plant species were on average more alike than between two plant species, the endophytic compositions could vary drastically, as suggested by the rarefaction plots showing that even >10 samples were often not sufficient to get the whole diversity of the fungal associated with a plant species. As a consequence, we found that a large part of the plant-fungus specificity arose from single samples rather than properties of the plant species: more samples should be included to robustly infer which fungi are specific to each plant species. Besides the variability between samples, we also found that fungal microbiota indeed varied significantly across sampling replicates from the same community, suggesting that there are very local effects in the assembly of root-associated fungal microbiomes (Dumbrell *et al.*, 2010; Kokkoris *et al.*, 2020). Altogether, the effect of both the sampled community and the replicates on the endophytic microbiota compositions reflects the importance of ecological specificity in plant-fungus interactions (Molina *et al.*, 1992).

Fungal sharing *versus* specialization of the main fungal lineages:

In local communities, although we detected significant differences in the fungal microbiota compositions of the main plant taxonomic groups, we also observed a large amount of fungal sharing between co-occurring plant species, including between phylogenetically distant plants (Figure III.7.3). Thus, we found little evidence for the state-

ment "ancient plants with ancient fungi", resulting from the observation of frequent interactions between liverworts and early-diverging Glomeromycotina lineages (Rimington *et al.*, 2018). Fungal sharing was particularly important for Glomeromycotina that present very little specialization, as already often detected in local communities of flowering plants (van der Heijden *et al.*, 2015; Sepp *et al.*, 2019). Conversely, the other fungal lineages, especially the Mucoromycotina and Cantharellales, were more specialized and more sporadic in their interactions with plants. Sebacinales and Helotiales presented intermediate levels of specialization, confirming that they are widespread endophytes (Weiß *et al.*, 2016). Such differences in specialization might reflect the different evolutionary origins of these plant-associated fungi. Indeed, Glomeromycotina are thought to be an ancestral plant symbiont that obligately associate with them (Pirozynski & Malloch, 1975; Selosse & Le Tacon, 1998): although some plants have lost their dependence on Glomeromycotina through time (Werner *et al.*, 2018), they often retain the ability to occasionally host sparse Glomeromycotina fungi (Cosme *et al.*, 2018; Brundrett & Tedersoo, 2018), which could explain why Glomeromycotina tend to colonize many plant species with very low specificity. Conversely, Sebacinales, Helotiales, and Cantharellales have more recently acquired their ability to interact with plants and many of these lineages are still saprotrophs (Miyauchi *et al.*, 2020). Thus, plant colonization can be more facultative for them and often require a minimal plant-fungus specificity to be established (van der Heijden *et al.*, 2015), which could explain the higher specialization we observed for these lineages (Figure III.7.5). However, we also found that plant-Mucoromycotina interactions were quite specialized, facultative, and variable according to the environment, which seems contradictory with the fact that Mucoromycotina are being increasingly recognized as likely ancestral plant symbionts (Hoysted *et al.*, 2018; Feijen *et al.*, 2018). If Mucoromycotina were indeed a major ancestral plant symbiont, why they are nowadays facultative symbionts limited to particular environmental conditions remains unclear.

These different levels of specialization of the main fungal lineages resulted in different network structures (Figure III.7.4). In particular, plant-Glomeromycotina networks tend to exhibit significant nestedness, confirming a pattern frequently observed in local communities of flowering plants (Montesinos-Navarro *et al.*, 2012; Chagnon *et al.*, 2012; Sepp *et al.*, 2019). In addition, our null models demonstrated that the nestedness in plant-Glomeromycotina networks cannot be explained by abundance-driven interactions only (Chagnon, 2016). Conversely, other plant-fungus networks, in particular those composed of Mucoromycotina or Cantharellales, tend to present less connected, un-nested, and even modular structures, reflecting the higher specificity of these plant-fungus interactions. Our results thus support the tendency toward un-nested structures of non-Glomeromycotina networks, as previously observed in local communities of flowering plants (Bahram *et al.*, 2014; Põlme *et al.*, 2018) or in a liverwort-Mucoromycotina network at the global scale (Rimington *et al.*, 2019). Although we cannot exclude that we missed a part of their diversity (*e.g.* the Tulasnellaceae), the fact that the Cantharellales tend to form modular networks with plants might be due to the existence of a large diversity of

ecologies with plants in this group, from mutualism to antagonism (*e.g.* the pathogens Rhizoctonia), which might create modularity in the network (Fontaine *et al.*, 2011). By separately looking at the main endophytic fungal lineages, our approach thus better captured their singularities, which can be missed when merging and studying all fungal groups in the same framework (Toju *et al.*, 2016). Improvement of our characterization of the individual fungal ecological niches would allow us to investigate even more specifically the patterns of interactions at a finer taxonomic scale (*e.g.* at the genus level for lineages like the Cantharellales that comprises a large ranges of ecologies).

Interestingly, despite the contrasted environmental conditions of our sampled communities, we found overall consistent structural patterns for each plant-fungus network across habitats (Figure III.7.4): this suggests that even if the environmental conditions impact the relative abundances of the endophytes (Figure III.7.2), they do not strongly influence the network structures, which would instead result from intrinsic properties of the fungal lineages. This result contrasts with a recent metanalysis of plant-fungus interactions that found that the mean annual precipitation had more influence on the level of nestedness of plant-fungus interaction networks than the fungal lineages involved (Põlme *et al.*, 2018). Such results could be explained by the fact that Põlme *et al.*, (2018) considered a large heterogeneity of types of networks (individual-based networks or networks looking specifically at a certain plant clade only – *e.g.* only the orchid-fungus networks), whereas we only compared species-based networks at the level of a local community. In addition, we found that the exotic and invasive plant species were relatively well connected to the other plant species in the disturbed habitat (Grand brûlé), although they also tend to interact with many specific fungi, which is probably due to the fact that they belong to different plant taxonomic groups (*e.g.* orchids) (Bunn *et al.*, 2015). Sampling more communities along the environment gradients in La Réunion would be necessary to robustly evaluate whether environment conditions affect or not the structures of these plant-fungus interaction networks.

Contrary to what was suggested by a recent metanalysis of plant-Glomeromycotina interactions at a global scale (Article 6), we found that adult lycopod sporophytes were well connected to other plant species by fungal sharing. However, compared to other plant species, we also noticed the propensity of Lycopodiaceae species to interact with lycopod-specific fungal OTUs (especially from Mucoromycotina and Sebacinales; Figure III.7.4 and Supplementary Figure 19). Next works should particularly focus on the achlorophyllous gametophytes of lycopods to investigate what are the fungi providing them nutrients. In our study, we only found one gametophyte of *Lycopodiella cernua* close to Plaine-des-Palmistes, and this gametophyte was abundantly and specifically colonized by Mucoromycotina. If a more thorough sampling confirms that lycopod-specific fungi link both mycoheterotrophic gametophytes and green adult sporophytes, it would suggest that the organic matter could transit from the sporophytes to the gametophytes and that would strongly reinforce the hypothesis of parental nurture (Leake *et al.*, 2008),

even though adult sporophytes are also sharing fungi with other plant species.

Conclusion:

Therefore, by exhaustively characterizing plant-fungus interactions in local communities, our study demonstrated the distinctiveness in terms of specialization and network structure of the main endophytic fungal lineages, probably underpinned by the singular ecologies of these plant-fungus symbioses. In addition, it also showed that microbial sharing is widespread in local communities, even among distantly related plants. This calls for future works in order to characterize the functions of the endophytic microbiota in plant roots, several studies already suggesting that they can be involved in a plethora of functions, from plant nutrition to protection (Newsham, 2011; Almario *et al.*, 2017). It therefore highlights the importance of systematically considering networks of interactions (the "network-biont") rather than isolated macroorganisms and their associated microbes (the "holobiont").

References:

Almario J, Jeena G, Wunder J, Langen G, Zuccaro A, Coupland G, Bucher M. 2017. Root-associated fungal microbiota of nonmycorrhizal *Arabis alpina* and its contribution to plant phosphorus nutrition. Proceedings of the National Academy of Sciences of the United States of America 114: E9403–E9412.

Babikova Z, Gilbert L, Bruce TJA, Birkett M, Caulfield JC, Woodcock C, Pickett JA, Johnson D. 2013. Underground signals carried through common mycelial networks warn neighbouring plants of aphid attack (N van Dam, Ed.). Ecology Letters 16: 835–843.

Bahram M, Harend H, Tedersoo L. 2014. Network perspectives of ectomycorrhizal associations. Fungal Ecology 7: 70–77.

Bascompte J, Jordano P, Melián CJ, Olesen JM. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences of the United States of America 100: 9383–9387.

Berendsen RL, Pieterse CMJ, Bakker PAHM. 2012. The rhizosphere microbiome and plant health. Trends in Plant Science 17: 478–486.

Berruti A, Desirò A, Visentin S, Zecca O, Bonfante P. 2017. ITS fungal barcoding primers *versus* 18S AMF-specific primers reveal similar AMF-based diversity patterns in roots and soils of three mountain vineyards. Environmental Microbiology Reports 9: 658–667.

Blüthgen NN, Menzel F, Blüthgen NN. 2006. Measuring specialization in species interaction networks. BMC Ecology 6.

Boullard B. 1979. Considérations sur la symbiose fongique chez les Ptéridophytes. Syllogeus 19 : 1-58.

Brundrett MC, Tedersoo L. 2018. Evolutionary history of mycorrhizal symbioses and global host plant diversity. New Phytologist 220: 1108–1115.

Bunn RA, Ramsey PW, Lekberg Y. 2015. Do native and invasive plants differ in their interactions with arbuscular mycorrhizal fungi? A meta-analysis (M van der Heijden, Ed.). Journal of Ecology 103: 1547–1556.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Chagnon PL. 2016. Seeing networks for what they are in mycorrhizal ecology. Fungal Ecology 24: 148–154.

Chagnon PL, Bradley RL, Klironomos JN. 2012. Using ecological network theory to evaluate the causes and consequences of arbuscular mycorrhizal community structure. New Phytologist 194: 307–312.

Chagnon PL, Bradley RL, Klironomos JN. 2015. Trait-based partner selection drives mycorrhizal network assembly. Oikos 124: 1609–1616.

Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28: 2106–2113.

Cosme M, Fernández I, van der Heijden MGA, Pieterse CMJ. 2018. Non-mycorrhizal plants: The exceptions that prove the rule. Trends in Plant Science 23: 577–587.

Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome 6: 226.

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, *et al.* 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349: 970–973.

Dormann CF. 2011. How to be a specialist? Quantifying specialisation in pollination networks. Network Biology 1: 1–20.

Dormann CF, Gruber B, Fründ J. 2008. Introducing the bipartite package: analysing ecological networks. R News 8: 8–11.

Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. 2010. Relative roles of niche and neutral processes in structuring a soil microbial community. ISME Journal 4: 337–345.

Feijen FA, Vos RA, Nuytinck J, Merckx VSFT. 2018. Evolutionary dynamics of mycorrhizal symbiosis in land plant diversification. Scientific Reports 8: 10698.

Field KJ, Pressel S, Duckett JG, Rimington WR, Bidartondo MI. 2015. Symbiotic options for the conquest of land. Trends in Ecology & Evolution 30: 477–486.

Field KJ, Rimington WR, Bidartondo MI, Allinson KE, Beerling DJ, Cameron DD, Duckett JG, Leake JR, Pressel S. 2016. Functional analysis of liverworts in dual symbiosis with Glomeromycota and Mucoromycotina fungi under a simulated Palaeozoic CO_2 decline. ISME Journal 10: 1514–1526.

Fontaine C, Guimarães PR, Kéfi S, Loeuille N, Memmott J, van der Putten WH, van Veen FJF, Thébault E. 2011. The ecological and evolutionary implications of merging different types of networks. Ecology Letters 14: 1170–1181.

Frank B. 1885. Ueber die auf Wurzelsymbiose beruhende Ernährung gewisser Bäume durch unterirdische Pilze. Berichte der Deutschen Botanischen Gesellschaft 3: 128–145.

Gotelli NJ. 2000. Null model analysis of species co-occurrence patterns. Ecology 81: 2606–2621.

van der Heijden MGA, Martin FM, Selosse MA, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytologist 205: 1406–1423.

Horn K, Franke T, Unterseher M, Schnittler M, Beenken L. 2013. Morphological and molecular analyses of fungal endophytes of achlorophyllous gametophytes of *Diphasiastrum alpinum* (Lycopodiaceae). American Journal of Botany 100: 2158–2174.

Hoysted GA, Bidartondo MI, Duckett JG, Pressel S, Field KJ. 2020. Phenology and function in lycopod–Mucoromycotina symbiosis. New Phytologist: 0–2.

Hoysted GA, Kowal J, Jacob A, Rimington WR, Duckett JG, Pressel S, Orchard S, Ryan MH, Field KJ, Bidartondo MI. 2018. A mycorrhizal revolution. Current Opinion in Plant Biology 44: 1–6.

Jacquemyn H, Brys R, Waud M, Busschaert P, Lievens B. 2015. Mycorrhizal networks and coexistence in species-rich orchid communities. New Phytologist 206: 1127–1134.

Jacquemyn H, Merckx VSFT, Brys R, Tyteca D, Cammue BPAA, Honnay O, Lievens B. 2011. Analysis of network architecture reveals phylogenetic constraints on mycorrhizal specificity in

the genus *Orchis* (Orchidaceae). New Phytologist 192: 518–528.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 14: 587–589.

Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Kiers ET, Duhamel M, Beesetty Y, Mensah JA, Franken O, Verbruggen E, Fellbaum CR, Kowalchuk GA, Hart MM, Bago A, *et al.* 2011. Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. Science 333: 880–882.

Kokkoris V, Lekberg Y, Antunes PM, Fahey C, Fordyce JA, Kivlin SN, Hart MM. 2020. Codependency between plant and arbuscular mycorrhizal fungal communities: what is the evidence? New Phytologist 228: 828–838.

Leake JR, Cameron DD, Beerling DJ. 2008. Fungal fidelity in the myco-heterotroph-to-autotroph life cycle of Lycopodiaceae: A case of parental nurture? New Phytologist 177: 572–576.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2015. Swarmv2: Highly-scalable and high-resolution amplicon clustering. PeerJ 2015: 1–12.

Manly BFJ. 2018. Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall/CRC.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17: 10.

Martos F, Munoz FF, Pailler T, Kottke I, Gonneau C, Selosse MA. 2012. The role of epiphytism in architecture and evolutionary constraint within mycorrhizal networks of tropical orchids. Molecular Ecology 21: 5098–5109.

McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. 2019. Methods for normalizing microbiome data: An ecological perspective (S Jarman, Ed.). Methods in Ecology and Evolution 10: 389–400.

McMurdie PJ, Holmes S. 2014. Waste not, want not: Why rarefying microbiome data is inadmissible (AC McHardy, Ed.). PLoS Computational Biology 10: e1003531.

Merckx VSFT. 2013. Mycoheterotrophy: An Introduction. In: Merckx VSFT, ed. Mycoheterotrophy. New York, NY: Springer New York, 1–17.

Miyauchi S, Kiss E, Kuo A, Drula E, Kohler A, Sánchez-García M, Morin E, Andreopoulos B, Barry KW, Bonito G, *et al.* 2020. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. Nature Communications 11: 1–17.

Molina R, Massicotte H, Trappe JM. 1992. Specificity phenomena in mycorrhizal symbioses: community-ecological consequences and practical implications. In: Allen, Routledge, eds. Mycorrhizal functioning, an integrative plant-fungal process. New York: Chapman and Hall, 357-423.

Montesinos-Navarro A, Segarra-Moragues JG, Valiente-Banuet A, Verdú M. 2012. The network structure of plant-arbuscular mycorrhizal fungi. New Phytologist 194: 536–547.

Montesinos-Navarro A, Segarra-Moragues JG, Valiente-Banuet A, Verdú M. 2015. Evidence for phylogenetic correlation of plant-AMF assemblages? Annals of Botany 115: 171–177.

Newsham KK. 2011. A meta-analysis of plant responses to dark septate root endophytes. New Phytologist 190: 783–793.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and Evolution 32: 268–274.

Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, Schilling JS, Kennedy PG. 2016. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecology 20: 241–248.

Nilsson RH, Larsson KH, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, *et al.* 2019. The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. Nucleic Acids Research 47: D259–D264.

Novotná A, Benítez Á, Herrera P, Cruz D, Filipczyková E, Suárez JP. 2018. High diversity of root-associated fungi isolated from three epiphytic orchids in southern Ecuador. Mycoscience 59: 24–32.

Oksanen J, Kindt R, Pierre L, O'Hara B, Simpson GL, Solymos P, Stevens MH. HH, Wagner H, Blanchet FG, Kindt R, *et al.* 2016. vegan: Community Ecology Package, R-package version 2.4-0. R-package version 2.2-1.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist 188: 223–241.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Patefield WM. 1981. An efficient method of generating random R × C tables with given row and column totals. Applied Statistics 30: 91.

Perez-Lamarque B, Selosse MA, Öpik M, Morlon H, Martos F. 2020. Cheating in arbuscular mycorrhizal mutualism: a network and phylogenetic analysis of mycoheterotrophy. New Phytologist 226: 1822–1835.

Petrolli R, Vieira CA, Jakalski M, Bocayuva MF, Valle C, Cruz ED V, Selosse MA, Martos F, Kasuya MCM. 2021. A fine-scale spatial analysis of fungal communities on tropical tree bark shows the epiphytic rhizosphere in orchids. New Phytologist (under review).

Philippot L, Raaijmakers JM, Lemanceau P, Van Der Putten WH. 2013. Going back to the roots: The microbial ecology of the rhizosphere. Nature Reviews Microbiology 11: 789–799.

Pirozynski KA, Malloch DW. 1975. The origin of land plants: A matter of mycotrophism. BioSystems 6: 153–164.

Põlme S, Bahram M, Jacquemyn H, Kennedy P, Kohout P, Moora M, Oja J, Öpik M, Pecoraro L, Tedersoo L. 2018. Host preference and network properties in biotrophic plant–fungal associations. New Phytologist 217: 1230–1239.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41: D590–D596.

R Core Team. 2020. R: A language and environment for statistical computing.

Rimington WR, Pressel S, Duckett JG, Bidartondo MI. 2015. Fungal associations of basal vascular plants: reopening a closed book? New Phytologist 205: 1394–1398.

Rimington WR, Pressel S, Duckett JG, Field KJ, Bidartondo MI. 2019. Evolution and networks in ancient and widespread symbioses between Mucoromycotina and liverworts. Mycorrhiza 29: 551–565.

Rimington WR, Pressel S, Duckett JG, Field KJ, Read DJ, Bidartondo MI. 2018. Ancient plants with ancient fungi: liverworts associate with early-diverging arbuscular mycorrhizal fungi. Proceedings of the Royal Society B: Biological Sciences 285: 20181600.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: A versatile open source tool for metagenomics. PeerJ 2016: e2584.

Schmid E, Oberwinkler F. 1993. Mycorrhiza-like interaction between the achlorophyllous gametophyte of *Lycopodium clavatum L.* and its fungal endophyte studied by light and electron microscopy. New Phytologist 124: 69–81.

Schneider-Maunoury L, Deveau A, Moreno M, Todesco F, Belmondo S, Murat C, Courty PE, Jąkalski M, Selosse MA. 2020. Two ectomycorrhizal truffles, *Tuber melanosporum* and *T. aestivum*, endophytically colonise roots of non-ectomycorrhizal plants in natural environments. New Phytologist 225: 2542–2556.

Selosse MA, Baudoin E, Vandenkoornhuyse P. 2004. Symbiotic microorganisms, a key for ecological success and protection of plants. Comptes Rendus - Biologies 327: 639–648.

Selosse MA, Dubois MP, Alvarez N. 2009. Do Sebacinales commonly associate with plant roots as endophytes? Mycological Research 113: 1062–1069.

Selosse MA, Schneider-Maunoury L, Martos F. 2018. Time to re-think fungal ecology? Fungal ecological niches are often prejudged. New Phytologist 217: 968–972.

Selosse MA, Le Tacon F. 1998. The land flora: a phototroph-fungus partnership? Trends in Ecology & Evolution 13: 15–20.

Sepp SK, Davison J, Jairus T, Vasar M, Moora M, Zobel M, Öpik M. 2019. Non-random association patterns in a plant–mycorrhizal fungal network reveal host–symbiont specificity. Molecular Ecology 28: 365–378.

Simard SW, Perry DA, Jones MD, Myrold DD, Durall DM, Molina R. 1997. Net transfer of carbon between ectomycorrhizal tree species in the field. Nature 388: 579–582.

Simmons BI, Sweering MJM, Schillinger M, Dicks L V., Sutherland WJ, Di Clemente R. 2019. bmotif: A package for motif analyses of bipartite networks. Methods in Ecology and Evolution 10: 695–701.

Smith SE, Read DJ. 2008. Mycorrhizal Symbiosis. Elsevier.

Strasberg D, Rouget M, Richardson DM, Baret S, Dupont J, Cowling RM. 2005. An assessment of habitat diversity and transformation on La Réunion Island (Mascarene Islands, Indian Ocean) as a basis for identifying broad-scale conservation priorities. Biodiversity and Conservation 14: 3015–3032.

Strullu-Derrien C, Selosse MA, Kenrick P, Martin FM. 2018. The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. New Phytologist 220: 1012–1030.

Taberlet P, Bonin A, Zinger L, Coissac E. 2018. DNA amplification and multiplexing. In: Environmental DNA. 41–57.

Tedersoo L, Mett M, Ishida TA, Bahram M. 2013. Phylogenetic relationships among host plants explain differences in fungal species richness and community composition in ectomycorrhizal symbiosis. New Phytologist 199: 822–831.

Thébault E, Fontaine C. 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329: 853–856.

Toju H, Guimarães PR, Olesen JM, Thompson JN. 2014. Assembly of complex plant-fungus networks. Nature Communications 5: 1–7.

Toju H, Guimarães PR, Olesen JM, Thompson JN. 2015. Below-ground plant–fungus network topology is not congruent with above-ground plant–animal network topology. Science Advances 1: e1500291.

Toju H, Tanabe AS, Ishii HS. 2016. Ericaceous plant-fungus network in a harsh alpine-subalpine environment. Molecular Ecology 25: 3242–3257.

Turenne CY, Sanche SE, Hoban DJ, Karlowsky JA, Kabani AM. 1999. Rapid identification of fungi by using the ITS2 genetic region and an automated fluorescent capillary electrophoresis system. Journal of Clinical Microbiology 37: 1846–1851.

Vandenkoornhuyse P, Ridgway KP, Watson IJ, Fitter AH, Young JPW. 2003. Co-existing grass species have distinctive arbuscular mycorrhizal communities. Molecular Ecology 12: 3085–3095.

Verbruggen E, van der Heijden MGA, Weedon JT, Kowalchuk GA, Rö-Ling WFM. 2012. Community assembly, species richness and nestedness of arbuscular mycorrhizal fungi in agricultural soils. Molecular Ecology 21: 2341–2353.

Vogt-Schilb H, Těšitelová T, Kotilínek M, Sucháček P, Kohout P, Jersáková J. 2020. Altered rhizoctonia assemblages in grasslands on ex-arable land support germination of mycorrhizal generalist, not specialist orchids. New Phytologist 227: 1200–1212.

Weiß M, Waller F, Zuccaro A, Selosse MA. 2016. Sebacinales - one thousand and one interactions with land plants. New Phytologist 211: 20–40.

Werner GDA, Cornelissen JHC, Cornwell WK, Soudzilovskaia NA, Kattge J, West SA, Kiers ET. 2018. Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown. Proceedings of the National Academy of Sciences 115: 5229–5234.

Werner GDA, Kiers ET. 2015. Partner selection in the mycorrhizal mutualism. New Phytologist 205: 1437–1442.

White T, Bruns T, Lee S, Taylor J. 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: PCR Protocols. 315–322.

Winther JL, Friedman WE. 2008. Arbuscular mycorrhizal associations in Lycopodiaceae. New Phytologist 177: 790–801.

Zanne AE, Tank DC, Cornwell WK, Eastman JM, Smith SA, FitzJohn RG, McGlinn DJ, O'Meara BC, Moles AT, Reich PB, *et al.* 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506: 89–92.

Supplementary data:

The supplementary data of this article are available through the link `https://bit.ly/3sm9qHa` or by scanning:

Chapter IV.

Discussion:

During my PhD, I have used a range of approaches to better understand the evolution of host-microbiota interactions. My works have included the development of a new model (Article 1) and the comparisons of the statistical performances of different approaches to study microbial inheritances (Article 3) or phylogenetic signals in species interactions (Article 4). I have applied these tools to various empirical datasets, ranging from the gut microbiota of primates (Articles 1 and 3) or spiders (Article 2), to the root-associated mycorrhizal microbiota (Articles 4, 5, 6, and 7). I have also worked on the application of models of species diversification to microbial groups (Article 5). All my works included phylogenetic-based approaches and most of them used the framework of bipartite networks (Articles 4, 5, 6, and 7). The microbiota datasets we used were mostly coming from metabarcoding experiments amplifying the SSU rRNA genes (for prokaryotes in Articles 1, 2, and 3, or fungi in Articles 5, 6, and 7) or the fungal ITS region (Articles 4 and 7). Finally, I have also performed fieldwork and molecular work to characterize mycorrhizal networks (Article 7). All these projects have been performed under the supervision of my PhD advisors and in collaboration with researchers at IBENS and ISYEB, as well as external collaborators who in particular gave us access to empirical datasets, including Maarja Öpik (Tartu University, Estonia), Henrik Krehenwinkel (Trier University, Germany), and Rosemary Gillespie (UC Berkeley, USA).

In this Discussion, we will first present a synthesis of the main findings of my PhD and the future works that could be carried related to them (section 1). Next, in a perspective part, we will more generally discuss the current limits and the future improvements of the metabarcoding datasets (section 2.1) and the quantitative approaches (section 2.2) for investigating the evolution of host-microbiota interactions. Then, we will synthesize different theories of host-microbiota evolutions and their support, complemented by some of our findings (section 2.3). Finally, we will conclude by discussing how our recent changes of lifestyle have impacted worldwide host-associated microbiota (section 2.4).

Contents of Chapter IV

1. Synthesis

In Chapter I, we have developed a model, HOME, for detecting transmitted microbial symbionts during the host clade diversification using DNA metabarcoding datasets. Using simulations, we found that our approach has an intermediate statistical power and very low type-I error. In other words, our approach can miss some of the transmitted microbes, but it is unlikely to infer many transmitted microbes that are actually acquired from the environment. Thus, HOME has the advantages (i) of not making any *a priori* on the microbial groups that could be transmitted (unlike Moeller *et al.*, 2016) and (ii) of working even when the divergence of the host clade is very recent (unlike Groussin *et al.*, 2017) and when using short DNA sequences that are slowly evolving. Indeed, compared with global-fit approaches like ParaFit (Legendre *et al.*, 2002) or PACo (Balbuena *et al.*, 2013), we found that HOME has fewer false-positives but lower power, especially when the number of segregating sites is low (*i.e.* when there is a low amount of information in the metabarcoding sequences). In addition, HOME is faster than the event-based approach ALE (Szöllősi *et al.*, 2013) in these conditions. Therefore, we recommend jointly use HOME with global-fit approaches like ParaFit: HOME will give a list of microbes that are very likely to be transmitted and ParaFit will add to this list some microbes that could be transmitted (but also many false positives). When applied to empirical datasets, we found contrasting results across the different animal groups we considered: while several bacterial lineages have been transmitted in the gut microbiota of primates (representing up to 10% of the total gut bacteria), this is likely not the case in the microbiota of spiders that showed very little evidence of transmission. Our results confirm that microbial transmission is frequent in mammal gut microbiota, as highlighted by previous experimental works (Moeller *et al.*, 2018) or co-phylogenetic analyses (Sanders *et al.*, 2014; Moeller *et al.*, 2016; Groussin *et al.*, 2017). Conversely, in arthropods, we know that microbial transmissions are very heterogeneous, with some host clades having mechanisms ensuring faithful transmissions (Engel & Moran, 2013), whereas others have microbes largely acquired from their environment (Kennedy *et al.*, 2020); the Hawaiian *Ariamnes* spiders seem to belong to this latter category. Interestingly, it means that the patterns of phylosymbiosis observed in both primates and spiders datasets are at least in part generated by different processes: vertical transmissions likely participate to the phylosymbiosis in primates, whereas in spiders, the heterogeneous pools of microbes (in particular the endosymbionts) likely mainly explain the phylosymbiosis, as closely related hosts are in contact with similar environmental pools of microbes. Therefore, our results illustrate the heterogeneity of microbiota evolution across animals and similar analyses should be pursued in other host clades to identify the "hotspots" of microbial transmissions among animals and plants.

In Chapter II, we have investigated the interplay between the evolutionary histories of hosts and their associated microbes. In Article 4, we have compared different methods for measuring phylogenetic signals in species interactions (*i.e.* whether closely re-

lated species tend to interact with similar partners) and have found that two widely used approaches, the phylogenetic bipartite linear model (PBLM; Ives & Godfray, 2006) and partial Mantel tests (*e.g.* Rezende *et al.*, 2007) are generally inaccurate for measuring phylogenetic signals. Conversely, despite intermediate statistical power, simple Mantel tests are likely the most accurate and fastest method available for measuring phylogenetic signals in species interactions. In addition, we have considered the adjustments of the Mantel tests for investigating phylogenetic signals in host-microbiota networks, by proposing a robust way to test for clade-specific phylogenetic signals and by testing that phylogenetic uncertainty is unlikely to strongly bias the results. We provided clear guidelines for future empirical applications and illustrated them on an orchid mycorrhizal network from La Réunion: we confirmed with our approaches that there is a significant phylogenetic signal in only a clade of epiphytic orchids, the Angraecinae, suggesting that host lineages colonizing a new habitat (here, epiphytism) might be constrained to specialize toward specific partners. Second, in Article 5, we investigated how land plants might have affected the diversification of the obligate mycorrhizal symbionts (Glomeromycotina). We found that this clade of fungi has relatively low diversification rates compared with other groups and that after a peak of diversification approximately 150 million years (Myr) ago, they experienced a strong decline in the rates of diversification. We suggested that this diversification slowdown might be related to the breakdown by many plant lineages of their symbiosis with Glomeromycotina. Indeed, plants have recently evolved alternative symbiotic strategies or have stopped relying on symbionts for their nutrition (Werner *et al.*, 2018). In the past 100 Myr, the proportion of plant clades relying on AMF decreased by $\sim$40% (Feijen *et al.*, 2018). In combination with abiotic events (like the Cenozoic climatic cooldown), it likely explains why Glomeromycotina diversification has slowed down. Because such a pattern can also result from various methodological artifacts (Moen & Morlon, 2014) linked to the use of SSU rRNA metabarcoding datasets to investigate the diversification of a microbial clade, we have performed a large range of sensitivity analyses to ensure the robustness of our results, including testing several species delineations, sampling fractions, and phylogenetic reconstructions, as well as simulations. We thus have provided an example of a framework to study the diversification of a microbial group while robustly controlling for potential biases. Overall, our results highlight the roles of the evolutionary histories of both hosts and microbes in the current patterns of diversity and species interactions.

Lastly, in Chapter III, we have focused on the evolution of cheating (mycoheterotrophy) in the arbuscular mycorrhizal symbiosis. We first performed a global analysis on plant-Glomeromycotina interactions and found that mycoheterotrophic plants tend to be specialized toward specialist and closely-related mycorrhizal fungi. This pattern of reciprocal specialization contrasts with the patterns of asymmetrical specializations, widespread in mutualistic mycorrhizal interactions. Our analyses therefore suggest that cheaters and their partners are isolated in the global network, probably because of constraints limiting their emergence, either physiological constraints (*e.g.* partner choice or

sanctions) or habitat constraints. Importantly, our results say something about whether or not mycoheterotrophic plants are also isolated in local communities where they live. Mycoheterotrophic plants are mostly recovered in understory habitats where carbon is not the limiting resource (Merckx, 2013; Gomes *et al.*, 2019a) and a recent study did not report any pattern of reciprocal specialization in these communities (Gomes *et al.*, 2019b). We can speculate that mycoheterotrophs could have emerged in such habitats only because they represent a negligible cost in terms of carbon for their fungal partners and the surrounding mycorrhizal plants (Kiers & van der Heijden, 2006). In other words, we suggest that the strong constraints limiting the emergence of mycoheterotrophy in the mycorrhizal mutualism could be relaxed when carbon is not the limiting resource. In addition, our analyses suggest that mycoheterotrophy can evolve through a progressive loss of the mycorrhizal partners associated with the autotrophic ancestors, which has been recently confirmed in the *Burmannia* clade (Zhao *et al.*, 2021). Finally, we found that a clade of initially mycoheterotrophic plants, the lycopods, were specifically associating with a clade of *Glomus*, which tends to support the hypothesis of parental nurture through the sharing of specific fungi between green sporophytes and achlorophyllous gametophytes. In other words, lycopods would not be true cheaters *per se*, as the mycoheterotrophic gametophytes would be nurtured by the sporophytes thanks to shared fungi (Field *et al.*, 2015). In Article 7, we investigated whether such patterns of specificity in lycopods were recovered in local communities that we sampled in La Réunion island. We found that, although lycopods-specific fungi exist, a lot of fungi are shared between lycopods and other plants. A specific and in-depth sampling of the sporophytes and gametophytes of lycopods in some local communities would help to better characterize whether private networks supporting parental nurture actually exist between sporophytes and gametophytes, or alternatively, if the patterns of strong specificity we observed in Article 6 resulted from an under-sampling bias. Altogether, these works support the idea that cheating is importantly constrained in the host-microbiota mutualisms; further works targeting specifically the cheaters and their partners in local communities will likely provide more information on the processes that allow cheaters to emerge in some communities.

2. Perspectives

2.1. The limits of metabarcoding to characterize microbiota evolution

In this section, we discuss the limits of DNA metabarcoding, the imminent improvement that meta-omics approaches will bring, and how complexifying our vision of host-associated microbiota would be needed to better understand their evolution.

Characterizing microbial communities using DNA metabarcoding has, by itself, plenty of limits, that can impact our ability to infer the evolution of host-microbiota interactions. First, when doing metabarcoding, one has to choose a barcoding region to amplify and a corresponding set of primers. This step already has important consequences on the characterization of microbiota composition (Bukin *et al.*, 2019): for instance, in Article 7, we found that the 18S rRNA gene or the ITS region often reported drastically different root-associated fungal communities, and it seems that even using a combination of two primer pairs is not enough to recover all the fungal symbionts (*e.g.* the Tulasnellaceae). Some microbial clades are thus likely excluded in all the following analyses; the same likely applies for archaea in primate guts that are almost absent in the datasets we used (Articles 1 and 3). Second, performing DNA metabarcoding of the ribosomal RNA operon, which is present in several copies in microbial genomes, can generate within microbial species variations of the barcoding genes. Consequently, when looking at nucleotidic substitutions within the barcoding sequences to reconstruct their evolution, one can track within microbial genomes divergences rather than divergences between microbial lineages (Pérez-Cobas *et al.*, 2020). We tackled this issue by selecting only the most abundant sequence per species/individuals (Articles 1-3) or by merging similar reads into OTUs (Articles 4-7), but more sophisticated approaches could be considered in the future (Pérez-Cobas *et al.*, 2020). Third, metabarcoding markers, like the SSU rRNA genes, are slowly evolving genes: they accumulated only few substitutions in the recent past (1% on average per 50 million years for the 16S rRNA gene of bacteria), such that they generally cannot be used to robustly reconstruct recent microbial evolutions. If the metabarcoding genes have accumulated no substitution since the host diverged, that will totally prevent us from reconstructing their evolutionary history (Article 2). In addition, if the substitution rate of the metabarcoding gene is lower than the speciation rate, species would accumulate at a higher speed than mutations, such that the gene will be improper for (i) delineating species and (ii) reconstructing robust phylogenetic trees (Article 5). These are serious issues that can strongly impact our abilities to study microbial evolution and bias our conclusions. Here, we have dealt with them by specifically developing a model to infer transmitted bacteria when the number of segregating sites is very low (Article 1) and by performing a range of sensitivity analyses when studying the diversification of a microbial group (Article 5). In the future, to avoid such issues, one could instead target metabarcoding genes that are more rapidly evolving or use sequencing technologies that enable to amplify longer reads, like Nanopore or PacBio sequencing, that can sequence the whole rRNA operon (Kolaříková *et al.*, 2021).

Metabarcoding datasets enable us to have a list of some of the microbial taxa present in a sample, but it can be challenging to get insights into the metabolic abilities and the functions of the microbial communities. One can extrapolate the microbial functions and niches based on their taxonomy using available databases; for instance, FUNGuild (Nguyen *et al.*, 2016) has allowed us to identify the fungal OTUs likely to be mycorrhizal (Article 7). However, such databases are particularly limited, especially when working with microbes of non-model host organisms. A more reliable approach is to use metagenomic or metatranscriptomic technics that enable a better characterization of the genes present or expressed respectively. First, such technics allow a better understanding of the microbiota functioning as a whole (Muegge *et al.*, 2011; Simon *et al.*, 2019) and could be used to improve our understanding of their evolution. Second, if metagenomic datasets can be assembled into metagenome-assembled genomes (MAGs), that could (i) provide more information for reconstructing their evolutionary history and (ii) give more insights into the functional niches of the different microbes present (Brochet *et al.*, 2021). Indeed, MAGs would enable more precise species delineations, not (arbitrarily) relying on only genetic divergences (Sukumaran & Knowles, 2017), and using multiple genes would greatly improve phylogenetic reconstructions. In addition, having MAGs would also provide a better understanding of what each microbial symbiont is doing: for instance, we suspect that arbuscular mycorrhizal fungi have different benefits for plants; some are efficiently providing nutrients to the plants, whereas others are better at modulating biotic or abiotic stresses (Chagnon *et al.*, 2013). Having such information in our hands would enable us to test how variations in traits (*e.g.* linked to host restrictiveness or host functioning) can affect the diversification of these microbial symbionts (Article 5). Similarly, having MAGs would allow us to investigate if certain gene acquisitions in microbial genomes (*e.g.* lateral gene transfer) have impacted the evolution of the corresponding symbiotic microbes (Hehemann *et al.*, 2010). Finally, MAGs would provide better insight into the genes that are involved in host-microbes interactions and communications and might be essential for the emergence and the functioning of symbiotic interactions (Delaux *et al.*, 2013; van der Heijden *et al.*, 2015). While the current sequencing coverage and assembly programs enable to characterize only the most abundant microbes, imminent breakthroughs in the field may rapidly generalize the use of MAGs for answering questions of host-microbiota evolutions.

Finally, in my PhD, we considered a rather static view of host-associated microbiota, as a list of microbes present in a plant or an animal. A better way to describe host-associated microbiota to study their evolution would be to consider (i) microbiota as dynamic ecosystems and (ii) the heterogeneity of interactors it contains. First, performing a temporal and spatial sampling of the host-associated microbiota would enable us to have a better understanding of the complex dynamics and variability of the host-microbiota interactions and the factors that can influence them (Koskella *et al.*, 2017). Second, we mostly neglected all the microbe-microbe relationships (*e.g.* mutualism, fa-

cilitation, competition, predation…) occurring within a microbiota. Unfortunately, such interactions remain poorly characterized, while they are likely a fundamental part of microbiota functioning and influence the patterns of evolution (Foster *et al.*, 2017). In some extreme cases, there are nested interactions within microbiota: for instance, Glomeromycotina and Mucoromycotina fungi are often associated with endosymbiotic bacteria (called mycoplasma-related endobacteria) that seem to play important roles in their functioning (Bonfante & Anca, 2009), like *Burkholderia* in *Gigaspora margarita* that supply nitrogen. Similarly, the phages associated with gut bacteria play major roles in the equilibrium of the bacterial communities and have likely co-evolved with them over long timescales (Gogarten *et al.*, 2021). Using co-occurrence networks could be a way, from metabarcoding datasets, to infer positive or negative microbe-microbe interactions (Faust & Raes, 2012). Altogether a better understanding of the host-associated microbiota would allow us to formulate clearer testable hypotheses on the drivers of host-microbiota evolutions.

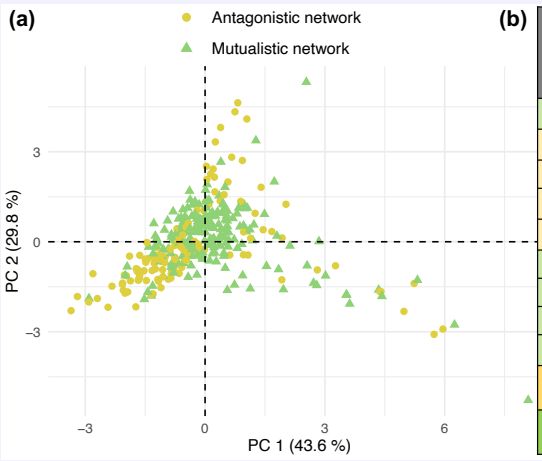
2.2. Toward a better modeling of host-microbiota evolution

Empirical applications of quantitative approaches of host-microbiota evolutions (*e.g.* models of microbial inheritances, methods for measuring phylogenetic signals in species interactions, models of microbial diversification) are all directly impacted by the limits inherent to the microbial metabarcoding datasets (see the previous section). The performance of these approaches (in terms of both statistical power and type-I error rate) would be greatly improved by increasing the lengths of the DNA barcodes and/or targeting other genes. However, these quantitative approaches also suffer from intrinsic limits that will require further attention and development.

Measuring phylogenetic signals in host-microbiota evolutions, with Mantel tests for instance, can be fast and useful as a first way of determining whether closely related species tend to interact with similar partners (phylosymbiosis; Article 4). However, it remains an overall measure of a general tendency and says nothing about the processes of microbiota evolution. Similarly, the structure of bipartite interaction networks is often studied using nestedness or modularity, but these patterns can be generated by various ecological or evolutionary processes (Fontaine *et al.*, 2011) and whether they have predictive power is debated (Box 1). Future efforts should be done to better apprehend the evolution of the interactions with process-driven models, that either consider microbial units separately (*e.g.* models of transmission) or all together (*e.g.* models of network evolutions).

Box 1: Do mutualistic *versus* antagonistic networks have different structures?

In Articles 6 and 7, we assumed that mutualistic networks were rather nested, whereas antagonistic ones were modular. Such expectations are mainly derived from plant-animal networks (Thébault & Fontaine, 2010), and the generality of such a dichotomy has recently been challenged (Michalska-Smith & Allesina, 2019). Analyses performed by Benoît Pichon and Rémy Le Goff (ENS students) have investigated whether we can indeed infer the antagonistic or mutualistic nature of species interactions from network structure. Using large databases of interaction networks covering a range of empirical systems (mainly plant-animal networks, but also host-microbiota ones) and machine learning classifiers, we found that although mutualistic networks are significantly more nested than antagonistic ones, looking at only nestedness and modularity is not enough to discriminate mutualistic and antagonistic networks (Figure IV.2.1a).
However, when considering motif frequencies (*i.e.* small-scale patterns of interactions within the networks), we succeed to classify with 80% of accuracy a network as mutualistic or antagonistic based on only its structure (Figure IV.2.1b and see Article 8 in Appendix). Our classification method linking network structures to natures of interactions could be used in the future to propose whether some indirectly observed interaction networks described using metabarcoding technics consist in mutualistic or antagonistic interactions. For instance, this could be particularly useful for plant endophytes with unclear ecologies.



(b)

Network ecology	# of networks	Prediction			
		Antagonistic network (high conf.)	Antagonistic network (low conf.)	Mutualistic network (high conf.)	Mutualistic network (low conf.)
Anemone-Fish	1	0	0	1	0
Bacteria-Phage	17	12	1	3	1
Herbivory	23	4	6	9	4
Host-Parasitism	77	46	14	10	7
Host-Parasitoid	10	4	1	4	1
Mycorrhiza	9	3	1	3	2
Plant-Ant	3	0	1	1	1
Pollination	174	7	7	139	21
Seed dispersal	29	6	0	20	3
Antagonistic network	127	53 %	17 %	20 %	10 %
Mutualistic network	216	7 %	4 %	76 %	13 %

Figure IV.2.1: Classifying mutualistic *versus* antagonistic networks based on their structure: (a) Projection of the empirical networks on the two principal components (PC1 and PC2) obtained using principal coordinate analysis (PCA) on global metrics (*i.e.* nestedness, modularity, connectance, and network size). Antagonistic (in yellow) and mutualistic networks (in green) are mixed on the projection, meaning that they cannot be separated based on a nested/modular dichotomy. (b) Conversely, when using motif frequencies, the nature of interactions (antagonistic *versus* mutualistic) of the empirical networks can be predicted using artificial neural networks. For each type of network ecology, we indicated the number of networks and whether they are predicted to be mutualistic (resp. antagonistic) with high/low confidence ("conf."). These analyses were performed by Benoît Pichon (see Article 8).

Many models for detecting transmitted microbes have already been developed, but two main points require further attention. First, event-based approaches, like ALE and HOME, rely on strong hypotheses of host-microbe evolutions to propose an evolutionary history for potentially transmitted microbes. HOME models only the process of nucleotidic substitution and horizontal host-switching, and neglects other important processes, like duplications. Conversely, ALE considers more processes (host switches, duplications, losses), but does not directly model DNA sequence evolutions, such that applying ALE when the number of segregating sites in the alignment is low can be tricky (Article 3). ALE is therefore very efficient at inferring vertical transmission when using long and variable DNA sequences (Dorrell *et al.*, 2021), but can be limited when using DNA metabarcodes that are slowly evolving. Further model developments could also include other processes like incomplete lineage sorting or hybridization (Boussau & Scornavacca, 2020) that can be particularly relevant for modeling the transmission of host-associated microbial communities. Second, both global-fit and event-based methods are actually measuring a pattern of cophylogeny. However, a cophylogenetic pattern between a microbial symbiont and its hosts may not be directly linked with strict vertical transmissions (de Vienne *et al.*, 2013). For instance, preferential host switching can generate congruent phylogenies (host-shift speciations; de Vienne *et al.*, 2007) and heterogeneous pools of microbes can also create a pattern of cophylogeny if closely related hosts are in contact with similar pools of microbes (Amato *et al.*, 2019). To exclude these processes, model validations must be carried: one may test for preferential host switching (*i.e.* whether host switches are indeed inferred to be more likely between closely related host species; Article 2) or may randomize the observations within geographic pools (which should break the cophylogenetic patterns if due to vertical transmissions; Article 3). In addition, one should verify that the ages of the host and microbial clades are matching (de Vienne *et al.*, 2013). Dating the age of a microbial clade can be particularly challenging as robust phylogenetic trees and absolute calibration points are rarely available, but one can at least check that the numbers of segregating sites in the microbial alignments across the host clade are coherent with the expected substitution rate of the considered metabarcoding genes. Therefore, inferring the evolutionary history of host-associated microbes is currently achievable when applying event-based approaches and model validations to exclude confounding processes.

Compared with models of transmissions, models of network evolution would represent a more integrative framework to consider the evolution of host-microbiota interactions.
First, one can model the evolution of the presence/absence of interactions. Such models would require hypotheses on how interactions evolve: for instance, we can assume that interactions with microbes (resp. host) are overall conserved in a host (resp. microbe) lineage and that interactions can randomly be acquired or lost through time (anagenetic changes). At speciation events, we can consider two scenarios: either the interactions are conserved by the two daughter lineages or they are split (cladogenetic changes). A

recent model developed by Braga *et al.* (2020) proposed an approach for reconstructing ancestral host-symbiont interactions: in short, they considered that the host repertoire of a symbiont (*i.e.* the sets of hosts a symbiont can interact with) might stochastically change by acquiring or losing hosts over time (anagenetic changes). Using Bayesian inference, they can reconstruct the ancestral host repertoire of the symbiont clade. Although this approach does not consider host speciation ('host units' are considered to be high taxonomic levels, *e.g.* classes or families, such that the MRCA of the symbionts is younger than the most recent host speciation), it represents an interesting approach for reconstructing ancestral clade-based networks (Guimarães, 2020). Note also that symbionts and hosts can be reverted in the model to reconstruct the ancestral microbial repertoire of a host clade. Future developments of the models could include biogeography (such that a host and its symbiont must be in sympatry to interact) and allow interactions between symbionts (*e.g.* competition) (Braga *et al.*, 2020).

Second, one can model the evolution of microbial abundances in the host-associated microbiota. For instance, we have started to develop such a model of the evolution of the abundances of a set of microbial units (OTUs) on a host phylogeny. We assume that p OTUs are present in the microbiota of n host species (for which a robust phylogeny is available) and that the divergences between these microbial OTUs are anterior to the MRCA of all the host species (*i.e.* there is no microbial divergence more recent than the root of the host phylogeny). To fulfill this hypothesis, the OTU delineation can correspond to broad delineations (*e.g.* 90% OTUs) or to high-level taxonomic delineations (*e.g.* bacterial phylum). Then, we assume that, from ancestral microbial abundances at the MRCA of the host clades (X_0), the OTU absolute abundances evolve on the host phylogeny according to a multivariate Brownian motion model, *i.e.* that the evolutions of the OTU abundances are functions of their own variance and their covariance with other OTUs. In other words, we account for the effect of possible negative or positive interactions between microbes (depicted by a variance-covariance matrix R). Under this multivariate Brownian motion model, the joint distribution of all microbial abundances across all host species follows a multivariate normal distribution and we can compute the likelihood, *i.e.* the probability of observing the OTU absolute abundances at present given the parameters of the models, X_0 and R. One can relatively easily estimate the parameter values X_0 and R that maximize this likelihood (Clavel *et al.*, 2015). However, with the current metabarcoding technics characterizing microbiota compositions, we do not have access to absolute microbial abundances, but only to relative microbial abundances in each extant host species (*i.e.* technically, we only have information about $p - 1$ relative abundances). In terms of likelihood, this strongly complicates its computation: one has to integrate the likelihood over all the possible values of absolute abundances (formula not shown), which becomes too numerically intensive. To tackle this issue, we are currently developing an alternative inference method based on artificial neural networks. In short, the idea is to simulate on a given host phylogeny many microbiota evolutions according to our model (with known X_0 and R) and to train the artificial neural networks to link relative microbial abundances at present (our input data) to the generating parameters

X_0 and R (the parameters we want to estimate). Some preliminary analyses conducted by Loréna Duret (ENS L3 student) are rather encouraging (see Figure IV.2.2). After validating our method using simulations, we would like to apply our approach to the gut microbiota of mammals or birds, which would allow us to estimate their ancestral microbiota composition and infer what are the microbial units evolving in a correlated way.

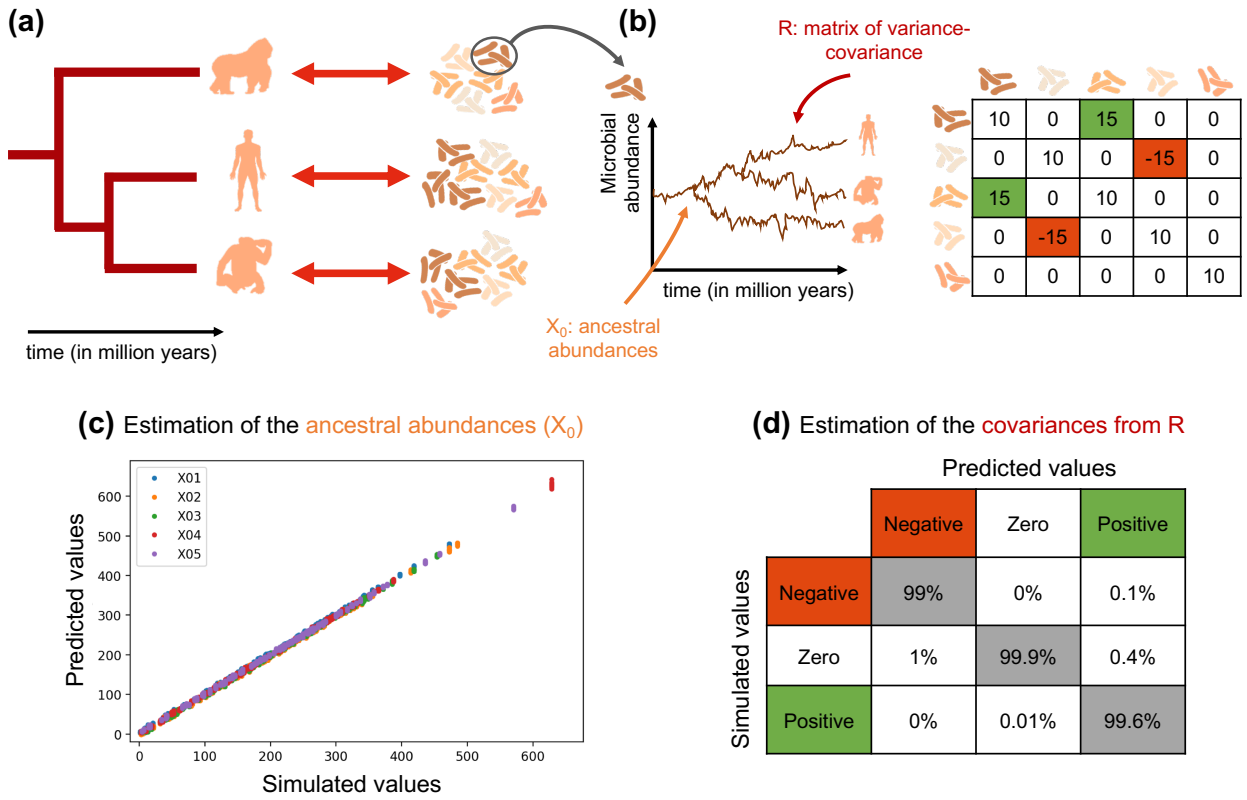


Figure IV.2.2: Modeling the evolution of the relative abundances of a set of microbial units on a host phylogeny. a-b: The model assumes that microbial absolute abundances evolve on a host phylogeny from initial abundances at the MRCA of the hosts (X_0). Microbial abundances evolve according to a multivariate Brownian motion described by the covariance matrix R (some microbial units are positively correlated (green) while others are negatively correlated (red)). We assume that we only have access to the relative microbial abundances of the extant host species. c-d: Example of an inference using artificial neural networks (Keras library in Python): 10,000 simulations were realized with $p = 5$ OTUs and $n = 100$ host species and the neural network was trained using 80% of them. Next, we tested the performance of the trained neural network by testing it on 20% of the remaining simulations: the ancestral abundances (X_0) are well recovered by our approach (c) as well as positive and negative covariances (R) between microbial units (d). These analyses were performed by Loréna Duret.

Altogether, models of network evolutions would also allow us to more efficiently link (sym)biotic interactions with species diversification. Indeed, when trying to link the diversification of Glomeromycotina with land plants (Article 5), we only had limited options: we looked whether temporal trends of Glomeromycotina matches with plant diversity both qualitatively (by comparing the trends) or quantitatively, by fitting environment-dependent birth-death models (with the fossil land plant diversity) or by correlating present-day speciation rates with current patterns of interactions with plants.

However, if one can reconstruct past host-microbe interactions, one will be able to directly test for a plethora of more precise hypotheses by designing trait-dependent diversification models inspired from the state-dependent speciation and extinction (SSE) models: Are generalist fungi experiencing more speciation than specialist ones? Have particular associations with plants (*e.g.* Angiosperms) driven the diversification of Glomeromycotina? Do microbial symbionts mediating cytoplasmic incompatibilities (*e.g. Wolbachia*) spur the speciation of their associated hosts?

However, one has to keep in mind that inferring past evolutionary histories from only present-day observations can be challenging. Many different processes can leave the same pattern at present, and some past events can leave no trace at all in the extant data. Therefore, models may fail at reconstructing past evolutionary histories. These concerns are well-known for ancestral trait reconstructions (Harmon, 2017) and species diversification models (Lambert & Stadler, 2013; Louca & Pennell, 2020), but likely also apply to models of microbial transmissions, where many reconciliated scenarios might have equal likelihoods (asymptotic unidentifiability; Solís-Lemus *et al.*, 2016). In addition, complexifying models (*i.e.* adding more parameters to estimate) when the amount of information is limited (like in DNA metabarcodes) creates practical identifiability issues (*i.e.* not enough information for correctly estimating the parameters). Therefore, complexifying the models of host-microbiota evolutions should advance in pair with our ability to generate adequate data for applying these models. Finally, a last important step in modeling is to perform model validations, to ensure that the proposed model correctly predict most of the aspects of the biological systems. Further works should be pursued in this direction to more systematically assess the robustness of the conclusions extracted from models fit to empirical systems.

2.3. What theory for the evolution of host-microbiota interactions?

In this section, we will present several theoretical models that have been proposed for the evolution of host-microbiota interactions and discuss their support, in the light of some of the findings of my PhD.

2.3.1. The hologenome theory of evolution

Microbial communities form intimate units with their animal and plant hosts. The most striking examples are the intimate relationships between eukaryotes and their organelles, like the mitochondria or the chloroplasts, that derived from a symbiosis between eukaryotic ancestors and their endosymbiotic bacteria ensuring respiration or photosynthesis respectively (Margulis, 1970). Given that extant animals and plants also intimately associate with a plethora of microbial symbionts, the term holobiont has been proposed to describe the units formed by a host and its microbial symbionts (Margulis

& Fester, 1991); and a hologenome designates the genetic material of both the hosts and its microbes. There is plenty of evidence that the animal or plant organism cannot function 'on its own', but instead, that the resulting holobiont forms a unit with emerging properties (see Introduction; Bordenstein & Theis, 2015). In addition, both changes in the host genomes (*e.g.* mutation) or in the microbial genomes (*e.g.* horizontal acquisition of new genes) play fundamental roles in the holobiont adaptation, with the rapid genomic changes of the microbial symbionts allowing the holobiont to rapidly adapt to changing environments (Rosenberg & Zilber-Rosenberg, 2018; Simon *et al.*, 2019). Therefore, given that microbial symbionts can be transmitted from host generations to host generations, the hologenome theory of evolution has stated that the holobiont may be seen as a unit of selection (Zilber-Rosenberg & Rosenberg, 2008). In other words, if holobionts form stable interactions that convergently and positively contribute to the holobiont fitness, they can form selectable units.

They are some evidence of selection at the level of the holobiont, in particular in some insect-bacteria systems or plant-endophytes systems (Clay *et al.*, 1993; Moran *et al.*, 2019). In such systems, microbial symbionts are faithfully transmitted across generations and the fitnesses of both parties are aligned, such that both the host and its microbes are under selective pressure to increase the reproductive success of the holobiont as a whole (Moran & Sloan, 2015). However, in many systems, evidence for holobiont-level selection is rather scarce (Moran & Sloan, 2015; Douglas & Werren, 2016). Therefore, three main criticisms have emerged against the generality of the hologenome theory of evolution because (i) holobionts are frequently not transmissible units, (ii) selective pressures are often not convergent in the different parties of the holobiont, and (iii) the holobiont may not be the right scale to study host-microbe interactions.

First, to be efficient, holobiont-level selection needs the holobiont to be conserved. Indeed, if host-microbe interactions are not durable over long-time scales, the selection pressures on the hosts and the microbes would likely be decoupled (Article 2). The large heterogeneity of the microbiota composition within host species and its important temporal variability suggest that at best, only a small part of the microbial symbionts are conserved over long time scales. In primates, our analyses (Articles 1 and 3) suggest that at best 10% of the bacterial gut symbionts are transmitted. Despite mixed transmission routes at the level of the whole microbiota, the 'restricted holobiont' formed by the hosts and its transmitted microbes might still act as a unit of selection. However, even if microbes are faithfully transmitted, another important factor is the ratio between generation times: the host generation time has to be sufficiently short regarding the microbial evolutionary timescales for selection at the holobiont level to occur (van Vliet & Doebeli, 2019).

Second, holobiont-level selection is particularly difficult to assess, given that measuring the selective pressures upon the host-associated microbes can indeed be particularly challenging (Mushegian & Ebert, 2016), such that the recurrence of holobiont-level selec-

tion in itself is questioned. Indeed, the fact that microbes can confer selective advantages to their hosts does not mean that the selection of the microbial traits acts as the level of the holobiont. For instance, detoxifying bacteria are often acquired by animal hosts to detoxify their diet, but the bacterial detoxifying traits are selected because of the toxin in the environment, and not because of the advantages it confers to the holobiont (Suzuki & Ley, 2020). In other words, some microbes can increase the host/holobiont fitness without being directly selected to do so. In addition, even when microbes are transmitted and their fitnesses aligned with that of their hosts, conflicts are rampant and cheating strategies often emerge (Moran & Sloan, 2015; see next section).

Third, are holobionts always forming biologically-relevant units? As stated by Zilber-Rosenberg & Rosenberg (2008), the holobiont fits within the framework of the 'superorganism'. In some cases, like animals with stable gut microbiota, or isolated plants, the resulting holobiont can indeed be a 'distinct biological entity, which forms of itself a complete whole' (Theis *et al.*, 2016). However, such a framework does not apply when considering animals that host transient (but functionally active) microbes or plants that are interlinked through shared mycorrhizal fungi ('wood-wide-webs'; Article 7) or by physical contacts (Vannier *et al.*, 2018). The extreme case corresponds to mycoheterotrophic plants which rely both directly on their fungal partners and indirectly on the surrounding autotrophic plants for carbon supply: What are the frontiers of the holobiont in such cases? By trying to propose a framework of 'superorganism', the concept of holobionts might actually mask the idea that what really matter in host-microbiota systems are the interactions by themselves. Indeed, we might rather stop considering individuals on their own, but instead, consider the (macro or micro)organisms jointly with their plethora of biotic interactions that define their functioning and their evolution.

2.3.2. Host-microbiota interactions, a reciprocal exploitation?

The idea of having a holobiont-level selection contrasts with the individual-centered or gene-centered views of evolution (Dawkins, 1976). Indeed, it is rather frequently argued that each interactor will often be under selective pressure to increase its own fitness irrespectively of the fitness of its partner (Figure IV.2.3a), in a selfish way (Sachs *et al.*, 2004; Queller & Strassmann, 2018). Even in the most faithfully transmitted and intimate partnerships, like the eukaryotic mitochondria, cytonuclear conflicts regularly arise when the eukaryotic nuclear genome and the mitochondrial genome are under opposite selective pressures (Saumitou-Laprade *et al.*, 1994). In this mindset of conflicts, some authors have then argued that host-microbe interactions should be seen as reciprocal exploitations (Law & Dieckmann, 1998): many animal-microbes or plant-microbes interactions are indeed consumer interactions where each species exploit a resource produced by its partner (Antonovics *et al.*, 2015). In addition, there are plenty of examples of selfish strategies, where hosts exploit their microbial symbionts only when needed (*e.g.* facultative mycorrhizal plants) or farm/enslave their microbes (*e.g.* gut bacteria of

ruminants, rhizobia), sometimes resulting in a 'symbiotic prison' for the microbes (Kiers & West, 2016). Finally, frequent cheating strategies emerge among host-microbe interactions, when one of the partners retrieves a higher benefit from the interaction at the expense of the other (Article 6).

Such host-microbe exploitative interactions are also thought to be more likely to emerge than immediate mutualisms (Sørensen *et al.*, 2019), such that most apparent mutualisms that we observe today would have arisen as exploitive parasitisms that later evolved as more mutualistic interactions thanks to trade-offs. For instance, host exploitation can easily start as the host capture and exploit a beneficial microbe. Then, as microbes consequently build up defenses to limit the negative effects of host exploitation, the costs of these defenses might become too important in the free-living state. Consequently, these microbes would have higher fitness when associating with hosts than when free-living, resulting in an apparent mutualism (Sørensen *et al.*, 2019). Similarly, mutualism can also evolve from microbial parasitism: Sachs *et al.* (2011) found that >75% of the host-associated bacterial symbionts are derived from parasitic ancestors. In such cases, faithful transmissions or absence of host choice would align the fitnesses of the host and its microbe, resulting in a transition from parasitism to mutualism (Figure IV.2.3e). In both scenarios, the host or the microbe may evolve additional dependences on their partners for other functions (evolutionary addiction), reinforcing the exploitative interactions, that will appear, over time, as a mutualistic symbiosis (Selosse *et al.*, 2014; Moran *et al.*, 2019).

Then, once mutualism is established, in a second time, mechanisms generally evolve to guarantee its stability (Sachs *et al.*, 2004; Sørensen *et al.*, 2019). Most animal-microbes or plant-microbes interactions are indeed not naïve mutualisms, and all interactors have generally developed strong controls to prevent cheating, such that the mutualistic interactions are often limited to a small range of physiological conditions where hosts and microbes tolerate each other (Figure IV.2.3b; Article 6). One could argue that in many interactions there is no apparent cost, and that no mechanisms preventing cheaters would need to be selected. However, as long as there is an intimate and durable interaction, parasitism is rampant. In host-associated microbiota, there are many examples of intimate microbes becoming parasites. For instance, several saprotrophic fungal lineages like the *Rhizoctonia* (Cantharellales) became plant pathogens (Veldre *et al.*, 2013), and similarly, gut commensal bacteria, like *Escherichia coli*, often horizontally acquire genes that turn them into pathogens (Wirth *et al.*, 2006). One could thus propose a 'Murphy's law of symbiosis', a rather pessimistic view of host-microbe interactions: "as long as two organisms are intimately and durably associated, anything that can go wrong will go wrong". Though exaggerated, this might explain why mechanisms preventing cheaters are widespread in most organisms (Sachs *et al.*, 2004).

In such an exploitative framework, and despite strong controls between partners,

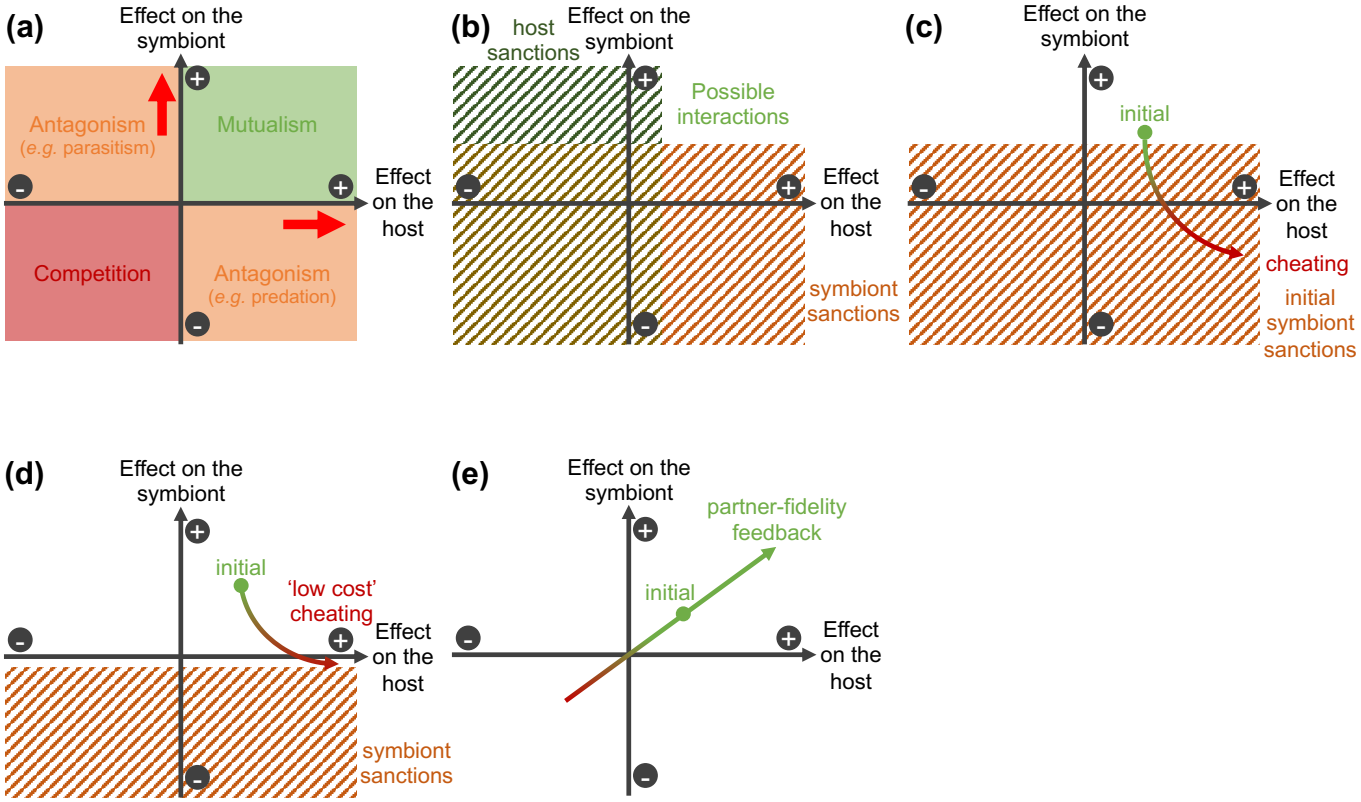


Figure IV.2.3: Host-microbe interactions, a reciprocal exploitation? (a) In a general case, selection is expected to act at the level of the individual, such that each partner (the host or the symbiont) is expected to be under selective pressure to increase its own fitness (red arrows). (b) If the hosts and their microbial symbionts have mechanisms preventing cheating (*e.g.* sanctions represented by hatched areas), the host-microbe interactions are limited to a small range of physiological conditions where hosts and microbes tolerate each other. These ranges of conditions can also vary according to the biotic and abiotic conditions (Frederickson, 2017). (c) If the host can evolve strategies to counteract the symbiont sanctions, host cheating can emerge (the host cheater increases its fitness while decreasing the fitness of its symbiont). (d) Under particular conditions, the cost of the host cheater for the symbiont might be low or negligible, such that "low cost" cheaters can more easily emerge and be tolerated. For instance, mycoheterotrophic cheating in plants only evolves in understory vegetations, when access to the light is low and carbon is likely not the limiting factor for the large surrounding autotrophic trees and their associated fungi (Gomes *et al.*, 2019a; Article 6). (e) In some conditions, *e.g.* if symbionts are faithfully transmitted, the fitnesses of both the host and the symbiont are aligned (partner-fidelity feedbacks) and increased mutualism is selected.

cheaters often manage to emerge among host-microbiota interactions (Figure IV.2.3c; Article 6). Are they nevertheless ecological and evolutionary successes? In the arbuscular mycorrhizal symbioses, mycoheterotrophic plants appear to be isolated from the global core of interactions (Article 6) and we noticed that most mycoheterotrophic lineages are relatively young and species-poor. One hypothesis is that cheating can evolve frequently in mutualisms but that they are evolutionarily unstable (Douglas, 2008); in other words, cheating would be an evolutionary dead-end. One way to test this hypothesis would be to measure whether mycoheterotrophic cheating lineages have lower diversification rates than lineages that have remained mutualistic autotrophs. We have started investi-

gating this in Neottieae, an orchid clade that contains autotrophic, mixotrophic (partially mycoheterotrophic), and mycoheterotrophic species. Preliminary analyses tend to find lower diversification rates in mycoheterotrophic clades, suggesting that cheating in this mutualism would be an evolutionary dead-end (Figure IV.2.4).

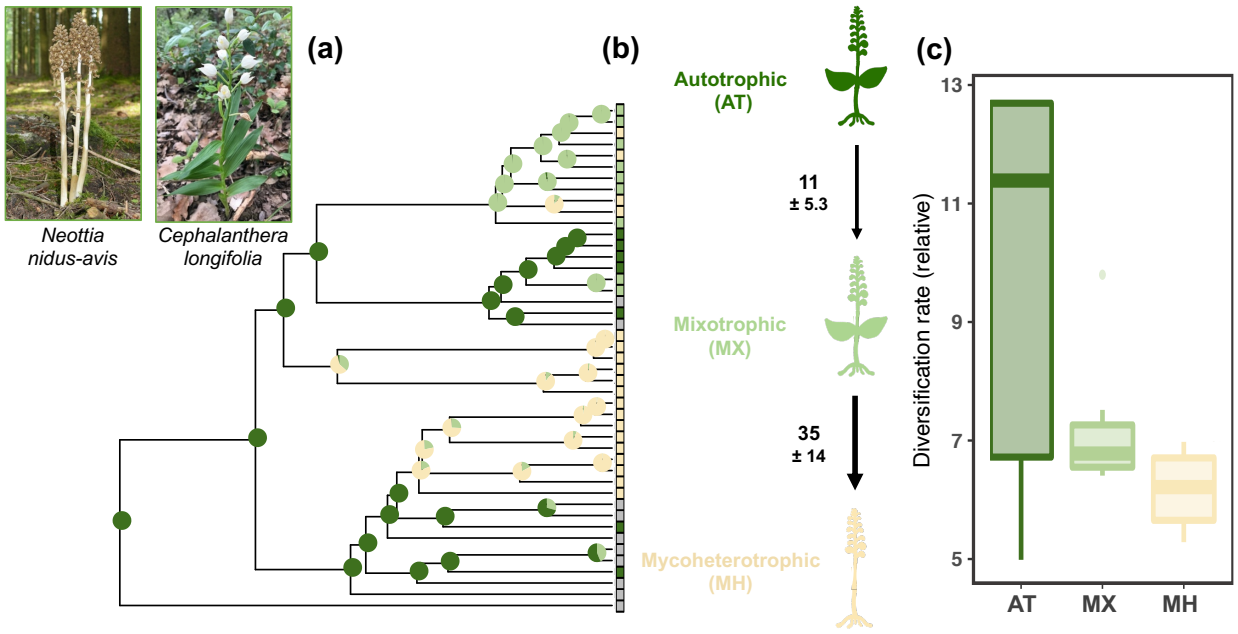


Figure IV.2.4: Mycoheterotrophy in Neottieae orchids might be an evolutionary dead-end. (a) Photos of a mycoheterotrophic (*Neottia nidus-avis*; Bernd Haynold) and mixotrophic (*Cephalanthera longifolia*; Benoît Perez-Lamarque) species. (b) Ancestral state reconstruction of the orchid nutrition using BayesTrait (Meade & Pagel, 2007). We found that ancestral Neottieae were autotrophic species that repeatedly evolved mixotrophy and mycoheterotrophy in a unidirectional way. Given the rates of transition, mixotrophy appears to be an unstable state in Neottieae, that rapidly shifts toward mycoheterotrophy. (c) Present-day diversification rates of the Neottieae estimated using ClaDS (Maliet *et al.*, 2019) are indicated as a function of the orchid nutrition: extant mixotrophic and mycoheterotrophic species appear to have lower diversification rates than autotrophic ones.

Therefore, host-microbiota interaction can be seen in the framework of reciprocal exploitation. Nevertheless, resulting mutualisms are not on the verge of breakdown yet (Frederickson, 2017), thanks to efficient and widespread constraints (Article 6). Consequently, mutualistic interactions are likely more evolutionary stable than parasitic ones, which may explain why cophylogenetic patterns in mutualisms are more frequent than in parasitisms (de Vienne *et al.*, 2013).

2.3.3. The host-associated microbiota, an ecosystem on a leash

A lot of emphases have been put on the functions ensured by the microbiota for the hosts (see section 1.2 in Introduction), because historically, host-associated microbiota have been mainly studied for the functions they provide to the hosts. However, for un-

derstanding the ecology and evolution of host-microbe interactions, host-associated microbiota has to be studied as an integrated biological system rather than through the lens of the host benefits only. Indeed, host-associated microbiota are often complex and permanently changing ecosystems modulated by their host. Therefore, it has been argued that one should use the tools and theories developed from community ecology to study the assembly, diversity, and stability of host-associated microbiota (Douglas & Werren, 2016; Koskella *et al.*, 2017). Compared with classical ecosystems, host microbiomes have a lifetime that can be relatively short and experience control from the host. Such ecosystems encapsulate at the same time microbe-microbe interactions, host-to-microbe interactions, and microbe-to-host interactions (Foster *et al.*, 2017). While the latter are generally well-studied, the two former are often overlooked. In terms of evolution, in species-rich microbiota like in the animal guts, microbes are under selective pressure to compete with other microbes within the microbiome, while "hosts evolve to keep the ecosystem on a leash" (Foster *et al.*, 2017). In other words, microbes are often not directly under pressure for benefiting their host, but rather for persisting within them, whereas hosts are under strong selective pressure for having efficient control mechanisms over their microbiota to ensure an overall beneficial outcome (Foster *et al.*, 2017; Moeller & Sanders, 2020).

Host controls act at three different levels: (i) by controlling microbiota immigration, (ii) by compartmentalizing microbes, and (iii) by monitoring them (Foster *et al.*, 2017). Controlling microbiota immigration includes the mechanisms favoring faithful transmissions or filtering of the environmental microbes, compartmentalization ensures a more specific screening of the microbes (*e.g.* in the gut diverticula or at the level of the mycorrhizal structure; Chomicki *et al.*, 2020), and monitoring includes all the mechanisms (see section 2.4 in Introduction) rewarding microbial traits beneficial for the hosts (the "carrot") and punishing non-beneficial ones (the "stick") (Shapira, 2016). Many of these mechanisms are precisely controlled by specific host genes, as illustrated by the shifts in gut microbiota compositions associated with particular human mutations (Goodrich *et al.*, 2016). They are likely to be phylogenetically conserved and consequently can result in a pattern of phylogenetic signal in microbiota compositions (phylosymbiosis; Article 4). Although many of these host controls are expected to be under selection, some traits, *e.g.* the ones responsible for the filtering of some microbes during microbiota assembly (pH, antimicrobial secretions, ...), might be evolving neutrally.

Modeling host-associated microbiota as "an ecosystem on the leash" can successively describe many aspects of the animal gut microbiota or the plant root microbiota, including the holobiont-level selection in some contexts (see section 2.3.1) and the propensity of exploitation among these interactions (see section 2.3.2). In the cases of mycorrhizal fungi linking several plants, it can be generalized in a model of an "ecosystem on leashes" where the microbes can simultaneously be controlled by several hosts, resulting in an evolving network of interactions (but see Box 2; Figure IV.2.5). Alternative models of host-microbiota systems include models of host controls only, where microbes are iso-

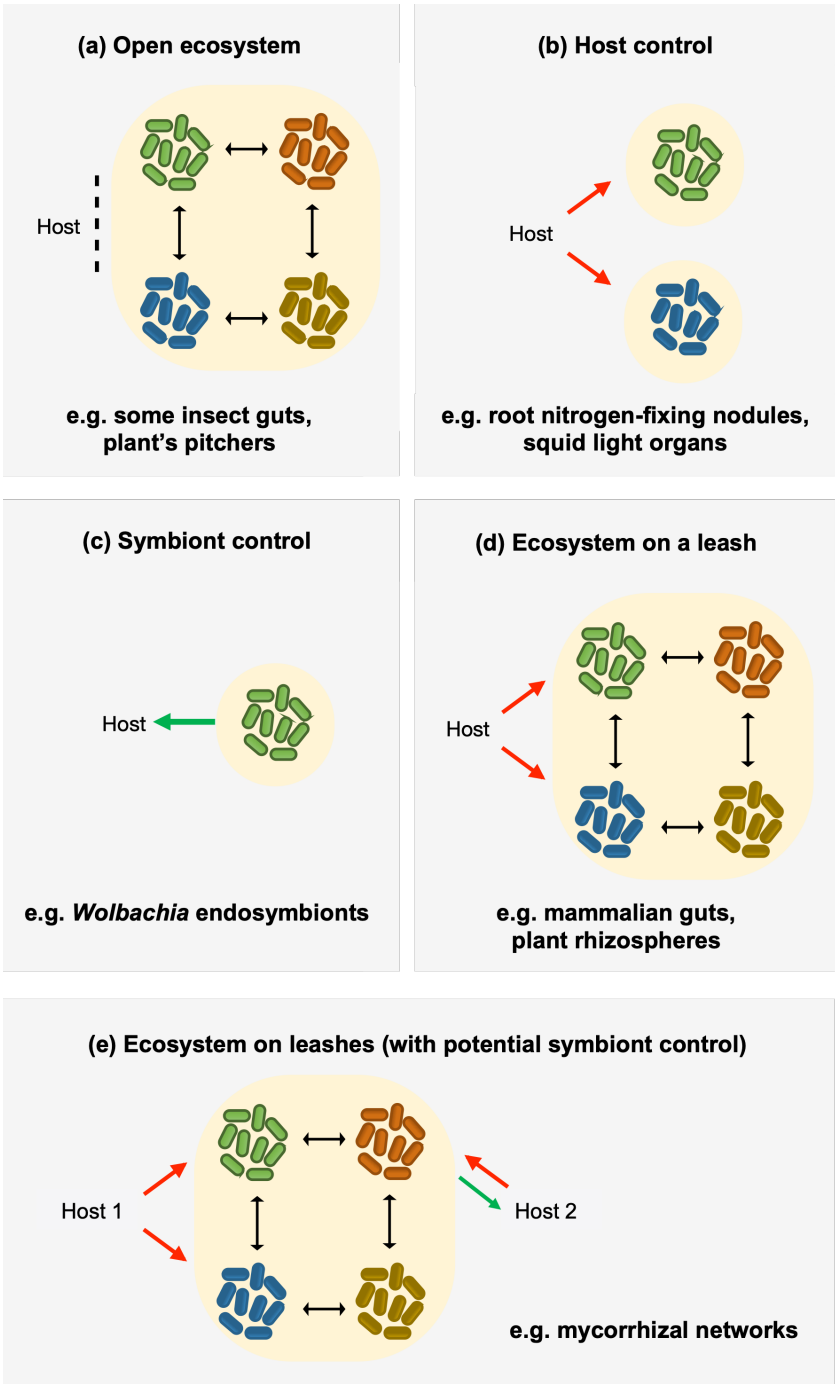


Figure IV.2.5: Different models of host-associated microbiota. (a) Open ecosystem model, where the host microbiome can be colonized by microbes without particular controls from the host. (b) Host control model, where the host tends to compartmentalize its symbionts and exerts strong controls upon them. (c) Symbiont control model, where a symbiont manipulates its host. (d) Ecosystem on a leash model: the hosts exerts controls upon its symbiotic microbes (*e.g.* by rewarding beneficial microbes and punishing cheaters), while microbes are mainly competing for surviving in the microbiome. Although often poorly characterized, such microbe-microbe interactions likely shape the composition and the functioning of the microbiota (Douglas & Werren, 2016). Future works will likely unravel the roles of functional redundancy and character displacement in these communities (Louca *et al.*, 2016; Foster *et al.*, 2017; Brochet *et al.*, 2021). (e) Model of an ecosystem on "leashes" with symbiont controls. Symbionts can simultaneously interact with several hosts, conferring them more leverage in their interactions with the hosts (*i.e.* symbiont controls). This figure is adapted from Foster *et al.* (2017).

lated from each other and specifically modulated by the host (*e.g.* the rhizobium in the root nodules of Fabaceae – an extreme case of host exploitation), models of symbiont controls (where a (parasitic or cheating) microbe manipulate the host for its own benefits, *e.g. Wolbachia* and other endosymbionts manipulating arthropods), or a model of an open ecosystem, where the host microbiome can be colonized by microbes without any control from the host (*e.g.* the gut microbiota of the arthropods mainly derived from their diet).

Box 2: Some limits when comparing animal and plant microbiota:

The works of my PhD mainly looked at the bacterial microbiota of animal guts and the root mycorrhizal communities of plants. Both systems were simultaneously considered in this Discussion, despite the existence of fundamental differences between them that prevent blind comparisons.
First, mycorrhizal fungi have generation times much longer than bacterial generation times, such that the evolutionary timescales of the mycorrhizal fungi and their plant hosts (especially the annual ones) are not as decoupled as the evolutionary timescales of the gut bacteria and their animal hosts.
Second, gut microbiota of animals are internalized within host organisms, whereas root microbiota are more exposed to the environments, as most of the mycorrhizal fungal organisms are freely exploring the surrounding soils. Therefore, we expect plant microbial symbionts to be less frequently specialized towards their hosts than the animal gut symbionts that are more likely to adapt to the specific gut microbiome conditions. Indeed, while microbes associated with mammals are mainly order-specific (Song *et al.*, 2020), we found a lot of microbial sharing between plant species that diverged >300 million years ago (Article 7). In addition, root-associated fungal communities appeared to be much more dependent on the abiotic environmental conditions (Article 7) than the animal communities that are generally quite resilient (Amato *et al.*, 2019). Thus, because of their internalization and the resulting host-restrictiveness of many of their associated bacteria (especially in mammals), animal gut microbiota are well modeled as "an ecosystem on a leash" where host controls (*e.g.* transmission mechanisms and monitoring) importantly shape microbiota composition and evolution (Figure IV.2.5). Conversely, plant roots do not offer a well-defined, separated ecosystem for their associated microbes, but rather offer an interface where a multitude of interactions with mycorrhizal fungi can take place in a compartmentalized way (there is generally a specific interaction structure, the mycorrhiza, between one plant and one fungus). Consequently, both host and symbiont can choose and monitor the interaction, *i.e.* mycorrhizal fungi have more leverage than gut bacteria over their hosts. Thus, the plant-mycorrhizal networks are rather well modeled as "an ecosystem on leashes" with important symbiont controls (Figure IV.2.5).

To conclude, this integrative framework will be particularly valuable in the future to explore remaining questions about host-microbiota evolution (Figure IV.2.6), like:

(i) What are the host traits affecting microbiota compositions? Are they actively selecting the microbes (host monitoring) or indirectly affecting the microbial colonization (*e.g.* the traits responsible for host filtering in the microbiome)? Are they under selection?

(ii) How frequently do microbes 'choose' their hosts *versus* passively colonize the microbiome niches? To what extent limitations in microbial dispersion affect host-associated microbiota assembly?

(iii) Are vertically transmitted microbes more likely to be host-restricted? What are the drivers of horizontal transmissions (host-switches)?

(iv) What is the relative importance of the different processes generating phylosymbiosis in host-associated microbiota across animal and plant kingdoms?

(v) Are microbe-microbe competitions negatively affecting the microbial-mediated host functions (because microbes are primarily under selection to survive and not to benefit their host) or positively affecting them (*e.g.* because of character displacement resulting in complementary functions)?

(vi) How frequent are the evolution of new dependences toward microbial partners and *vice versa*? And what are the drivers of such dependences (niche expansion, Black Queen hypothesis, or evolutionary addiction)? Do they affect their diversifications in a predictable way?

(vii) Can host-associated microbes directly spur host differentiation and speciation? Do they coevolve?

(viii) Are alternative models (open ecosystems or models of host or symbiont control, including those resulting from cheating emergences) less stable over long timescales?

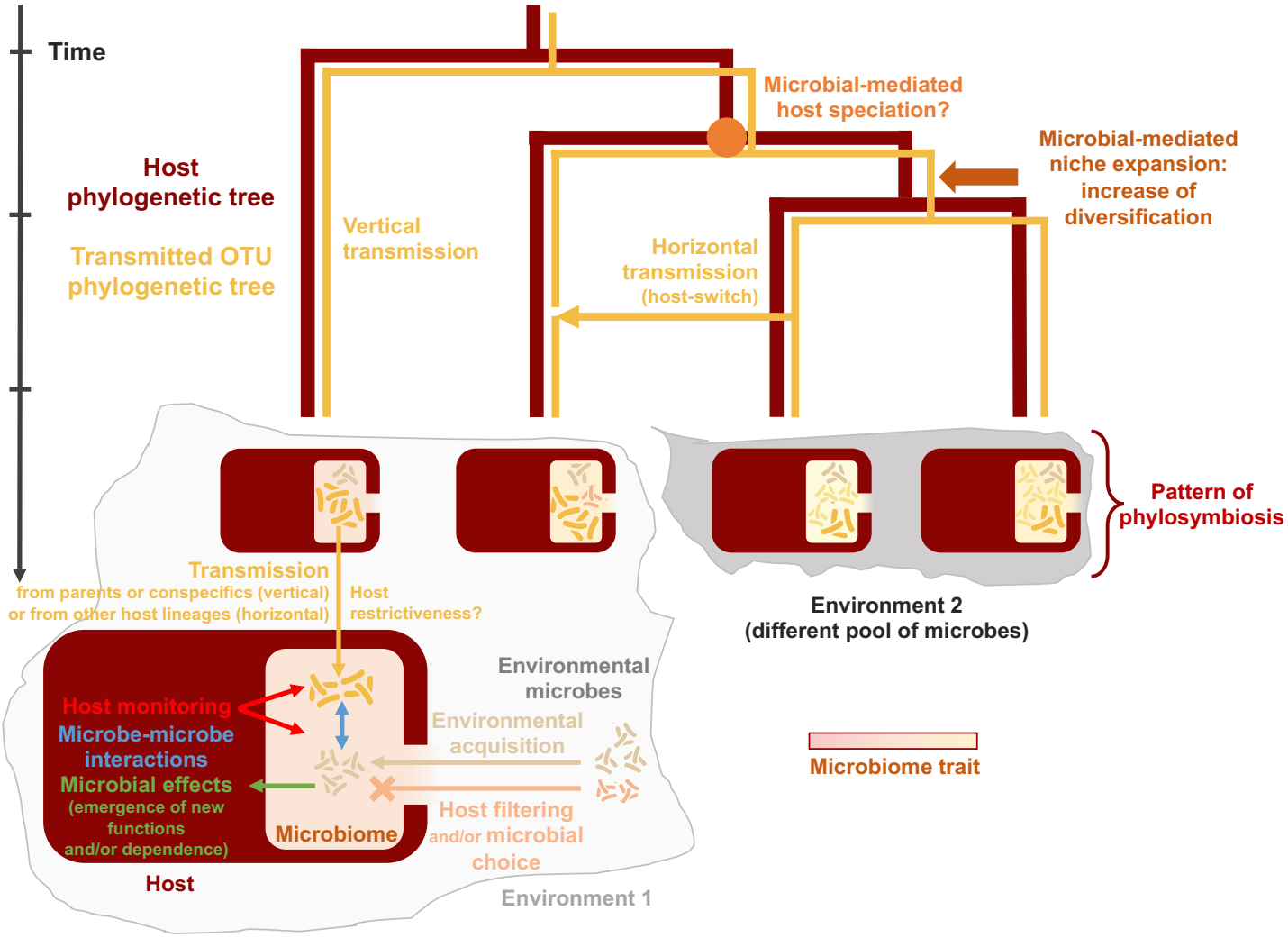


Figure IV.2.6: An integrative framework to study the evolution of host-microbiota interactions. Host organisms (brown rectangles) present a microbiome niche (more or less compartmentalized and with particular host-mediated conditions, *e.g.* pH or antimicrobial secretions, represented by a gradient of colors). These microbiomes can be colonized by some microbes, either transmitted from other hosts (*e.g.* parents and conspecifics) or acquired from their environment, while other microbes do not colonize it (because of host filtering or microbial 'choice'). Plenty of microbe-microbe interactions shape the functioning of the whole microbial community (blue arrow), which participates in the host functioning (green arrow). These microbes and their resulting effects are (more or less specifically) controlled by the hosts (host monitoring, which prevents cheating), and reciprocally, microbes can also sometimes have leverage over their host by modulating their effect (bidirectional controls). Over long timescales, microbiota compositions might be affected by changes in host traits (host filtering), microbial vertical and horizontal transmissions, or changes in the environmental pools of available microbes (represented here by two landmasses, 'environment 1' and 'environment 2'), which can generate a pattern of phylosymbiosis. Dependence on both sides can also evolve (*e.g.* because of the emergence of new functions, or through evolutionary addiction). Host-microbe interactions can also impact their diversifications (*e.g.* microbes can eventually spur the host speciation - 'speciation by symbiosis' - or increase host diversification by expanding their niches, and *vice versa*).

2.4. The Anthropocene: a major crisis for host-associated microbiota?

As we have discussed, host-associated microbiota play fundamental roles in the functioning of most animals and plants and such host-microbe interactions result from billions of years of evolution. Since their emergence, multicellular organisms dwell in a microbial world that deeply shapes their ecology and their evolution. There is a range of host-microbe interactions, from labile and opportunistic ones, to durable and mutualistic associations that have coevolved for millions of years. However, in the last century, for the first time, human populations have gained the ability to extract themselves or any other animal or plant organisms from their microbial worlds. In extreme cases, animals and plants can be grown in axenic conditions, but a simple use of antibiotic or antifungal compounds to cure host diseases is generally enough for generating large and durable perturbations of the host-associated microbiota. Consequently, over a few decades, humans have lost a non-negligible part of their microbial symbionts: compared with other great apes, humans have a lower microbial diversity in their gut (Moeller *et al.*, 2014; Gaulke *et al.*, 2018), which is significantly associated with the recent changes of lifestyle, the westernization (Yatsunenko *et al.*, 2012; Nishida & Ochman, 2019). Similar trends are also often found in captive animals (McKenzie *et al.*, 2017). In crops, the intensive use of fertilizers has led to a reduced dependency of the plants toward mycorrhizal fungi (Plenchette *et al.*, 2005), and in anthropogenic ecosystems in general, we observe that plants that are still associated with mycorrhizal fungi tend to present less microbial diversity (Figure IV.2.7).

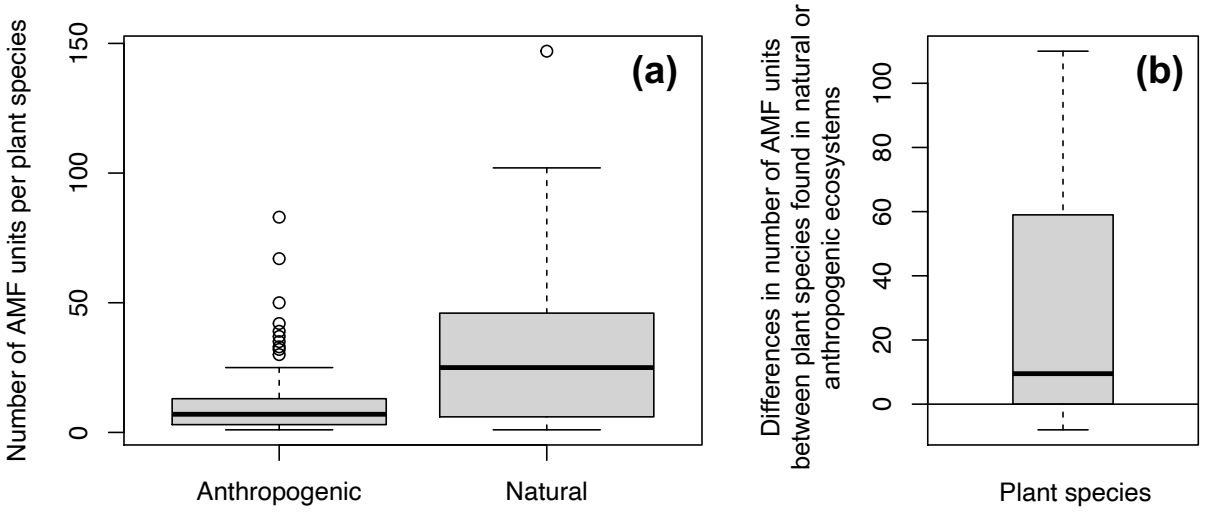


Figure IV.2.7: Decline in the arbuscular mycorrhizal fungal diversity in plants from anthropogenic ecosystems. We investigated in the MaarjAM database (Öpik *et al.*, 2010) the fungal diversity associated with different plant species (total number of arbuscular mycorrhizal fungal 'species-like' units – virtual taxa) found in anthropogenic *versus* natural ecosystems. (a) Arbuscular mycorrhizal fungal (AMF) units are significantly less abundant in plant species from anthropogenic ecosystems. (b) By looking at the plant species present in both anthropogenic and natural ecosystems, we also found an important decline in their AMF diversity (positive differences).

Unfortunately, in humans, eradicating pathogenic microbes has also led to strong impacts on the beneficial ones. Perturbations of the gut microbiota, called dysbiosis, due to westernization, often result in chronic inflammatory diseases (*e.g.* allergy or inflammatory bowel diseases), which have exponentially increased in developed countries over the last decades (Kaplan & Ng, 2017). The hygiene hypothesis has been proposed to describe this idea that a lack of microbial exposure can lead to an incorrect development of the human organism and subsequent new diseases (Rook *et al.*, 2013). One way to deal with this arising issue is to look at the microbes that are associated with chronic inflammatory diseases when absent in human guts, in order to develop therapeutic strategies (prebiotics or probiotics) that promote their presence in the gut microbiota. Dozens of bacterial strains are identified as beneficial gut bacteria with potential therapeutic uses. In collaboration with Claire Cherbuy and Cassandre Bedu-Ferrari from the Institut Micalis (INRAE), we have looked at the evolution of some of these beneficial bacterial species in primates and found that some of them, like *Roseburia intestinalis*, have likely been vertically transmitted during primate evolution (Figure IV.2.8). Therefore, the chronic inflammatory diseases diagnosed in westernized humans result at least partially from the breakdown of their million years of evolution with beneficial gut bacteria (Rook *et al.*, 2013).

To conclude, improving our understanding of the functioning and evolution of host-associated microbiota will give us the opportunity to prevent or limit the deleterious impacts of modern lifestyles on plant-associated or animal-associated microbiota, as well as improve our ability to 'engineer' these microbial communities to remediate them.

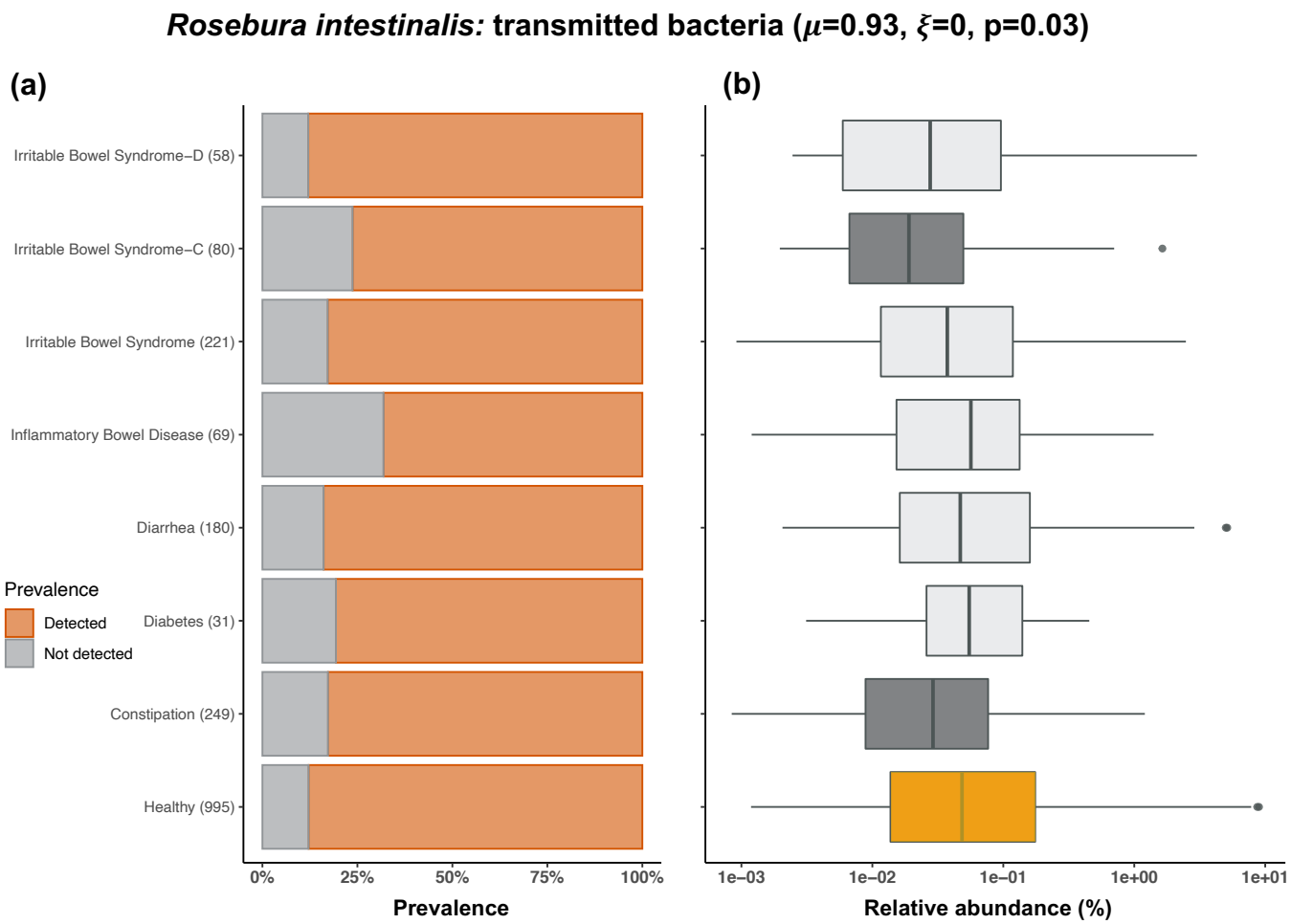


Figure IV.2.8: *Roseburia intestinalis*, a transmitted bacterium that is frequently negatively correlated with westernized human pathologies:. We investigated whether important human gut bacteria were transmitted in primate guts using the data from Amato *et al.* (2019) and HOME (Article 1). Then, we examined the prevalence and abundance of these bacteria in human gut microbiota, using the data from the American Gut Project (McDonald *et al.*, 2018) in the GMrepo database (Wu *et al.*, 2020), as a function of human pathologies. Some of the transmitted bacteria, like *Roseburia intestinalis*, are significantly associated with dysbiosis and human pathologies. Indeed, *Roseburia intestinalis* tends to be less prevalent in individuals presenting inflammatory bowel diseases or syndromes than in healthy humans (a). In addition (b), when present, *Roseburia intestinalis* tend to be significantly less abundant in individuals presenting some pathogenies (colored in dark greys) than in healthy individuals (in yellow).

Appendix

The Article 8 entitled *"Inferring the nature of interspecific interactions based on the structure of ecological networks"* is available through the link `https://bit.ly/3tyf7U9` or by scanning:

and its associated supplementary data are available through the link `https://bit.ly/3tn9dVJ` or by scanning:

References

Akin DE, Borneman WS. 1990. Role of rumen fungi in fiber degradation. Journal of Dairy Science 73: 3023–3032.

Almario J, Jeena G, Wunder J, Langen G, Zuccaro A, Coupland G, Bucher M. 2017. Root-associated fungal microbiota of nonmycorrhizal *Arabis alpina* and its contribution to plant phosphorus nutrition. Proceedings of the National Academy of Sciences of the United States of America 114: E9403–E9412.

Almeida-Neto M, Guimarães P, Guimarães PR, Loyola RD, Ulrich W. 2008. A consistent metric for nestedness analysis in ecological systems: reconciling concept and measurement. Oikos 117: 1227–1239.

Althoff DM, Segraves KA, Johnson MTJ. 2014. Testing for coevolutionary diversification: Linking pattern with process. Trends in Ecology and Evolution 29: 82–89.

Amato KR, G. Sanders J, Song SJ, Nute M, Metcalf JL, Thompson LR, Morton JT, Amir A, J. McKenzie V, Humphrey G, et al. 2019. Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. The ISME Journal 13: 576–587.

Antonovics J, Bergmann J, Hempel S, Verbruggen E, Veresoglou S, Rillig M. 2015. The evolution of mutualism from reciprocal parasitism: more ecological clothes for the Prisoner's Dilemma. Evolutionary Ecology 29: 627–641.

Armstrong EE, Perez-Lamarque B, Bi K, Chen C, Becking LE, Lim JY, Linderoth T, Krehenwinkel H, Gillespie R. 2020. A holobiont view of island biogeography: Unraveling patterns driving the nascent diversification of a Hawaiian spider and its microbial associates. bioRxiv.

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T, Mende DR, Fernandes GR, Tap J, Bruls T, Batto JM, et al. 2011. Enterotypes of the human gut microbiome. Nature 473: 174–180.

Asakuma S, Hatakeyama E, Urashima T, Yoshida E, Katayama T, Yamamoto K, Kumagai H, Ashida H, Hirose J, Kitaoka M. 2011. Physiology of consumption of human milk oligosaccharides by infant gut-associated bifidobacteria. Journal of Biological Chemistry 286: 34583–34592.

Babikova Z, Gilbert L, Bruce TJA, Birkett M, Caulfield JC, Woodcock C, Pickett JA, Johnson D. 2013. Underground signals carried through common mycelial networks warn neighbouring plants of aphid attack (N van Dam, Ed.). Ecology Letters 16: 835–843.

Bago B, Bécard G. 2002. Bases of the obligate biotrophy of arbuscular mycorrhizal fungi. In: Mycorrhizal Technology in Agriculture. Basel: Birkhäuser Basel, 33–48.

Bailly-Bechet M, Martins-Simões P, Szöllősi GJ, Mialdea G, Sagot M-F, Charlat S. 2017. How long does *Wolbachia* remain on board? Molecular Biology and Evolution 34: 1183–1193.

Balbuena JA, Míguez-Lozano R, Blasco-Costa I. 2013. PACo: A novel procrustes application to cophylogenetic analysis (CS Moreau, Ed.). PLoS ONE 8: e61048.

Balbuena JA, Pérez-Escobar ÓA, Llopis-Belenguer C, Blasco-Costa I. 2020. Random Tanglegram Partitions (Random TaPas): An Alexandrian approach to the cophylogenetic gordian knot (M Charleston, Ed.). Systematic Biology 69: 1212–1230.

De Bary A. 1853. Untersuchungen uber die Brandpilze und die durch sie verursachten Krankheiten der Pflanzen, mit Rücksicht auf das Getreide und andere Nutzpflanzen. Berlin: GWF Muller.

De Bary A. 1879. Die erscheinung der symbiose: Vortrag gehalten auf der versammlung deutscher naturforscher und aerzte zu cassel. Trubner.

Bascompte J, Jordano P. 2013. Mutualistic networks. Princeton: Princeton University Press.

Bascompte J, Jordano P, Melian CJ, Olesen JM. 2003. The nested assembly of plant-animal mutualistic networks. Proceedings of the National Academy of Sciences 100: 9383–9387.

Batstone RT, Carscadden KA, Afkhami ME, Frederickson ME. 2018. Using niche breadth theory to explain generalization in mutualisms. Ecology 99: 1039–1050.

Batstone RT, O'Brien AM, Harrison TL, Frederickson ME. 2020. Experimental evolution makes microbes more cooperative with their local host genotype. Science 370: 476–478.

Bauer MA, Kainz K, Carmona-Gutierrez D, Madeo F. 2018. Microbial wars: Competition in ecological niches and within the microbiome. Microbial Cell 5: 215–219.

Beckett SJ. 2016. Improved community detection in weighted bipartite networks. Royal Society Open Science 3: 140536.

Begum N, Qin C, Ahanger MA, Raza S, Khan MI, Ashraf M, Ahmed N, Zhang L. 2019. Role of arbuscular mycorrhizal fungi in plant growth regulation: Implications in abiotic stress tolerance. Frontiers in Plant Science 10: 1068.

Benucci GMN, Burnard D, Shepherd LD, Bonito G, Munkacsi AB. 2020. Evidence for co-evolutionary history of early diverging Lycopodiaceae plants with fungi. Frontiers in Microbiology 10: 1–12.

Berasategui A, Salem H, Paetz C, Santoro M, Gershenzon J, Kaltenpoth M, Schmidt A. 2017. Gut microbiota of the pine weevil degrades conifer diterpenes and increases insect fitness. Molecular Ecology 26: 4099–4110.

Berendsen RL, Pieterse CMJ, Bakker PAHM. 2012. The rhizosphere microbiome and plant health. Trends in Plant Science 17: 478–486.

Bidartondo MI, Read DJ, Trappe JM, Merckx VSFT, Ligrone R, Duckett JG. 2011. The dawn of symbiosis between plants and fungi. Biology Letters 7: 574–577.

Bittleston LS, Pierce NE, Ellison AM, Pringle A. 2016. Convergence in multispecies interactions. Trends in Ecology and Evolution 31: 269–280.

Blum JE, Fischer CN, Miles J, Handelsman J. 2013. Frequent replenishment sustains the beneficial microbiome of *Drosophila melanogaster*. mBio 4: e00860-13.

Bonfante P, Anca IA. 2009. Plants, mycorrhizal fungi, and bacteria: A network of interactions. Annual Review of Microbiology 63: 363–383.

Bonfante P, Genre A. 2015. Arbuscular mycorrhizal dialogues: Do you speak 'plantish' or 'fungish'? Trends in Plant Science 20: 150–154.

Bordenstein SR, Theis KR. 2015. Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. PLoS Biology 13: 1–23.

Bouffaud ML, Poirier MA, Muller D, Moënne-Loccoz Y. 2014. Root microbiome relates to plant host evolution in maize and other Poaceae. Environmental Microbiology 16: 2804–2814.

Boullard B. 1979. Considérations sur la symbiose fongique chez les Ptéridophytes. Syllogeus 19 : 1-58.

Boussau B & Scornavacca C. 2020. Reconciling gene trees with species trees. In: Scornavacca C, Delsuc F, Galtier N, eds. Phylogenetics in the Genomic Era. Authors open access book, 0–23.

Braga MP, Guimarães PR, Wheat CW, Nylin S, Janz N. 2018. Unifying host-associated diversification processes using butterfly–plant networks. Nature Communications 9: 5155.

Braga MP, Landis MJ, Nylin S, Janz N, Ronquist F. 2020. Bayesian inference of ancestral host-parasite interactions under a phylogenetic model of host repertoire evolution. Systematic Biology 69: 1149–1162.

Bright M, Bulgheresi S. 2010. A complex journey: transmission of microbial symbionts. Nat Rev Microbiol. 8: 218–230.

Brochet S, Quinn A, Mars RAT, Neuschwander N, Sauer U, Engel P. 2021. Niche partitioning facilitates coexistence of closely related gut bacteria. bioRxiv: 2021.03.12.434400.

Brucker RM, Bordenstein SR. 2012. Speciation by symbiosis. Trends in Ecology and Evolution 27: 443–451.

Brucker RM, Bordenstein SR. 2013. The hologenomic basis of speciation: Gut bacteria cause hybrid lethality in the genus Nasonia. Science 341: 667–669.

Brundrett MC. 2002. Coevolution of roots and mycorrhizas of land plants. New Phytologist 154: 275–304.

Brundrett MC, Tedersoo L. 2018. Evolutionary history of mycorrhizal symbioses and global host plant diversity. New Phytologist 220: 1108–1115.

Brune A. 2014. Symbiotic digestion of lignocellulose in termite guts. Nature Reviews Microbiology 12: 168–180.

Bukin YS, Galachyants YP, Morozov I V., Bukin S V., Zakharenko AS, Zemskaya TI. 2019. The effect of 16S rRNA region choice on bacterial community metabarcoding results. Scientific Data

6: 190007.

Burns AR, Guillemin K. 2017. The scales of the zebrafish: host–microbiota interactions from proteins to populations. Current Opinion in Microbiology 38: 137–141.

Caetano DS, O'Meara BC, Beaulieu JM. 2018. Hidden state models improve state-dependent diversification approaches, including biogeographical models. Evolution 72: 2308–2324.

Callahan BJ, McMurdie PJ, Holmes SP. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal 11: 2639–2643.

Chagnon PL. 2016. Seeing networks for what they are in mycorrhizal ecology. Fungal Ecology 24: 148–154.

Chagnon PL, Bradley RL, Klironomos JN. 2012. Using ecological network theory to evaluate the causes and consequences of arbuscular mycorrhizal community structure. New Phytologist 194: 307–312.

Chagnon PL, Bradley RL, Maherali H, Klironomos JN. 2013. A trait-based framework to understand life history of mycorrhizal fungi. Trends in Plant Science 18: 484–491.

Chang Y, Desirò A, Na H, Sandor L, Lipzen A, Clum A, Barry K, Grigoriev I V., Martin FM, Stajich JE, et al. 2019. Phylogenomics of Endogonaceae and evolution of mycorrhizas within Mucoromycota. New Phytologist 222: 511–525.

Chomicki G, Kiers ET, Renner SS. 2020a. The evolution of mutualistic dependence. Annual Review of Ecology, Evolution, and Systematics 51: 409–432.

Chomicki G, Weber M, Antonelli A, Bascompte J, Kiers ET. 2019. The impact of mutualisms on species richness. Trends in Ecology and Evolution 34: 698–711.

Chomicki G, Werner GDA, West SA, Kiers ET. 2020b. Compartmentalization drives the evolution of symbiotic cooperation: Compartmentalisation drives symbiosis. Philosophical Transactions of the Royal Society B: Biological Sciences 375.

Chung H, Pamp SJ, Hill JA, Surana NK, Edelman SM, Troy EB, Reading NC, Villablanca EJ, Wang S, Mora JR, et al. 2012. Gut immune maturation depends on colonization with a host-specific microbiota. Cell 149: 1578–1593.

Clavel J, Escarguel G, Merceron G. 2015. mvMORPH: An R package for fitting multivariate evolutionary models to morphometric data (T Poisot, Ed.). Methods in Ecology and Evolution 6: 1311–1319.

Clay K, Marks S, Cheplick GP. 1993. Effects of insect herbivory and fungal endophyte infection on competitive interactions among grasses. Ecology 74: 1767–1777.

Colman DR, Toolson EC, Takacs-Vesbach CD. 2012. Do diet and taxonomy influence insect gut bacterial communities? Molecular Ecology 21: 5124–5137.

Condamine FL, Rolland J, Morlon H. 2013. Macroevolutionary perspectives to environmental change (H Maherali, Ed.). Ecology Letters 16: 72–85.

Conrath U, Beckers GJM, Flors V, García-Agustín P, Jakab G, Mauch F, Newman M-A, Pieterse CMJ, Poinssot B, Pozo MJ, et al. 2006. Priming: Getting ready for battle. Molecular Plant-Microbe Interactions 19: 1062–1071.

Cosme M, Fernández I, Van der Heijden MGA, Pieterse CMJ. 2018. Non-mycorrhizal plants: The exceptions that prove the rule. Trends in Plant Science 23: 577–587.

Costea PI, Hildebrand F, Manimozhiyan A, Bäckhed F, Blaser MJ, Bushman FD, De Vos WM, Ehrlich SD, Fraser CM, Hattori M, et al. 2017. Enterotypes in the landscape of gut microbial community composition. Nature Microbiology 3: 8–16.

D'Hérelle F. 1917. Sur un microbe invisible antagoniste des bacilles dysentériques. C R Acad Sci Paris: 373–375.

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

Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T, et al. 2015. Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. Science 349: 970–973.

Dawkins R. 1976. The Selfish Gene. New York: Oxford University Press.

Dearnaley JDW, Martos F, Selosse MA. 2012. Orchid mycorrhizas: Molecular ecology, physiology, evolution and conservation aspects. In: Fungal Associations, 2nd Edition. 207–230.

Degnan PH, Pusey AE, Lonsdorf E V, Goodall J, Wroblewski EE, Wilson ML, Rudicell RS, Hahn BH, Ochman H. 2012. Factors associated with the diversification of the gut microbial communities within chimpanzees from Gombe National Park. Proceedings of the National Academy of Sciences of the United States of America 109: 13034–13039.

Delaux P-M, Séjalon-Delmas N, Bécard G, Ané J-M. 2013. Evolution of the plant–microbe symbiotic 'toolkit'. Trends in Plant Science 18: 298–304.

Delsuc F, Metcalf JL, Wegener Parfrey L, Song SJ, González A, Knight R. 2014. Convergence of gut microbiomes in myrmecophagous mammals. Molecular Ecology 23: 1301–1317.

Desirò A, Duckett JG, Pressel S, Villarreal JC, Bidartondo MI. 2013. Fungal symbioses in hornworts: A chequered history. Proceedings of the Royal Society B: Biological Sciences 280: 20130207.

Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, Knight R. 2010. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proceedings of the National Academy of Sciences of the United States of America 107: 11971–11975.

Donald J, Roy M, Suescun U, Iribar A, Manzi S, Péllissier L, Gaucher P, Chave J. 2020. A test of community assembly rules using foliar endophytes from a tropical forest canopy (B Singh, Ed.). Journal of Ecology 108: 1605–1616.

Dorrell RG, Villain A, Perez-Lamarque B, Audren de Kerdrel G, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR, et al. 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. Proceedings of the National Academy of Sciences 118: e2009974118.

Douglas AE. 2008. Conflict, cheats and the persistence of symbioses. New Phytologist 177: 849–858.

Douglas AE, Werren JH. 2016. Holes in the hologenome: Why host-microbe symbioses are not holobionts. mBio 7: 1–7.

Duffy J, Dowling HF. 1979. Fighting infection: Conquests of the twentieth century. Journal of Interdisciplinary History 9: 581.

Duron O, Bouchon D, Boutin S, Bellamy L, Zhou L, Engelstadter J, Hurst GD. 2008. The diversity of reproductive parasites among arthropods: *Wolbachia* do not walk alone. BMC Biology 6: 27.

Engel P, Moran NA. 2013. The gut microbiota of insects - diversity in structure and function. FEMS Microbiology Reviews 37: 699–735.

Erlandson S, Wei X, Savage J, Cavender-Bares J, Peay K. 2018. Soil abiotic variables are more important than Salicaceae phylogeny or habitat specialization in determining soil microbial community structure. Molecular Ecology 27: 2007–2024.

Escherich T. 1886. Die darmbakterien des sauglings und ihre beziehungen zur physiologie der Verdauung. F. Enke.

Everard A, Belzer C, Geurts L, Ouwerkerk JP, Druart C, Bindels LB, Guiot Y, Derrien M, Muccioli GG, Delzenne NM, et al. 2013. Cross-talk between *Akkermansia muciniphila* and intestinal epithelium controls diet-induced obesity. Proceedings of the National Academy of Sciences of the United States of America 110: 9066–71.

Farré-Maduell E, Casals-Pascual C. 2019. The origins of gut microbiome research in Europe: From Escherich to Nissle. Human Microbiome Journal 14: 100065.

Faust K, Raes J. 2012. Microbial interactions: From networks to models. Nature Reviews Microbiology 10: 538–550.

Feijen FA, Vos RA, Nuytinck J, Merckx VSFT. 2018. Evolutionary dynamics of mycorrhizal symbiosis in land plant diversification. Scientific Reports 8: 10698.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17: 368–376.

Felsenstein J. 2004. Inferring phylogenies. Sunderland: Sinauer Associates, Inc.

Ferriere R, Bronstein JL, Rinaldi S, Law R, Gauduchon M. 2002. Cheating and the evolutionary stability of mutualisms. Proceedings of the Royal Society B: Biological Sciences 269: 773–780.

Field KJ, Leake JR, Tille S, Allinson KE, Rimington WR, Bidartondo MI, Beerling DJ, Cameron DD. 2015. From mycoheterotrophy to mutualism: Mycorrhizal specificity and functioning in *Ophioglossum vulgatum* sporophytes. New Phytologist 205: 1492–1502.

Field KJ, Rimington WR, Bidartondo MI, Allinson KE, Beerling DJ, Cameron DD, Duckett JG, Leake JR, Pressel S. 2016. Functional analysis of liverworts in dual symbiosis with Glomeromycota and Mucoromycotina fungi under a simulated Palaeozoic CO_2 decline. ISME Journal 10: 1514–1526.

Fleming A. 1929. On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of *B. influenzae*. 1929. British journal of experimental pathology 10: 226.

Flint HJ, Scott KP, Duncan SH, Louis P, Forano E. 2012. Microbial degradation of complex carbohydrates in the gut. Gut Microbes 3: 289–306.

Fontaine C, Guimarães PR, Kéfi S, Loeuille N, Memmott J, van der Putten WH, van Veen FJFF, Thébault E. 2011. The ecological and evolutionary implications of merging different types of networks. Ecology Letters 14: 1170–1181.

Fortuna MA, Stouffer DB, Olesen JM, Jordano P, Mouillot D, Krasnov BR, Poulin R, Bascompte J. 2010. Nestedness versus modularity in ecological networks: Two sides of the same coin? Journal of Animal Ecology 79: 811–817.

Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. 2017. The evolution of the host microbiome as an ecosystem on a leash. Nature 548: 43–51.

Foster KR, Wenseleers T. 2006. A general model for the evolution of mutualisms. Journal of Evolutionary Biology 19: 1283–1293.

Frank B. 1885. Ueber die auf Wurzelsymbiose beruhende Ernährung gewisser Bäume durch unterirdische Pilze. Berichte der Deutschen Botanischen Gesellschaft 3: 128–145.

Frank B. 1892. Die Ernahrung der Kiefer durch ihre Mykorhiza-Pilze. Ber Dtsch Bot Ges: 577–583.

Franzenburg S, Walter J, Künzel S, Wang J, Baines JF, Bosch TCG, Fraune S. 2013. Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proceedings of the National Academy of Sciences of the United States of America 110.

Frederickson ME. 2017. Mutualisms are not on the verge of breakdown. Trends in Ecology and Evolution 32: 727–734.

Gadkar V, David-schwartz R, Kunik T, Kapulnik Y. 2001. Arbuscular mycorrhizal fungal colonization. Factors involved in host recognition. Plant Physiology 127: 1493–1499.

Gaulke CA, Arnold HK, Humphreys IR, Kembel SW, O'Dwyer JP, Sharpton TJ. 2018. Ecophylogenetics clarifies the evolutionary association between mammals and their gut microbiota (A Martiny and DA Relman, Eds.). mBio 9: 1–14.

Genini J, Morellato LPC, Guimarães PR, Olesen JM. 2010. Cheaters in mutualism networks (I Bartomeus, Ed.). Biology Letters 6: 494–497.

Giraud T, Refrégier G, Le Gac M, de Vienne DM, Hood ME. 2008. Speciation in fungi. Fungal Genetics and Biology 45: 791–802.

Gogarten JF, Rühlemann M, Archie E, Tung J, Akoua-Koffi C, Bang C, Deschner T, Muyembe-Tamfun J-J, Robbins MM, Schubert G, et al. 2021. Primate phageomes are structured by superhost phylogeny and environment. Proceedings of the National Academy of Sciences 118: e2013535118.

Gomes SIF, van Bodegom PM, Merckx VSFT, Soudzilovskaia NA. 2019a. Environmental drivers for cheaters of arbuscular mycorrhizal symbiosis in tropical rainforests. New Phytologist 223: 1575–1583.

Gomes SI, Fortuna MA, Bascompte J, Merckx VSFT. 2019b. Plant cheaters preferentially target arbuscular mycorrhizal fungi that are highly connected to mutualistic plants. bioRxiv: 867259.

Goodrich JK, Davenport ER, Waters JL, Clark AG, Ley RE. 2016. Cross-species comparisons of host genetic associations with the microbiome. Science 352: 532–535.

Gotelli NJ. 2000. Null model analysis of species co-occurrence patterns. Ecology 81: 2606–2621.

Groussin M, Mazel F, Sanders JG, Smillie CS, Lavergne S, Thuiller W, Alm EJ. 2017. Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. Nature Communications 8: 14319.

Guarner F, Malagelada J. 2003. Gut flora in health and disease. The Lancet 361: 1831.

Guimarães PR. 2020. The structure of ecological networks across levels of organization. Annual Review of Ecology, Evolution, and Systematics 51: 433–460.

Guimarães PR, Rico-Gray V, Oliveira PS, Izzo TJ, dos Reis SF, Thompson JN. 2007. Interaction intimacy affects structure and coevolutionary dynamics in mutualistic networks. Current Biology 17: 1797–1803.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S, McHardy AC, Dangl JL, Knight R, Ley R, et al. 2015. Microbiota and host nutrition across plant and animal kingdoms. Cell Host and Microbe 17: 603–616.

Hadfield JD, Krasnov BR, Poulin R, Nakagawa S. 2014. A tale of two phylogenies: Comparative analyses of ecological interactions. The American Naturalist 183: 174–187.

Hallier E. 1869. Die Parasiten der Infectionskrankheiten. Zeitschr. Parasitenkd. Jena 117: 291.

Hammer TJ, Janzen DH, Hallwachs W, Jaffe SP, Fierer N. 2017. Caterpillars lack a resident gut microbiome. Proceedings of the National Academy of Sciences of the United States of America 114: 9641–9646.

Hammer TJ, Sanders JG, Fierer N. 2019. Not all animals need a microbiome. FEMS Microbiology Letters 366: 69–73.

Harmon LJ. 2017. Phylogenetic Comparative Methods.

Harrison XA, McDevitt AD, Dunn JC, Griffiths S, Benvenuto C, Birtles R, Boubli JP, Bown K, Bridson C, Brooks D, et al. 2020. Host-associated fungal communities are determined by host phylogeny and exhibit widespread associations with the bacterial microbiome. bioRxiv: 2020.07.07.177535.

Heather JM, Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. Genomics 107: 1–8.

Hehemann JH, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G. 2010. Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. Nature 464: 908–912.

van der Heijden MGAA, Martin FM, Selosse MA, Sanders IR. 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. New Phytologist 205: 1406–1423.

Heijtz RD, Wang S, Anuar F, Qian Y, Björkholm B, Samuelsson A, Hibberd ML, Forssberg H, Pettersson S. 2011. Normal gut microbiota modulates brain development and behavior. Proceedings of the National Academy of Sciences of the United States of America 108: 3047–3052.

Hembry DH, Yoder JB, Goodman KR. 2014. Coevolution and the diversification of life. American Naturalist 184: 425–438.

Hempel S, Gotzenberger L, Kuhn I, Michalski SG, Rillig MC, Zobel M, Moora M. 2013. Mycorrhizas in the Central European flora: Relationships with plant life history traits and ecology. Ecology 94: 1389–1399.

Henry LM, Peccoud J, Simon JC, Hadfield JD, Maiden MJC, Ferrari J, Godfray HCJ. 2013. Horizontally transmitted symbionts and host colonization of ecological niches. Current Biology 23: 1713–1717.

Hird SM, Sánchez C, Carstens BC, Brumfield RT. 2015. Comparative gut microbiota of 59 neotropical bird species. Frontiers in Microbiology 6: 1403.

Hommola K, Smith JE, Qiu Y, Gilks WR. 2009. A permutation test of host-parasite cospeciation. Molecular Biology and Evolution 26: 1457–1468.

Hooper LV, Littman DR, Macpherson AJ. 2012. Interactions between the microbiota and the immune system. Science 336: 1268–1273.

Hosokawa T, Kikuchi Y, Nikoh N, Shimada M, Fukatsu T. 2006. Strict host-symbiont cospeciation and reductive genome evolution in insect gut bacteria (J Eisen, Ed.). PLoS Biology 4: 1841–1851.

Hosokawa T, Kikuchi Y, Shimada M, Fukatsu T. 2007. Obligate symbiont involved in pest status of host insect. Proceedings of the Royal Society B: Biological Sciences 274: 1979–1984.

Howard RJ. 1996. Cultural control of plant diseases: A historical perspective. Canadian Journal of Plant Pathology 18: 145–150.

Hoysted GA, Kowal J, Jacob A, Rimington WR, Duckett JG, Pressel S, Orchard S, Ryan MH, Field KJ, Bidartondo MI. 2018. A mycorrhizal revolution. Current Opinion in Plant Biology 44: 1–6.

Hu G, Zhang L, Yun Y, Peng Y. 2019. Taking insight into the gut microbiota of three spider species: No characteristic symbiont was found corresponding to the special feeding style of spiders. Ecology and Evolution 9: 8146–8156.

Huelsenbeck JP, Rannala B, Larget B. 2000. A Bayesian framework for the analysis of cospeciation. Evolution 54: 352–364.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. 2016. A new view of the tree of life. Nature Microbiology 1: 16048.

Hugenholtz P. 2002. Exploring prokaryotic diversity in the genomic era. Genome Biology 3: reviews0003.1.

Huxley J. 1942. Evolution. The Modern Synthesis. London: George Alien & Unwin Ltd.

Ives AR, Godfray HCJ. 2006. Phylogenetic analysis of trophic associations. The American Naturalist 168: E1–E14.

Jacquemyn H, Merckx VSFT. 2019. Mycorrhizal symbioses and the evolution of trophic modes in plants (R Shefferson, Ed.). Journal of Ecology 107: 1567–1581.

Jacquemyn H, Merckx VSFT, Brys R, Tyteca D, Cammue BPAA, Honnay O, Lievens B. 2011. Analysis of network architecture reveals phylogenetic constraints on mycorrhizal specificity in the genus *Orchis* (Orchidaceae). New Phytologist 192: 518–528.

Janzen DH. 1980. When is it coevolution? Evolution 34: 611–612.

Johnson NC, Graham JH, Smith FA. 1997. Functioning of mycorrhizal associations along the mutualism-parasitism continuum. New Phytologist 135: 575–586.

Jordano P. 2010. Coevolution in multispecific interactions among free-living species. Evolution: Education and Outreach 3: 40–46.

Jousselin E, Elias M. 2019. Testing host-plant driven speciation in phytophagous insects : a phylogenetic perspective. ArXiv: 1910.09510, Peer-reviewed and recommended by PCI Evolutionary Biology.

Kaplan GG, Ng SC. 2017. Understanding and preventing the global increase of inflammatory bowel disease. Gastroenterology 152: 313-321.e2.

Kelly D, Conway S, Aminov R. 2005. Commensal gut bacteria: Mechanisms of immune modulation. Trends in Immunology 26: 326–333.

Kennedy SR, Tsau S, Gillespie R, Krehenwinkel H. 2020. Are you what you eat? A highly transient and prey-influenced gut microbiome in the grey house spider Badumna longinqua. Molecular Ecology 29: 1001–1015.

Kiers ET, Duhamel M, Beesetty Y, Mensah JA, Franken O, Verbruggen E, Fellbaum CR, Kowalchuk GA, Hart MM, Bago A, et al. 2011. Reciprocal rewards stabilize cooperation in the mycorrhizal symbiosis. Science 333: 880–882.

Kiers ET, van der Heijden MGA. 2006. Mutualistic stability in the arbuscular mycorrhizal symbiosis: Exploring hypotheses of evolutionary cooperation. Ecology 87: 1627–1636.

Kiers ET, Palmer TM, Ives AR, Bruno JF, Bronstein JL. 2010. Mutualisms in a changing world: An evolutionary perspective. Ecology Letters 13: 1459–1474.

Kiers ET, Rousseau RA, West SA, Denlson RF. 2003. Host sanctions and the legume-rhizobium mutualism. Nature 425: 78–81.

Kiers ET, West SA. 2016. Evolution: Welcome to symbiont prison. Current Biology 26: R66–R68.

Kikuchi Y, Hosokawa T, Fukatsu T. 2007. Insect-microbe mutualism without vertical transmission: A stinkbug acquires a beneficial gut symbiont from the environment every generation. Applied and Environmental Microbiology 73: 4308–4316.

Kneip C, Lockhart P, Voß C, Maier UG. 2007. Nitrogen fixation in eukaryotes - New models for symbiosis. BMC Evolutionary Biology 7: 55.

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

Koch R. 1876. Die aetiologie der milzbrand-krankheit, begrundet auf die entwicklungsgeschichte des Bacillus antracis. Beitr. Biol. Pflanz.: 277–310.

Kohl KD. 2020. Ecological and evolutionary mechanisms underlying patterns of phylosymbiosis in host-associated microbial communities. Philosophical Transactions of the Royal Society B: Biological Sciences 375: 20190251.

Kokkoris V, Lekberg Y, Antunes PM, Fahey C, Fordyce JA, Kivlin SN, Hart MM. 2020. Codependency between plant and arbuscular mycorrhizal fungal communities: what is the evidence? New Phytologist 228: 828–838.

Kolaříková Z, Slavíková R, Krüger C, Krüger M, Kohout P. 2021. PacBio sequencing of Glomeromycota rDNA: a novel amplicon covering all widely used ribosomal barcoding regions and its applicability in taxonomy and ecology of arbuscular mycorrhizal fungi. New Phytologist: nph.17372.

Koskella B, Hall LJ, Metcalf CJE. 2017. The microbiome beyond the horizon of ecological and evolutionary theory. Nature Ecology and Evolution 1: 1606–1615.

Kwong WK, Medina LA, Koch H, Sing KW, Soh EJY, Ascher JS, Jaffé R, Moran NA. 2017. Dynamic microbiome evolution in social bees. Science Advances 3: 1–17.

Lambert A, Stadler T. 2013. Birth-death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theoretical Population Biology 90: 113–128.

Lapeyrie F. 1990. The role of ectomycorrhizal fungi in calcareous soil tolerance by 'symbiocalcicole' woody plants. Annales des Sciences Forestières 47: 579–589.

Law R, Dieckmann U. 1998. Symbiosis through exploitation and the merger of lineages in evolution. Proceedings of the Royal Society B: Biological Sciences 265: 1245–1253.

Leake JR, Cameron DD, Beerling DJ. 2008. Fungal fidelity in the myco-heterotroph-to-autotroph life cycle of Lycopodiaceae: A case of parental nurture? New Phytologist 177: 572–576.

Leake J, Johnson D, Donnelly D, Muckle G, Boddy L, Read D. 2004. Networks of power and influence: the role of mycorrhizal mycelium in controlling plant communities and agroecosystem functioning. Canadian Journal of Botany 82: 1016–1045.

Lebeis SL, Paredes SH, Lundberg DS, Breakfield N, Gehring J, McDonald M, Malfatti S, Del Rio TG, Jones CD, Tringe SG, et al. 2015. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. Science 349: 860–864.

Leftwich PT, Clarke NVE, Hutchings MI, Chapman T. 2017. Gut microbiomes and reproductive isolation in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America 114: 12767–12772.

Legendre P, Desdevises Y, Bazin E. 2002. A statistical test for host-parasite coevolution (RDM Page, Ed.). Systematic Biology 51: 217–234.

Legendre P, Legendre L. 2012. Numerical ecology. Elsevier.

Lehnert M, Krug M, Kessler M. 2017. A review of symbiotic fungal endophytes in lycophytes and ferns – a global phylogenetic and ecological perspective. Symbiosis 71: 77–89.

Leopold DR, Busby PE. 2020. Host genotype and colonist arrival order jointly govern plant microbiome composition and function. Current Biology 30: 3260-3266.e5.

Leung JM, Graham AL, Knowles SCL. 2018. Parasite-microbiota interactions with the vertebrate gut: Synthesis through an ecological lens. Frontiers in Microbiology 9: 843.

Lewitus E, Bittner L, Malviya S, Bowler C, Morlon H. 2018. Clade-specific diversification dynamics of marine diatoms since the Jurassic. Nature Ecology and Evolution 2: 1715–1723.

Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al. 2008. Evolution of mammals and their gut microbes. Science 320: 1647–1651.

Li XL, Marschner H, George E. 1991. Acquisition of phosphorus and copper by VA-mycorrhizal hyphae and root-to-shoot transport in white clover. Plant and Soil 136: 49–57.

Lim SJ, Bordenstein SR. 2020. An introduction to phylosymbiosis. Proceedings of the Royal Society B: Biological Sciences 287: 20192900.

Louca S, Jacques SMS, Pires APF, Leal JS, Srivastava DS, Parfrey LW, Farjalla VF, Doebeli M. 2016. High taxonomic variability despite stable functional structure across microbial communities. Nature Ecology & Evolution 1: 0015.

Louca S, Pennell MW. 2020. Extant timetrees are consistent with a myriad of diversification histories. Nature 580: 502–505.

Lozupone C, Knight R. 2005. UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71: 8228–8235.

Maddison WP, Midford PE, Otto SP. 2007. Estimating a binary character's effect on speciation and extinction (T Oakley, Ed.). Systematic Biology 56: 701–710.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2014. Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ 2014: e593.

Maliet O, Hartig F, Morlon H. 2019. A model with many small shifts for estimating species-specific diversification rates. Nature Ecology & Evolution 3: 1086–1092.

Maliet O, Loeuille N, Morlon H. 2020. An individual based model for the eco-evolutionary emergence of bipartite interaction networks (T Poisot, Ed.). Ecology Letters: ele.13592.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research 27: 209–220.

Mao D-P, Zhou Q, Chen C-Y, Quan Z-X. 2012. Coverage evaluation of universal bacterial primers using the metagenomic datasets. BMC Microbiology 12: 66.

Marchesi JR, Ravel J. 2015. The vocabulary of microbiome research: a proposal. Microbiome 3: 1–3.

Margulis L. 1970. Origin of eukaryotic cells. Yale University Press.

Margulis L, Fester R. 1991. Symbiosis as a source of evolutionary innovation: speciation and morphogenesis.

Martos F, Munoz FF, Pailler T, Kottke I, Gonneau C, Selosse MA. 2012. The role of epiphytism in architecture and evolutionary constraint within mycorrhizal networks of tropical orchids. Molecular Ecology 21: 5098–5109.

Mayr E. 1942. Systematics and the Origin of Species. New York: Columbia University Press.

Mazel F, Davis KM, Loudon A, Kwong WK, Groussin M, Parfrey LW. 2018. Is host filtering the main driver of phylosymbiosis across the Tree of Life? (H Bik, Ed.). mSystems 3: 1–15.

McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, Aksenov AA, Behsaz B, Brennan C, Chen Y, et al. 2018. American Gut: an open platform for citizen science microbiome research (CS Greene, Ed.). mSystems 3.

McFall-Ngai MJ. 2002. Unseen forces: The influence of bacteria on animal development. Developmental Biology 242: 1–14.

McFall-Ngai M, Hadfield MG, Bosch TCG, Carey H V., Domazet-Lošo T, Douglas AE, Dubilier N, Eberl G, Fukami T, Gilbert SF, et al. 2013. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences 110: 3229–3236.

McKenzie VJ, Song SJ, Delsuc F, Prest TL, Oliverio AM, Korpita TM, Alexiev A, Amato KR, Metcalf JL, Kowalewski M, et al. 2017. The effects of captivity on the mammalian gut microbiome. In: Integrative and Comparative Biology. Oxford Academic, 690–704.

Meade A, Pagel M. 2007. BayesTraits. Computer program and documentation available at `http://www.evolution.rdg.ac.uk/BayesTraits.html`, 1216-1223.

Merckx VSFT. 2013. Mycoheterotrophy: An Introduction. In: Merckx VSFT, ed. Mycoheterotrophy. New York, NY: Springer New York, 1–17.

Michalska-Smith MJ, Allesina S. 2019. Telling ecological networks apart by their structure: A computational challenge (T Bollenbach, Ed.). PLoS Computational Biology 15: 1–13.

Miyauchi S, Kiss E, Kuo A, Drula E, Kohler A, Sánchez-García M, Morin E, Andreopoulos B, Barry KW, Bonito G, et al. 2020. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. Nature Communications 11: 1–17.

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

Moeller AH, Degnan PH, Pusey AE, Wilson ML, Hahn BH, Ochman H. 2012. Chimpanzees and humans harbour compositionally similar gut enterotypes. Nature Communications 3: 1179.

Moeller AH, Gomes-Neto JC, Mantz S, Kittana H, Segura Munoz RR, Schmaltz RJ, Ramer-Tait AE, Nachman MW. 2019. Experimental evidence for adaptation to species-specific gut microbiota in house mice. mSphere 4: e00387-19.

Moeller AH, Li Y, Ngole EM, Ahuka-Mundeke S, Lonsdorf E V, Pusey AE, Peeters M, Hahn BH, Ochman H. 2014. Rapid changes in the gut microbiome during human evolution. Proceedings of the National Academy of Sciences of the United States of America 111: 16431–16435.

Moeller AH, Peeters M, Ndjango J-BJB, Li Y, Hahn BH, Ochman H. 2013. Sympatric chimpanzees and gorillas harbor convergent gut microbial communities. Genome Research 23: 1715–1720.

Moeller AH, Sanders JG. 2020. Roles of the gut microbiota in the adaptive evolution of mammalian species. Philosophical Transactions of the Royal Society B: Biological Sciences 375: 20190597.

Moeller AH, Suzuki TA, Lin D, Lacey EA, Wasser SK, Nachman MW. 2017. Dispersal limitation promotes the diversification of the mammalian gut microbiota. Proceedings of the National Academy of Sciences of the United States of America 114: 13768–13773.

Moeller AH, Suzuki TA, Phifer-Rixey M, Nachman MW. 2018. Transmission modes of the mammalian gut microbiota. Science 362: 453–457.

Moen D, Morlon H. 2014. Why does diversification slow down? Trends in Ecology and Evolution 29: 190–197.

Molina R, Massicotte H, Trappe JM. 1992. Specificity phenomena in mycorrhizal symbioses: community-ecological consequences and practical implications. In: Allen, Routledge, eds. Mycorrhizal functioning, an integrative plant-fungal process. New York: Chapman and Hall, 357-423.

Montesinos-Navarro A, Segarra-Moragues JG, Valiente-Banuet A, Verdú M. 2012. The network structure of plant-arbuscular mycorrhizal fungi. New Phytologist 194: 536–547.

Moran NA. 2002. The ubiquitous and varied role of infection in the lives of animals and plants. In: American Naturalist. 160: 1–8.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annual Review of Genetics 42: 165–190.

Moran NA, Ochman H, Hammer TJ. 2019. Evolutionary and ecological consequences of gut microbial communities. Annual Review of Ecology, Evolution, and Systematics 50: 451–475.

Moran NA, Sloan DB. 2015. The Hologenome concept: Helpful or hollow? PLoS Biology 13: e1002311.

Morlon H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters 17: 508–525.

Morlon H, Hartig F, Robin S. 2020. Prior hypotheses or regularization allow inference of diversification histories from extant timetrees. bioRxiv: 2020.07.03.185074.

Morlon H, Kemps BD, Plotkin JB, Brisson D. 2012. Explosive radiation of a bacterial species group. Evolution 66: 2577–2586.

Morlon H, Lewitus E, Condamine FL, Manceau M, Clavel J, Drury J. 2016. RPANDA: An R package for macroevolutionary analyses on phylogenetic trees. Methods in Ecology and Evolution 7: 589–597.

Morlon H, Parsons TL, Plotkin JB. 2011. Reconciling molecular phylogenies with the fossil record. Proceedings of the National Academy of Sciences 108: 16327–16332.

Morlon H, Potts MD, Plotkin JB. 2010. Inferring the dynamics of diversification: A coalescent approach (PH Harvey, Ed.). PLoS Biology 8: e1000493.

Morris BEL, Henneberger R, Huber H, Moissl-Eichinger C. 2013. Microbial syntrophy: Interaction for the common good. FEMS Microbiology Reviews 37: 384–406.

Morris JJ, Lenski RE, Zinser ER. 2012. The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. mBio 3: e00036-12.

Muegge BD, Kuczynski J, Knights D, Clemente JC, Fontana L, Henrissat B, Knight R, Gordon JI, González A, Fontana L, et al. 2011. Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. Science 332: 970–974.

Mujica MI, Pérez MF, Jakalski M, Martos F, Selosse MA. 2020. Soil P reduces mycorrhizal colonization while favors fungal pathogens: observational and experimental evidence in *Bipinnula* (Orchidaceae). FEMS Microbiology Ecology 96.

Mushegian AA, Ebert D. 2016. Rethinking "mutualism" in diverse host-symbiont communities. BioEssays 38: 100–108.

Nee S. 2006. Birth-death models in macroevolution. Annual Review of Ecology, Evolution, and Systematics 37: 1–17.

Nguyen NH, Song Z, Bates ST, Branco S, Tedersoo L, Menke J, Schilling JS, Kennedy PG. 2016. FUNGuild: An open annotation tool for parsing fungal community datasets by ecological guild. Fungal Ecology 20: 241–248.

Nishida AH, Ochman H. 2018. Rates of gut microbiome divergence in mammals. Molecular Ecology 27: 1884–1897.

Nishida AH, Ochman H. 2019. A great-ape view of the gut microbiome. Nature Reviews Genetics 20: 195–206.

Noë R, Hammerstein P. 1994. Biological markets: supply and demand determine the effect of partner choice in cooperation, mutualism and mating. Behavioral Ecology and Sociobiology 35: 1–11.

Ochman H, Elwyn S, Moran NA. 1999. Calibrating bacterial evolution. Proceedings of the National Academy of Sciences 96: 12638–12643.

Ochman H, Worobey M, Kuo CH, Ndjango JBN, Peeters M, Hahn BH, Hugenholtz P. 2010. Evolutionary relationships of wild hominids recapitulated by gut microbial communities. PLoS Biology 8: 3–10.

Olesen JM, Bascompte J, Dupont YL, Jordano P. 2007. The modularity of pollination networks. Proceedings of the National Academy of Sciences 104: 19891–19896.

Öpik M, Davison J, Moora M, Zobel M. 2014. DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. Botany 92: 135–147.

Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010. The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). New Phytologist 188: 223–241.

van Opstal EJ, Bordenstein SR. 2019. Phylosymbiosis impacts adaptive traits in *Nasonia* wasps. mBio 10: e00887-19.

Pagel M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. Proceedings of the Royal Society B: Biological Sciences 255: 37–45.

Pasteur L. 1880. Sur les maladies virulentes: et en particulier sur la maladie appelée vulgairement choléra des poules. C. R. Acad. Sci. 90: 249.

Pasteur L. 1885a. Méthode pour prévenir la rage après morsure. Comptes rendus de l'Académie des Sciences: 765–774.

Pasteur L. 1885b. Intérêt des bactéries pour un organisme vivant. C. R. Acad. Sci.

Pellmyr O, Huth CJ. 1994. Evolutionary stability of mutualism between yuccas and yucca moths. Nature 372: 257–260.

Pérez-Cobas AE, Gomez-Valero L, Buchrieser C. 2020. Metagenomic approaches in microbial ecology: An update on whole-genome and marker gene sequencing analyses. Microbial

Genomics 6: 1–22.

Pirozynski KA, Malloch DW. 1975. The origin of land plants: A matter of mycotrophism. BioSystems 6: 153–164.

Plenchette C, Clermont-Dauphin C, Meynard JM, Fortin JA. 2005. Managing arbuscular mycorrhizal fungi in cropping systems. Canadian Journal of Plant Science 85: 31–40.

Poisot T. 2015. When is co-phylogeny evidence of coevolution? Parasite diversity and diversification: Evolutionary ecology meets phylogenetics: 420–433.

Põlme S, Bahram M, Jacquemyn H, Kennedy P, Kohout P, Moora M, Oja J, Öpik M, Pecoraro L, Tedersoo L. 2018. Host preference and network properties in biotrophic plant–fungal associations. New Phytologist 217: 1230–1239.

Pons J, Barraclough TG, Gomez-Zurita J, Cardoso A, Duran DP, Hazell S, Kamoun S, Sumlin WD, Vogler AP. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects (M Hedin, Ed.). Systematic Biology 55: 595–609.

De Queiroz K. 2007. Species concepts and species delimitation. Systematic Biology 56: 879–886.

Queller DC, Strassmann JE. 2018. Evolutionary conflict. Annual Review of Ecology, Evolution, and Systematics 49: 73–93.

Quince C, Curtis TP, Sloan WT. 2008. The rational exploration of microbial diversity. ISME Journal 2: 997–1006.

R Core Team. 2020. R: A language and environment for statistical computing.

Rabosky DL. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees (S-O Kolokotronis, Ed.). PLoS ONE 9: e89543.

Rabosky DL, Lovette IJ. 2008. Density-dependent diversification in North American wood warblers. Proceedings of the Royal Society B: Biological Sciences 275: 2363–2371.

Rafferty NE, Ives AR. 2013. Phylogenetic trait-based analyses of ecological networks. Ecology 94: 2321–2333.

Read DJ. 1991. Mycorrhizas in ecosystems. Experientia 47: 376–391.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rezende EL, Lavabre JE, Guimarães PR, Jordano P, Bascompte J. 2007. Non-random coextinctions in phylogenetically structured mutualistic networks. Nature 448: 925–928.

Rimington WR, Pressel S, Duckett JG, Bidartondo MI. 2015. Fungal associations of basal vascular plants: reopening a closed book? New Phytologist 205: 1394–1398.

Roberts G, Sherratt TN. 1998. Development of cooperative relationships through increasing investment. Nature 394: 175–179.

Rodriguez RJ, White JF, Arnold AE, Redman RS. 2009. Fungal endophytes: Diversity and functional roles: Tansley review. New Phytologist 182: 314–330.

Rook GAW, Lowry CA, Raison CL. 2013. Microbial 'Old Friends', immunoregulation and stress resilience. Evolution, Medicine and Public Health 2013: 46–64.

Roossinck MJ. 2019. Evolutionary and ecological links between plant and fungal viruses. New Phytologist 221: 86–92.

Rosenberg E, Zilber-Rosenberg I. 2018. The hologenome concept of evolution after 10 years. Microbiome 6: 78.

Roy M, Watthana S, Stier A, Richard F, Vessabutr S, Selosse MA. 2009. Two mycoheterotrophic orchids from Thailand tropical dipterocarpacean forests associate with a broad diversity of ectomycorrhizal fungi. BMC Biology 7: 51.

Rudman SM, Greenblum S, Hughes RC, Rajpurohit S, Kiratli O, Lowder DB, Lemmon SG, Petrov DA, Chaston JM, Schmidt P. 2019. Microbiome composition shapes rapid genomic adaptation of *Drosophila melanogaster*. Proceedings of the National Academy of Sciences of the United States of America 116: 20025–20032.

Russell JA, Moreau CS, Goldman-Huertas B, Fujiwara M, Lohman DJ, Pierce NE. 2009. Bacterial gut symbionts are tightly linked with the evolution of herbivory in ants. Proceedings of the National Academy of Sciences of the United States of America 106: 21236–21241.

Russell JA, Sanders JG, Moreau CS. 2017. Hotspots for symbiosis: Function, evolution, and specificity of ant-microbe associations from trunk to tips of the ant phylogeny (Hymenoptera: Formicidae). Myrmecological News 24: 43–69.

Sachs JL, Mueller UG, Wilcox TP, Bull JJ. 2004. The evolution of cooperation. The Quarterly review of biology 51: 211–44.

Sachs JL, Skophammer RG, Regus JU. 2011. Evolutionary transitions in bacterial symbiosis. Proceedings of the National Academy of Sciences of the United States of America 108: 10800–10807.

Sanders JG, Beichman AC, Roman J, Scott JJ, Emerson D, McCarthy JJ, Girguis PR. 2015. Baleen whales host a unique gut microbiome with similarities to both carnivores and herbivores. Nature Communications 6: 8285.

Sanders JG, Powell S, Kronauer DJC, Vasconcelos HL, Frederickson ME, Pierce NE. 2014. Stability and phylogenetic correlation in gut microbiota: lessons from ants and apes. Molecular Ecology 23: 1268–1283.

Saumitou-Laprade P, Cuguen J, Vernet P. 1994. Cytoplasmic male sterility in plants: molecular evidence and the nucleocytoplasmic conflict. Trends in Ecology and Evolution 9: 431–435.

Schardl CL, Craven KD, Speakman S, Stromberg A, Lindstrom A, Yoshida R. 2008. A novel test for host-symbiont codivergence indicates ancient origin of fungal endophytes in grasses (R Page and J Sullivan, Eds.). Systematic Biology 57: 483–498.

Schneider-Maunoury L, Deveau A, Moreno M, Todesco F, Belmondo S, Murat C, Courty PE, Jąkalski M, Selosse MA. 2020. Two ectomycorrhizal truffles, *Tuber melanosporum* and *T. aestivum*, endophytically colonise roots of non-ectomycorrhizal plants in natural environments. New Phytologist 225: 2542–2556.

Ségurel L, Bon C. 2017. On the evolution of lactase persistence in humans. Annual Review of Genomics and Human Genetics 18: 297–319.

Selosse MA. 2000. La symbiose: structures et fonctions, rôle écologique et évolutif. Vuibert.

Selosse MA, Baudoin E, Vandenkoornhuyse P. 2004. Symbiotic microorganisms, a key for ecological success and protection of plants. Comptes Rendus - Biologies 327: 639–648.

Selosse MA, Bessis A, Pozo MJ. 2014. Microbial priming of plant and animal immunity: symbionts as developmental signals. Trends in Microbiology 22: 607–613.

Selosse MA, Dubois MP, Alvarez N. 2009. Do Sebacinales commonly associate with plant roots as endophytes? Mycological Research 113: 1062–1069.

Selosse MA, Richard F, He X, Simard SW. 2006. Mycorrhizal networks: des liaisons dangereuses? Trends in Ecology and Evolution 21: 621–628.

Selosse MA, Rousset F. 2011. The plant-fungal marketplace. Science 333: 828–829.

Selosse MA, Roy M. 2009. Green plants that feed on fungi: facts and questions about mixotrophy. Trends in Plant Science 14: 64–70.

Selosse MA, Schneider-Maunoury L, Martos F. 2018. Time to re-think fungal ecology? Fungal ecological niches are often prejudged. New Phytologist 217: 968–972.

Selosse MA, Le Tacon F. 1998. The land flora: a phototroph-fungus partnership? Trends in Ecology & Evolution 13: 15–20.

Sender R, Fuchs S, Milo R. 2016. Are we really vastly outnumbered? Revisiting the ratio of bacterial to host cells in humans. Cell 164: 337–340.

Séne S, Selosse MA, Forget M, Lambourdière J, Cissé K, Diédhiou AG, Rivera-Ocasio E, Kodja H, Kameyama N, Nara K, et al. 2018. A pantropically introduced tree is followed by specific ectomycorrhizal symbionts due to pseudo-vertical transmission. ISME Journal 12: 1806–1816.

Sepp SK, Davison J, Jairus T, Vasar M, Moora M, Zobel M, Öpik M. 2019. Non-random association patterns in a plant–mycorrhizal fungal network reveal host–symbiont specificity. Molecular Ecology 28: 365–378.

Shade A, Jacques MA, Barret M. 2017. Ecological patterns of seed microbiome diversity, transmission, and assembly. Current Opinion in Microbiology 37: 15–22.

Shapira M. 2016. Gut microbiotas and host evolution: scaling up symbiosis. Trends in Ecology and Evolution 31: 539–549.

Sharon G, Segal D, Ringo JM, Hefetz A, Zilber-Rosenberg I, Rosenberg E. 2010. Commensal bacteria play a role in mating preference of *Drosophila melanogaster*. Proceedings of the National Academy of Sciences 107: 20051–20056.

Shaw G, Leake JR, Baker AJM, Read DJ. 1990. The biology of mycorrhiza in the Ericaceae: XVII. The role of mycorrhizal infection in the regulation of iron uptake by ericaceous plants. New Phytologist 115: 251–258.

Simard SW, Perry DA, Jones MD, Myrold DD, Durall DM, Molina R. 1997. Net transfer of carbon between ectomycorrhizal tree species in the field. Nature 388: 579–582.

Simmons BI, Cirtwill AR, Baker NJ, Wauchope HS, Dicks L V., Stouffer DB, Sutherland WJ. 2019. Motifs in bipartite ecological networks: uncovering indirect interactions. Oikos 128: 154–170.

Simon JC, Marchesi JR, Mougel C, Selosse MA. 2019. Host-microbiota interactions: From holobiont theory to analysis. Microbiome 7: 1–5.

Smith KA. 2012. Louis Pasteur, the father of immunology? Frontiers in Immunology 3: 68.

Smith K, McCoy KD, Macpherson AJ. 2007. Use of axenic animals in studying the adaptation of mammals to their commensal intestinal microbiota. Seminars in Immunology 19: 59–69.

Smith SE, Read DJ. 2008. Mycorrhizal Symbiosis. Elsevier.

Solís-Lemus C, Yang M, Ané C. 2016. Inconsistency of species tree methods under gene flow. Systematic Biology 65: 843–851.

Song SJ, Sanders JG, Delsuc F, Metcalf J, Amato K, Taylor MW, Mazel F, Lutz HL, Winker K, Graves GR, et al. 2020. Comparative analyses of vertebrate gut microbiomes reveal convergence between birds and bats. mBio 11: 1–14.

Sørensen MES, Lowe CD, Minter EJA, Wood AJ, Cameron DD, Brockhurst MA. 2019. The role of exploitation in the establishment of mutualistic microbial symbioses. FEMS Microbiology Letters 366.

Stadler T. 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proceedings of the National Academy of Sciences 108: 6187–6192.

Stefani F, Bencherif K, Sabourin S, Hadj-Sahraoui AL, Banchini C, Séguin S, Dalpé Y. 2020. Taxonomic assignment of arbuscular mycorrhizal fungi in an 18S metagenomic dataset: a case study with saltcedar (*Tamarix aphylla*). Mycorrhiza 30: 243–255.

Strullu-Derrien C, Selosse MA, Kenrick P, Martin FM. 2018. The origin and evolution of mycorrhizal symbioses: from palaeomycology to phylogenomics. New Phytologist 220: 1012–1030.

Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences of the United States of America 114: 1607–1611.

Suzuki TA, Ley RE. 2020. The role of the microbiota in human genetic adaptation. Science 370: eaaz6827.

Suzuki TA, Worobey M. 2014. Geographical variation of human gut microbial composition. Biology Letters 10: 20131037.

Szöllősi GJ, Rosikiewicz W, Boussau B, Tannier E, Daubin V. 2013a. Efficient exploration of the space of reconciled gene trees. Systematic Biology 62: 901–912.

Szöllősi GJ, Tannier E, Daubin V, Boussau B. 2015. The inference of gene trees with species trees. Systematic Biology 64: e42–e62.

Szöllősi GJ, Tannier E, Lartillot N, Daubin V. 2013b. Lateral gene transfer from the dead. Systematic Biology 62: 386–397.

Taylor DL, Hollingsworth TN, McFarland JW, Lennon NJ, Nusbaum C, Ruess RW. 2014. A first comprehensive census of fungi in soil reveals both hyperdiversity and fine-scale niche partitioning. Ecological Monographs 84: 3–20.

Tedersoo L, Mett M, Ishida TA, Bahram M. 2013. Phylogenetic relationships among host plants explain differences in fungal species richness and community composition in ectomycorrhizal symbiosis. New Phytologist 199: 822–831.

Thébault E, Fontaine C. 2008. Does asymmetric specialization differ between mutualistic and trophic networks? Oikos 0: 080227085440234–0.

Thébault E, Fontaine C. 2010. Stability of ecological communities and the architecture of mutualistic and trophic networks. Science 329: 853–856.

Theis KR, Dheilly NM, Klassen JL, Brucker RM, Baines JF, Bosch TCG, Cryan JF, Gilbert SF, Goodnight CJ, Lloyd EA, et al. 2016. Getting the hologenome concept right: an eco-evolutionary framework for hosts and their microbiomes (JA Gilbert, Ed.). mSystems 1: 1–6.

Thompson JN. 2005. The geographic mosaic of coevolution. University of Chicago Press.

Thomson BD, Robson AD, Abbott LK. 1986. Effect of phosphorus in the formation of mycorrhizas by *Gigaspora calospora* and *Glomus fasciculatum* in relation to root carbohydrates. New Phytologist 103: 751–765.

Tissier H. 1905. Etude d'une varieté d'infection intestinale chez le nourrisson. Annales de l'Institut Pasteur 19ème année: 273–316.

Toju H, Guimarães PR, Olesen JM, Thompson JN. 2014. Assembly of complex plant-fungus networks. Nature Communications 5: 1–7.

Toju H, Tanabe AS, Ishii HS. 2016. Ericaceous plant-fungus network in a harsh alpine-subalpine environment. Molecular Ecology 25: 3242–3257.

Truyens S, Weyens N, Cuypers A, Vangronsveld J. 2015. Bacterial seed endophytes: genera, vertical transmission and interaction with plants. Environmental Microbiology Reports 7: 40–50.

Ubeda C, Djukovic A, Isaac S. 2017. Roles of the intestinal microbiota in pathogen protection. Clinical & Translational Immunology 6: e128.

Van Valen L. 1973. A new evolutionary law. Evolutionary Theory 1: 1–30.

Vandenkoornhuyse P, Ridgway KP, Watson IJ, Fitter AH, Young JPW. 2003. Co-existing grass species have distinctive arbuscular mycorrhizal communities. Molecular Ecology 12: 3085–3095.

Vannier N, Mony C, Bittebiere A-K, Michon-Coudouel S, Biget M, Vandenkoornhuyse P. 2018. A microorganisms' journey between plant generations. Microbiome 6: 79.

Vavre F, Kremer N. 2014. Microbial impacts on insect evolutionary diversification: From patterns to mechanisms. Current Opinion in Insect Science 4: 29–34.

Veldre V, Abarenkov K, Bahram M, Martos F, Selosse MA, Tamm H, Kõljalg U, Tedersoo L. 2013. Evolution of nutritional modes of Ceratobasidiaceae (Cantharellales, Basidiomycota) as revealed from publicly available ITS sequences. Fungal Ecology 6: 256–268.

de Vienne DM, Giraud T, Shykoff JA. 2007. When can host shifts produce congruent host and parasite phylogenies? A simulation approach. Journal of Evolutionary Biology 20: 1428–1438.

de Vienne DM, Refrégier G, López-Villavicencio M, Tellier A, Hood ME, Giraud T. 2013. Cospeciation vs host-shift speciation: Methods for testing, evidence from natural associations and relation to coevolution. New Phytologist 198: 347–385.

van Vliet S, Doebeli M. 2019. The role of multilevel selection in host microbiome evolution. Proceedings of the National Academy of Sciences of the United States of America 116: 20591–20597.

Weeks AR, Turelli M, Harcombe WR, Reynolds KT, Hoffmann AA. 2007. From parasite to mutualist: Rapid evolution of *Wolbachia* in natural populations of Drosophila. PLoS Biology 5: 0997–1005.

Wehner J, Powell JR, Muller LAH, Caruso T, Veresoglou SD, Hempel S, Rillig MC. 2014. Determinants of root-associated fungal communities within Asteraceae in a semi-arid grassland (M van der Heijden, Ed.). Journal of Ecology 102: 425–436.

Weiß M, Waller F, Zuccaro A, Selosse MA. 2016. Sebacinales - one thousand and one interactions with land plants. New Phytologist 211: 20–40.

Werner GDA, Cornelissen JHC, Cornwell WK, Soudzilovskaia NA, Kattge J, West SA, Kiers ET. 2018. Symbiont switching and alternative resource acquisition strategies drive mutualism breakdown. Proceedings of the National Academy of Sciences 115: 5229–5234.

Werner GDA, Kiers ET. 2015. Order of arrival structures arbuscular mycorrhizal colonization of plants. New Phytologist 205: 1515–1524.

Werner GDA, Strassmann JE, Ivens ABF, Engelmoer DJP, Verbruggen E, Queller DC, Noë R, Johnson NC, Hammerstein P, Kiers ET. 2014. Evolution of microbial markets. Proceedings of the National Academy of Sciences of the United States of America 111: 1237–1244.

Weyl EG, Frederickson ME, Yu DW, Pierce NE. 2010. Economic contract theory tests models of mutualism. Proceedings of the National Academy of Sciences of the United States of America 107: 15712–15716.

Wilson D. 1995. Endophyte: The evolution of a term, and clarification of its use and definition. Oikos 73: 274.

Wilson AW, Hosaka K, Mueller GM. 2017. Evolution of ectomycorrhizas as a driver of diversification and biogeographic patterns in the model mycorrhizal mushroom genus Laccaria. New Phytologist 213: 1862–1873.

Winther JL, Friedman WE. 2008. Arbuscular mycorrhizal associations in Lycopodiaceae. New Phytologist 177: 790–801.

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, et al. 2006. Sex and virulence in *Escherichia coli*: An evolutionary perspective. Molecular Microbiology 60: 1136–1151.

Wu S, Sun C, Li Y, Wang T, Jia L, Lai S, Yang Y, Luo P, Dai D, Yang YQ, et al. 2020. GMrepo: A database of curated and consistently annotated human gut metagenomes. Nucleic Acids Research 48: D545–D553.

Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP, et al. 2012. Human gut microbiome viewed across age and geography. Nature 486: 222–227.

Yeoh YK, Dennis PG, Paungfoo-Lonhienne C, Weber L, Brackin R, Ragan MA, Schmidt S, Hugenholtz P. 2017. Evolutionary conservation of a core root microbiome across plant phyla along a tropical soil chronosequence. Nature Communications 8: 215.

Young JPW, Haukka KE. 1996. Diversity and phylogeny of rhizobia. New Phytologist 133: 87–94.

Youngblut ND, Reischer GH, Dauser S, Walzer C, Stalder G, Farnleitner AH, Ley RE. 2020. Strong influence of vertebrate host phylogeny on gut archaeal diversity. bioRxiv: 2020.11.10.376293.

Youngblut ND, Reischer GH, Walters W, Schuster N, Walzer C, Stalder G, Ley RE, Farnleitner AH. 2019. Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. Nature Communications 10: 1–15.

Yun JH, Roh SW, Whon TW, Jung MJ, Kim MS, Park DS, Yoon C, Nam Y Do, Kim YJ, Choi JH, 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 C, Derrien M, Levenez F, Brazeilles R, Ballal SA, Kim J, Degivry MC, Quéré G, Garault P, Van Hylckama Vlieg JET, et al. 2016. Ecological robustness of the gut microbiota in response to ingestion of transient food-borne microbes. ISME Journal 10: 2235–2245.

Zhao Z, Li X, Liu MF, Merckx VSFT, Saunders RMK, Zhang D. 2021. Specificity of assemblage, not fungal partner species, explains mycorrhizal partnerships of mycoheterotrophic *Burmannia* plants. ISME Journal: 1–14.

Zilber-Rosenberg I, Rosenberg E. 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiology Reviews 32: 723–735.

RÉSUMÉ

De nombreuses études récentes ont permis de caractériser la composition des communautés microbiennes, appelées microbiotes, hébergées par plantes et animaux. Le but de ma thèse est de faire avancer notre compréhension de l'évolution des microbiotes associés aux espèces hôtes animales ou végétales, en utilisant comme données les arbres phylogénétiques des hôtes et des séquences d'ADN microbien de *metabarcoding* caractérisant leurs microbiotes. Pour cela, nous avons développé de nouvelles méthodes quantitatives, collecté des données ainsi que réalisé une série d'analyses. Nous avons considéré à la fois les microbes des animaux et des plantes, et tout particulièrement les interactions mycorhizennes. Dans le chapitre I, nous nous sommes intéressés à l'évolution du microbiote au cours de la diversification d'un clade d'hôtes. Nous avons développé une approche quantitative afin d'inférer les microbes transmis. À partir de la phylogénie des hôtes et de séquences d'ADN de leurs microbiotes, regroupées en unité taxonomique opérationnelle (OTU), notre approche utilise la variation nucléotidique au sein des OTUs pour détecter ceux qui sont transmis lors de la diversification des hôtes. Appliquée aux microbiotes de primates et araignées, nous avons trouvé que >5% des bactéries intestinales des primates étaient transmises verticalement, tandis qu'il n'y a vraisemblablement pas de transmission chez les araignées, confirmant l'hétérogénéité de l'évolution des interactions hôtes-microbes chez les animaux. Enfin, nous avons comparé les performances de notre modèle à celles d'autres approches existantes et montré que notre modèle est moins enclin aux faux-positifs lorsque la variation nucléotidique intra-OTU est faible. Dans le chapitre II, nous avons examiné les liens entre les histoires évolutives des hôtes et de leurs microbes associés. Nous avons plus particulièrement cherché à répondre à deux questions : « Dans quelle mesure les patrons d'interactions hôtes-microbiotes sont influencés par leurs histoires évolutives ? » et « comment l'histoire évolutive des hôtes influence-t-elle la diversification de leurs microbes associés ? ». La première question nous a amené à comparer les méthodes disponibles pour estimer le signal phylogénétique dans les réseaux d'interactions, afin de déterminer par exemple si des espèces de plantes proches ont tendance à interagir avec les mêmes champignons mycorhiziens. Nous avons trouvé qu'une approche fréquemment utilisée génère beaucoup de faux-positifs et qu'à l'inverse, les tests de Mantel donnent des résultats assez satisfaisants. Nous avons enfin exploré la seconde question en évaluant comment les plantes ont pu affecter la diversification des champignons endomycorhiziens (Glomeromycotina). Nos analyses suggèrent que ces symbiontes obligatoires ont récemment subit un ralentissement de leur diversification, qui peut être lié à l'évolution, chez de nombreuses plantes, de stratégies alternatives à l'endomycorhize. Dans le chapitre III, nous nous sommes focalisés sur l'évolution de la tricherie dans le mutualisme hôte-microbiote, et plus particulièrement dans la symbiose mycorhizienne. Nous avons exploré les contraintes limitant l'émergence de la tricherie chez les plantes (mycohétérotrophie) en analysant les patrons d'interactions endomycorhizennes à l'échelle mondiale. Nous avons ensuite étudié si des contraintes similaires s'appliquaient dans les communautés locales où vivent des plantes initialement mycohétérotrophes (les lycopodes), échantillonnées sur l'île de la Réunion. Nous en avons déduit qu'il existe généralement de fortes contraintes limitant la tricherie dans cette symbiose, mais que ces contraintes peuvent être relâchées au sein des communautés où la tricherie a lieu. Ainsi, ma thèse illustre que l'utilisation de méthodes quantitatives combinées à des données de *metabarcoding*, malgré leurs limites respectives, permet de mieux caractériser l'évolution des interactions hôtes-microbiotes.

MOTS CLÉS : microbiotes, symbiose, cophylogénie, coévolution, diversification, réseau mycorhizien

SUMMARY

A plethora of recent studies have characterized the composition and functional role of microbial communities hosted by animals and plants, called microbiota. The overall goal of my PhD is to advance our understanding of how microbiota evolve with their host species, using data comprised of the phylogenetic relationships between host species and metabarcoding microbial sequences characterizing their microbial communities. We developed new quantitative tools, collected data, and performed a series of analyses, all directed to this common overarching goal. We considered both microbiota-animal and microbiota-plant systems, with a specific focus on mycorrhizal interactions. In Chapter I, we study the evolution of the microbiota during the diversification of host clades. We develop a quantitative approach for inferring the modes of microbial inheritance as host clades diversify. Given a host phylogeny and the microbiota of present-day species, each characterized by a list of short DNA sequences clustered into operational taxonomic units (OTUs), our approach uses nucleotide variability within OTUs to detect OTUs that are vertically transmitted. We apply this approach to two distinct systems, the gut microbiota of primates and a clade of Hawaiian spiders. We find that >5% of bacteria in primate guts are vertically transmitted, whereas there is no evidence of vertical transmission in spiders, confirming that host-microbiota evolutionary dynamics are highly heterogeneous across the animal kingdom. Finally, we compare the performances of our model to other available approaches and find that it is less prone to false-positives when the nucleotide variability within OTUs is low. In Chapter II, we examine the interplay between the evolutionary history of host and host-associated microbial clades. We focus on two specific questions: "To what extent does evolutionary history influence which microbial species interact with which host species?" and "How does the evolutionary history of hosts influence the diversification of host-associated microbial clades?". The first question leads us to compare different methods for estimating phylogenetic signals in host-microbiota interactions, i.e. whether closely related species share similar sets of partners, with an application on plant-mycorrhizal interactions. We find that one of the most widely used approaches often detects phylogenetic signals when it should not and that Mantel tests perform best. We explore the second question by studying the diversification of the arbuscular mycorrhizal fungi (Glomeromycotina) in the past 500 million years and evaluating how land plants might have affected the diversification of these obligate mycorrhizal symbionts. Our analyses support that these fungi have experienced a recent diversification slowdown that might be linked to the shrinkage of their mycorrhizal niches as plant lineages evolve alternative symbiotic strategies. In Chapter III, we focus on the evolution of cheating in host-microbiota mutualisms, by taking the mycorrhizal symbiosis as a case study. We explore constraints on the evolutionary emergence of cheating in plants (mycoheterotrophy) by analyzing the patterns of plant-mycorrhizal fungus interactions at the global scale. Next, we investigate whether similar constraints are found in local mycorrhizal networks including initially mycoheterotrophic plants (Lycopodiaceae) that we have sampled in La Réunion island. We conclude that there are overall strong constraints limiting the emergence of cheaters in the mycorrhizal symbiosis, but these constraints might be relaxed in the local communities where cheating occurs. Overall, my thesis illustrates how new or recent computational tools, in combination with metabarcoding sequencing data, allow studying how microbiota evolve with their hosts. We discuss the challenges and promise of this comparative approach to host-microbiota evolution.

KEYWORDS : host-associated microbiota, symbiosis, co-phylogeny, co-evolution, diversification, mycorrhizal network