From models to data : understanding biodiversity patterns from environmental DNA data

Guilhem Sommeria-Klein

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Cotutelle internationale avec :

Présentée et soutenue par :
Guilhem Sommeria-Klein
Le jeudi 14 septembre 2017

Titre :

From models to data:
understanding biodiversity patterns from environmental DNA data

---

École doctorale et discipline ou spécialité :
ED SEVAB : Écologie, biodiversité et évolution

Unité de recherche :
Laboratoire Evolution et Diversité Biologique (EDB)

Directeur(s) de Thèse :

Jérôme Chave
Hélène Morlon

Rapporteurs :
Christopher Quince
Corinne Vacher

Autre(s) membre(s) du jury :

Sébastien Brosse
Francesco Ficetola

Acknowledgements
Remerciements

Je souhaiterais ici chaleureusement remercier toutes les personnes qui ont contribué au bon déroulement de cette thèse de façon plus ou moins directe, par leur aide, leurs conseils ou leur soutien.

Merci tout d'abord à Jérôme Chave, mon directeur de thèse à Toulouse, pour m'avoir ouvert les portes de l'écologie, une discipline dont j'ignorais tout ou presque au début de ma thèse. Ayant été jusqu'alors principalement confronté à la recherche théorique, j'ai découvert un univers où les données, leur production, leur analyse et leur interprétation, sont le nerf de la guerre. Mon adaptation ne s'est pas faite en un jour, et je mesure aujourd'hui le chemin parcouru. Je te remercie Jérôme de tes efforts pour m'intégrer à la discipline en début de thèse, notamment en me permettant de partir deux fois faire du terrain aux Nouragues, une expérience inoubliable. Merci pour tes conseils toujours avisés et tes relectures attentives. Ton exigence, ton enthousiasme et ton énergie inépuisables, ton recul sur la discipline et ta large culture scientifique m'ont inspiré tout au long de ma thèse.

Merci à Hélène Morlon, ma directrice de thèse à Paris, de m'avoir accueilli pour ma dernière année de thèse. Outre que cette année supplémentaire m'a été précieuse pour terminer ma thèse dans de bonnes conditions, cette immersion dans la macro-évolution a été pour moi l'occasion de découvrir une autre façon d'aborder la recherche en écologie et évolution. Merci Hélène pour tes conseils, ton soutien et ta patience. Et merci de m'avoir fait profiter de cette atmosphère unique, où dynamisme, rigueur et efficacité riment si bien avec convivialité, bienveillance et décontraction. Nous n'avons pas eu l'occasion de travailler ensemble autant que je l'aurais souhaité au cours de cette thèse, et je me réjouis par conséquent de pouvoir continuer à le faire dans le futur.

I am very grateful to Christopher Quince and Corinne Vacher for taking the time to review this thesis. Thank you very much for your careful reading and your constructive criticism. I also thank Sébastien Brosse and Francesco Ficetola for agreeing to be part of the jury.

Merci à mes co-auteurs et collaborateurs, sur le travail desquels une grande partie de cette thèse est basée.

Merci en particulier à Lucie Zinger, dont le travail d'analyse et d'interprétation des données utilisées dans cette thèse a été crucial. Lucie, tu as été un précieux pont avec la biologie tout au long de cette thèse pour le physicien de formation que je suis, et je te suis infiniment reconnaissant pour tes nombreuses explications et conseils. Merci à Pierre Taberlet et Eric Coissac de m'avoir fait bénéficier de leur expertise state-of-the-art dans la production et l'analyse des données metabarcoding. Merci à Heidy Schimann pour son aide dans l'interprétation biologique des données, ainsi que pour son aide

logistique sur le terrain. Merci à Amaia Iribar pour son aide indispensable dans la préparation du terrain, sa contribution à la production des données, ainsi que pour sa remarquable efficacité pour surmonter les difficultés technique, logistique et administrative de tout ordre, le tout dans une bonne humeur inaltérable. Merci à Sophie Manzi et Eliane Louisanna pour leur aide sur le terrain et leur travail de wetlab. Merci à Vincent Schilling pour sa contribution aux analyses bioinformatiques, son aide sur le terrain, et pour son humour en tant que camarade de carbet aux Nouragues. Merci également à Saint Omer Cazal et Audrey Sagne pour leur aide sur le terrain à Paracou et Arbocel, et à Daniel Boutaud pour ses très utiles porte-piochons tout-terrain. Merci enfin à Elodie Courtois et Blaise Tymen de m'avoir fait découvrir la Guyane et les Nouragues en début de thèse, et pour ces moments partagés sur le terrain.

Je remercie en outre Antoine Fouquet et Jean-Pierre Vacher de m'avoir fait confiance pour l'analyse de leurs données. Merci également à Hélène Holota pour son efficacité, sa disponibilité et sa gentillesse, à Blaise Tymen pour son aide avec les données Lidar, et à Mélanie Roy, Antoine Fouquet, Gaël Grenouillet et Lounès Chikhi, entre autres, pour d'enrichissantes discussions et pour leurs conseils.

Je voudrais ensuite remercier un certain nombre de personnes qui, si elles n'ont pas directement contribué au contenu de cette thèse, ont éclairé mes journées de travail à Toulouse et Paris et ont fait le sel de ces quatre années de vie.

Merci à mes co-bureaux toulousains Jessica et Félix pour toutes ces longues discussions, et pour avoir supporté mon humour avec bienveillance en toute circonstance. Merci à mes quatre « camarades de promo » à EDB, Arthur, Paul, Isabelle et Jean-Pierre, avec qui cela a été un immense plaisir de partager ces trois années à Toulouse. Merci à Boris, Blaise, Mathieu, Olivia, Léa, Josselin, Aurèle, Luc, Nico, Camille, Céline, Marine, Lucie, Alice, Sébastien, Jade, Kévin, Jan, Fabian, Isabel … pour tous ces moments partagés, et à tous les membres d'EDB jeunes et moins jeunes pour l'ambiance remarquable du labo.

Merci à Simon et aux locataires successifs de l'inénarrable maison de la culture François Magendie, à Louise et ses plongées dans le monde du théâtre, à Etienne et Mathilde et leurs « écoles d'été » hippies, à Guillem, Claire, Hélène, Lucie, Lisa-Lou et aux autres, pour avoir brillamment peuplé ces années toulousaines. Merci à Jean-Pierre et Sébastien de m'avoir initié à l'herpéto. Merci à Alex pour son accueil à Montpellier et ces débats passionnés sur la science et l'écologie. Merci aux Américains : Marc et Léo et leur inspirant « Silicon Valley spirit », et Matthieu, fidèle birding buddy. Et merci à Simon et Florian pour leurs – trop rares – immersions dans l'informatique quantique et les réseaux d'énergie.

Un grand merci également à mes co-bureaux parisiens. Marc, Odile, Julien, Eric, Leandro, Olivier, votre accueil chaleureux dans et hors du labo a grandement facilité ma transition parisienne.

Merci enfin aux anciens, aux vieux de la vieille, Grenoblois d'ici et d'ailleurs : Thibault, Mathieu, Arantxa, David, Lucas, Aurore, Carl, Vio … Je vous compte parmi les amis, mais c'est déjà presque la famille.

Je remercie pour finir, last but not least, mes parents et mon frère, pour leur soutien sans faille et ô combien indispensable.

Table of Contents

Introduction

I. What drives the assembly of ecological communities?

Science consists in finding patterns in a collection of isolated observations so as to gain understanding of the processes that generated them. Natural sciences began with attempts at classifying the diversity of the living organisms into categories (Aristotle, IVth cent. BC), and this classification has been developed and perfected over the centuries into the modern binomial nomenclature (Linnaeus, 1753). But classification efforts were not limited to the description of species. Associations of species, and in particular plant associations, were named using the same model, and were carefully described based on their taxonomic composition and the abiotic properties of their environment (Braun-Blanquet & Pavillard, 1922). Even though forest plant associations were observed shifting through time, this phenomenon was described as mirroring the life cycle of individual organisms, from 'youth' to 'senescence' (Clements, 1916). The underlying idea was that the organization of the living world obeyed static and deterministic rules, which were to be uncovered. This idea was encouraged by the discovery of the elegant laws that govern physics and chemistry.

By contrast, early discoveries on evolution and biogeography (Darwin, 1859; Wallace, 1876) brought the idea that chance and history have played an overwhelming role in shaping the modern living world. Gleason (1926) and Tansley (1935) were the first to contend that the diversity of plant associations was not well described by discrete vegetation types, and that species associations were rather the transient outcome of random dispersal events, constrained by abiotic conditions and species interactions. Later, Hutchinson (1961), MacArthur (1972), Diamond (1975), Hubbell (1979), Ricklefs (1987), and Brown (1995), among others, have successively elaborated on this idea, laying the foundations of modern community ecology. The term 'community' refers to all the organisms coexisting in a given location and at a given time. It may also refer to a taxonomic subgroup of these organisms, such as a 'plant community'.

The question of the relative role played by deterministic and stochastic processes in shaping ecological communities remains central to ecology. In this section, I first argue that addressing this question is key to our ability to preserve natural ecosystems and to predict their response to human perturbations. I then briefly review the mechanisms of community assembly that have been proposed.

1. Motivations

The increasing awareness of the threats posed to natural ecosystems by human activities has added a sense of urgency to the study of ecological processes. Indeed, the fate of the Earth's biodiversity, and beyond it, of the ecosystems on which human societies rely for food, water, clean air, health, and raw materials, has become a major source of concern (Daily, 1997). As a consequence, theoretical advances in ecology can no longer be considered in isolation from their practical implications. In particular, many predictions relevant to policy-making strongly depend on assumptions regarding the mechanisms of community assembly. Thus, data-driven understanding of community assembly is critical to well-informed policy-making. Three examples are given below: the prediction of ecosystem stability and state shifts in response to human perturbations, the prediction of the impact of climate change, and the conservation of biodiversity.

Measuring ecosystem stability to perturbations is a subject of active research, as is the relationship between biodiversity and ecosystem stability (McCann, 2000; Tilman *et al.*, 2006; Loreau & de Mazancourt, 2013). In this context, natural ecosystems are commonly represented as stable communities held together by species interactions, in part because this representation lends itself well to theoretical approaches (Arnoldi *et al.*, 2016). Drawing on this framework, it has been hypothesized that the response of ecosystems to perturbations may bear a similarity with that of physical systems exhibiting critical phase transitions (cf. Fig. 1; Scheffer *et al.*, 2012). Accordingly, 'tipping points', sudden and difficult-to-reverse shifts in a system's state in response to

perturbation, should be expected (Brook *et al.*, 2013). Moreover, such state shifts could be possibly predicted in advance through the identification of early-warning signals (Carpenter *et al.*, 2011; Scheffer *et al.*, 2012). While this type of non-linear behaviour has been evidenced in lake ecosystems (Carpenter *et al.*, 2011), it remains difficult to study empirically, and knowledge of community assembly processes is key to provide realistic assumptions for the theoretical prediction of possible tipping points.

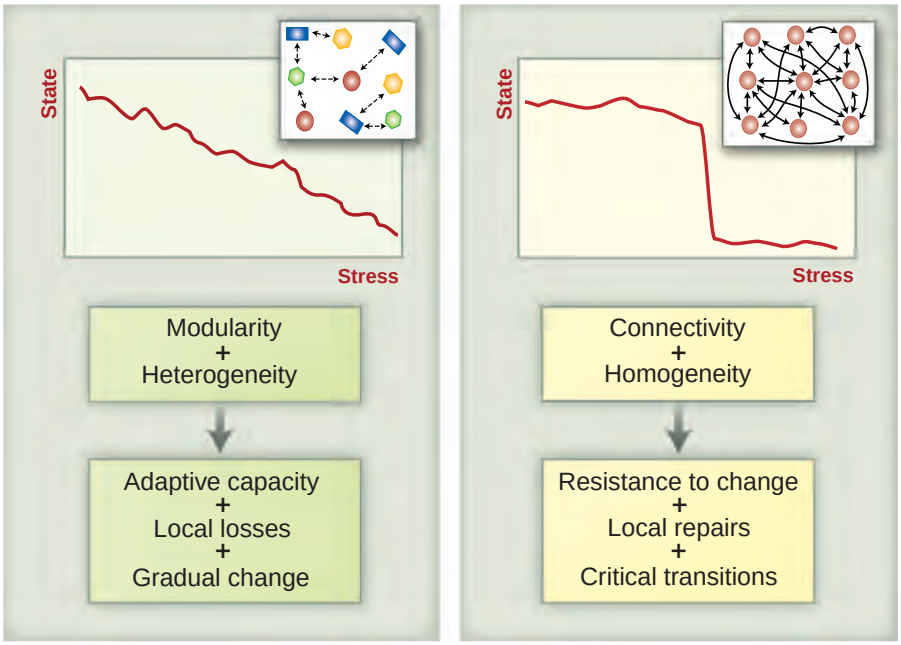

Figure 1. The response of ecosystems to human-induced stress is commonly studied using a network representation of ecological communities, envisioned as stable entities held together by interactions. Depending on network connectivity and modularity, the response may be linear (**left**) or exhibit a tipping point (**right**). Data-driven knowledge of community assembly processes is much needed to inform such models. Adapted from Scheffer *et al.* (2012).

Climate change has become the foremost threat to many ecosystems, especially those that are less directly impacted by human activities. Species distribution modelling is an important tool to predict the effect of climate change on biodiversity (Miller, 2010). It consists in inferring the abiotic requirements of individual species from their observed geographic distribution, and predicting their future distribution based on predicted changes in abiotic conditions. The need to take into account processes other

than abiotic requirements, such as species interactions, dispersal limitation, adaptation, and phenotypic plasticity, has long been acknowledged (Guisan & Thuiller, 2005), nevertheless most predictions are still obtained while ignoring these processes (Wisz *et al.*, 2013). Another approach to predicting the effect of climate change on ecosystems is through the dynamical simulation of ecosystems, either by simulating each organism individually or using coarser models (Fisher *et al.*, 2014). Building such models, especially at the level of individual organisms, requires a clear understanding of the processes relevant to community assembly and dynamics.

Lastly, knowledge of community assembly is necessary to guide conservation efforts. Assumptions on the mechanisms of community assembly play a key role in the debate on the optimal design of natural reserves (Cabeza & Moilanen, 2001) or on species sensitivity to extinction (Tilman *et al.*, 1994). Such assumptions are also required to estimate the amount of biodiversity harboured in species-rich and poorly known ecosystems. A straightforward way to proceed is to assume that the relationship between the number of individuals and the number of species, observed for a sample of individuals, holds for the entire ecosystem. This reasoning implies that community assembly can be regarded as random at the scale of the ecosystem. It has been for instance applied to Amazonian trees, yielding an estimated total of 16,000 tree species extrapolated from about 5,000 observed species (ter Steege *et al.*, 2013).

2. Deterministic processes

The deterministic processes of community assembly can be decomposed into two major components: abiotic filtering and biotic interactions.

'Abiotic filtering' is a metaphor referring to the fact that species can only establish themselves in locations where abiotic conditions suit their needs: hence, any given location hosts only a subset of the species that would have the ability to reach it (Kraft *et al.*, 2015). While this concept is very general, it has its roots in the study of plant community assembly (Noble & Slatyer, 1977). In this context, abiotic filters may

include temperature, precipitation, soil nutrients, soil pH, soil grain size, soil water content, soil depth and bedrock.

Biotic interactions refer to any type of interaction between organisms, either between or within species, and can be broadly categorized into competition, predation, parasitism, commensalism and mutualism (Schemske *et al.*, 2009). Biotic interactions may facilitate or hinder the establishment of a species in a community depending on the type of interaction, and as such their action on community assembly may be referred to as 'biotic filtering'. Biotic and abiotic filtering are sometimes jointly referred to as 'habitat filtering' (Maire *et al.*, 2012). Indirect biotic interactions across trophic levels may have complex and non-trivial outcomes. For instance, if we assume that a trophic network can be decomposed into discrete trophic levels, increasing abundances among the species belonging to a given trophic level (e.g., carnivores) lead to decreasing abundances in the trophic level immediately below (e.g., herbivores), and in turn to increasing abundances one level lower (primary producers), a process known as a 'trophic cascade' (Paine, 1980; Polis *et al.*, 2000). Interspecific interaction may also take the form of a modification of surrounding abiotic conditions by organisms, for instance by so-called 'ecosystem engineer' species (Wright *et al.*, 2002), or simply through shading in the case of plants, thus blurring the line between abiotic and biotic filtering.

Within a single trophic level, competition is considered to be the dominant type of biotic interactions (Chesson, 2000). The 'competitive exclusion principle' states that the coexistence of two species competing for the same resource is not stable (Gause, 1932; MacArthur, 1958; Hutchinson, 1961; Armstrong & McGehee, 1980). Indeed, if one of the species has an even slight competitive advantage, it will eventually outcompete the other. Thus, any set of coexisting species is expected to exhibit differences in the way they exploit their habitat. This has led to the concept of 'niche', which refers in its broader meaning to the relationship between a species and its habitat, including its resource use, its interactions with other species, and the way its occupies its habitat both spatially and temporally (cf. Fig. 2; Grinnell, 1917; Hutchinson, 1957; Chase & Leibold, 2003). A species' niche may be represented as a hypervolume in the space of all available resources and possible habitat uses.

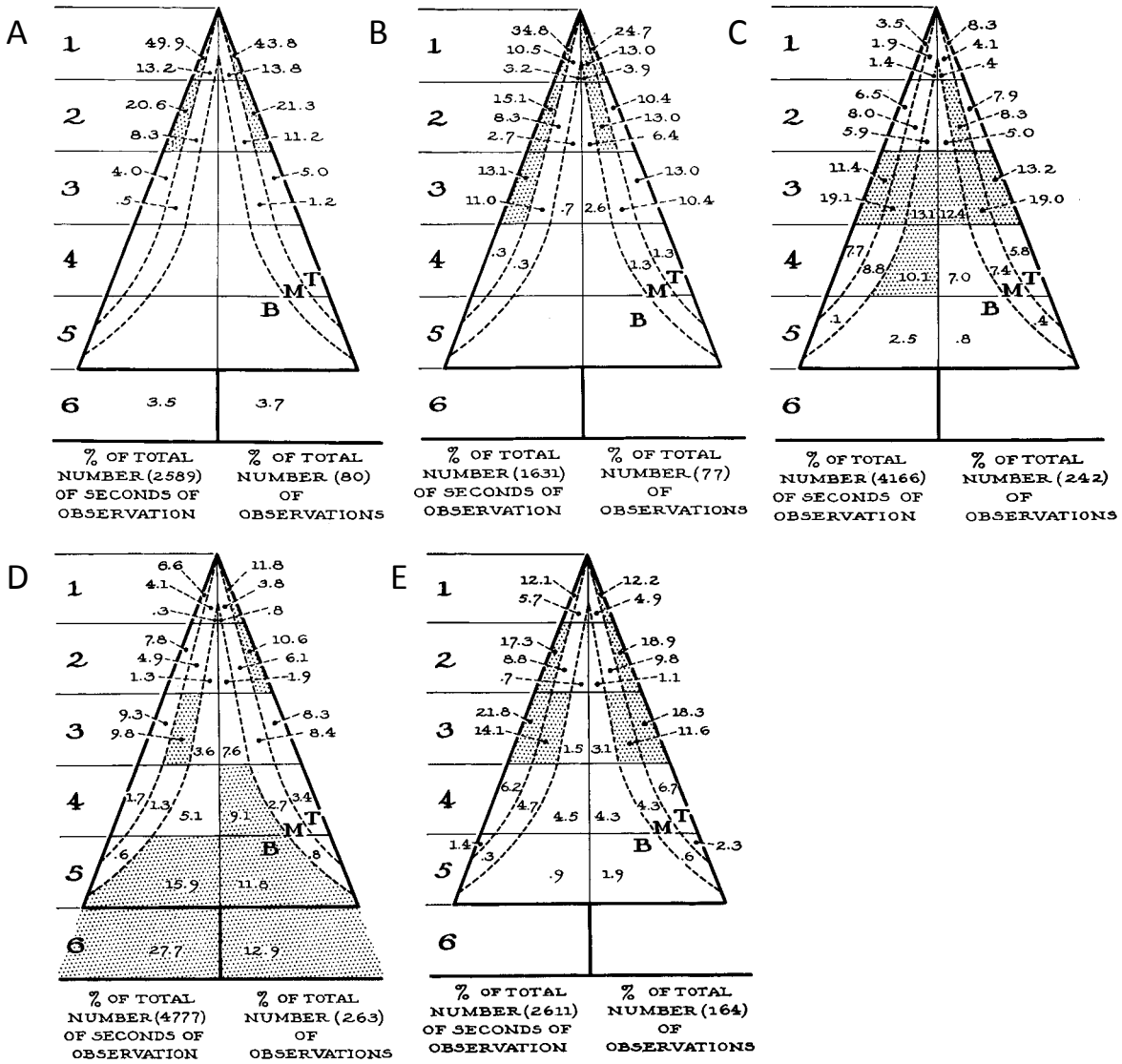

Figure 2. A classical example of niche partitioning: habitat preferences among closely related warbler species in the boreal forests of North America. (A) Cape May, (B) Blackburnian, (C) Bay-breasted, (D) Yellow-rumped, and (E) Black-throated Green warblers favour different tree layers and different tree heights when foraging for insects during the breeding season. Adapted from MacArthur (1958).

In spite of theoretical predictions, the coexistence of many similar species competing for a common resource in homogeneous environments is a common occurrence in nature. This is for instance the case in species-rich communities such as tropical forest trees and oceanic phytoplankton communities. This apparent paradox has been called the 'paradox of the plankton' (Hutchinson, 1961). Thus, additional

mechanisms need to be considered to account for species coexistence in such communities (Tilman, 1982; Chesson, 2000). Even though a vast number of potential mechanisms of species coexistence has been proposed (Palmer, 1994), they can be roughly divided into 'equalizing' mechanisms, that reduce competitive differences between species, and 'stabilizing' mechanisms, that balance the effect of interspecific competition (Chesson, 2000).

Intraspecific competition represents one stabilizing mechanism. It has indeed been found empirically that competition among conspecific individuals is often at least as intense as among different species (Connell, 1983). Predation and parasitism are another important cause of negative intraspecific interactions among prey or host species. Indeed, the fact that predators and parasites tend to specialize on one or a few species induces a 'negative density-dependence', i.e. favours lower population densities. This effect, known as the Janzen-Connell effect, was first proposed for tropical forest trees (Connell, 1970; Janzen, 1970). Lastly, spatial and temporal fluctuations in environmental conditions are also a stabilizing mechanism favouring species coexistence (Chase & Leibold, 2003; see section I.4 below).

Competition, predation and parasitism act also as equalizing mechanisms. Indeed, interspecific competition eliminates less competitive species from the community, while predation and parasitism effectively offsets the competitive advantage of the most successful species (Chesson, 2000). The importance of equalizing mechanisms and intraspecific competition in species-rich communities has prompted some ecologists to propose that competitive differences between organisms could be altogether neglected in such systems (Hubbell, 2001), as discussed in the following.

3. Stochastic processes

However complex and fascinating the interplay of species' niches is, community assembly cannot be fully understood without considering the influence of geography and history on community composition (MacArthur, 1972; Ricklefs, 1987). Firstly, the

capacity to disperse is finite in all species: offspring are more likely to be found near parent individuals. Thus, community composition in a given location is dependent on the pool of species that are within dispersal distance of that location, and on random dispersal events. The limited dispersal of individuals generates spatial clusters in the distribution of a species (Houchmandzadeh, 2009), and thus causes spatial variations in community composition even in the absence of other mechanisms. Secondly, if there are no competitive differences between two competing species, the fact than one is locally common and the other rare is due to chance alone. The relative abundances of the two species are expected to fluctuate randomly over time, until one eventually goes extinct. Thus, over a sufficiently long period of time, competitive exclusion is expected to take place even in the absence of competitive differences. The larger the number of competing species in a given location, the lower the average population of each species is, and the faster the community will lose species to random demographic fluctuations. This process has been called demographic or ecological drift, by analogy to the process of genetic drift in population genetics (Etienne & Alonso, 2007).

The 'neutrality' assumption is defined as the absence of any competitive differences among individuals, irrespective of the species they belong to (Watterson, 1974; Caswell, 1976). Since dispersal limitation and demographic drift take place independently of any competitive differences between organisms, they are often referred to as 'neutral' processes, even though they are also present in non-neutral systems. Under a dynamics governed by dispersal limitation and demographic drift, ecological communities never reach equilibrium: their composition indefinitely shifts over time. Nevertheless, if the total number of individuals, the species richness, and the dispersal capacity of individuals remain constant over time, community structure reaches a stationary state that can be described statistically as a function of these parameters.

MacArthur & Wilson (1967) were the first to build dispersal limitation and demographic drift into a model, which they used as a foundation for a 'theory of island biogeography' aimed at explaining species richness on islands. They reasoned that the number of species found on a coastal island results from an equilibrium between

immigration of new species from the mainland and species extinction on the island through demographic drift, even though they did not explicitly interpret their theory as neutral. The two processes are stochastic and their relative frequency determines the number of species found on the island at any given time (cf. Fig. 3). They further assumed that the immigration rate depends on the distance to the mainland, and that the extinction rate depends on the island's area, thus enabling empirical comparison of their theory to observations (Simberloff & Wilson, 1969).

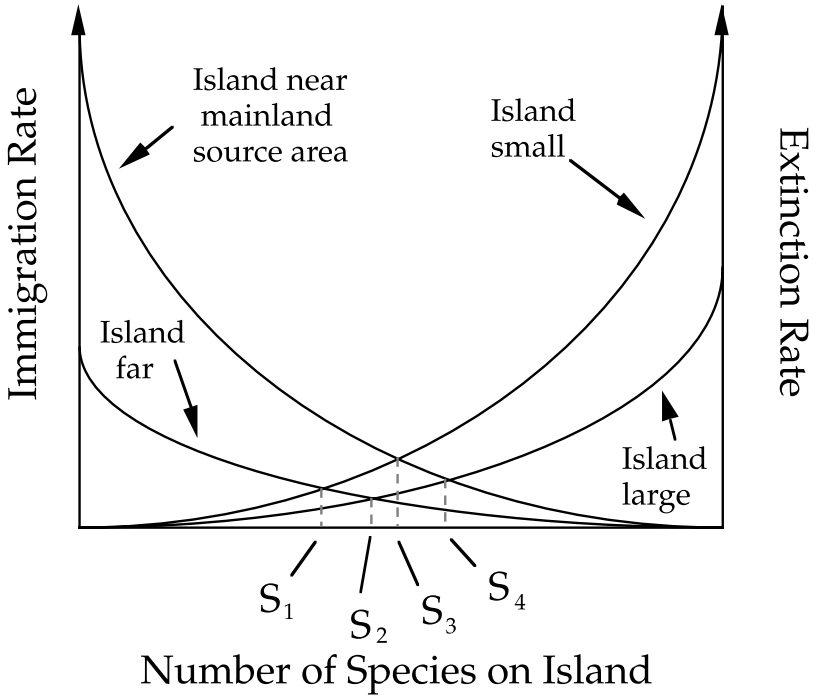

Figure 3. MacArthur & Wilson (1967) were the first to combine dispersal limitation and demographic drift into a simple model, that aims at explaining the number of species found on islands. They assumed that the number of species results from a dynamic equilibrium between stochastic immigration and extinction, which are dependent on distance to the mainland and on island size, respectively. Adapted from Hubbell (2001).

The theory of island biogeography was later expanded to better account for empirical observations (Brown & Kodric-Brown, 1977). It was also proposed that it might apply more generally to any patch of isolated habitat (Brown, 1978). In parallel, Watterson (1974) and Caswell (1976) used the mathematical tools of population

genetics to model neutral communities at the level of individual organisms instead of the level of species, thus providing a more mechanistic description of the processes, but without including dispersal limitation. Hubbell (1979, 1997, 2001) eventually combined both ideas into an influential neutral model, which he used as a basis to propose a 'unified neutral theory of biodiversity and biogeography'. His theory not only states the importance of demographic drift and dispersal limitation for community assembly, but also proposes that they may be the dominating mechanisms in some species-rich communities, especially tropical forest trees and coral reefs. Indeed, strong interspecific competition and predation could act as equalizing mechanisms between species in these communities, as mentioned earlier, and combine with strong intraspecific competition to make all individuals of all species effectively equivalent (Scheffer & van Nes, 2006). Another hypothesis is that in highly diversified communities, complex interspecific interactions could average out at the scale of the community, leading to an 'emergent neutrality' (Holt, 2006).

In Hubbell's model, the mainland's species reservoir, called the 'metacommunity', undergoes a demographic drift where random extinctions are offset by random speciation events. The island, or 'local community', also undergoes a demographic drift, but random extinctions are offset by the dispersal, or immigration, of individuals from the source metacommunity. Since the model is neutral, all individuals are considered to have the same dispersal capacity, irrespective of the species they belong to. The scope of the theory is not limited to isolated habitat patches: the local community may represent any spatially delineated ecological community, while the metacommunity represents the regional pool of species constituted by the aggregation of all local communities. The model is controlled by two parameters, the frequency of speciation events in the metacommunity, which determines the regional species richness, and the frequency of immigration into the local community. The immigration flux into the local community modulates its connectivity with the metacommunity: the stronger the immigration flux, the more species-rich and the more similar to the metacommunity the local community is. Hubbell's model and subsequent related neutral models (Etienne & Alonso, 2007) are amenable to several quantitative

predictions, and thus to statistical testing (this is discussed in more details in sections II.1 and III.4).

Two distinct neutrality assumptions can be distinguished in Hubbell's neutral theory: one regarding the metacommunity dynamics, over an evolutionary timescale, and one regarding the local community dynamics, over the timescale of an individual's lifetime. Predictions regarding local community structure, namely the relationship between area and species richness, the decay of taxonomic similarity with distance, and the distribution of relative species abundances (see section II.1), integrate both assumptions. They are in good qualitative agreement with empirical data (Hubbell, 2001), nevertheless most datasets exhibit quantitative departure from neutrality (McGill *et al.*, 2006). The assumption of a neutral diversification dynamics in the metacommunity can be tested separately, and has been shown to be unrealistic. Indeed, the mean species age predicted by Hubbell's model are not consistent with empirical measurements (Ricklefs, 2003, 2006), and the shape of the predicted phylogenetic trees does not match that of empirically reconstructed trees (Davies *et al.*, 2011). Hence, recent approaches have instead focused on testing separately the assumption of local neutral assembly through immigration, with contrasting results depending on the system (Sloan *et al.*, 2006; Jabot *et al.*, 2008; Ofiteru *et al.*, 2010; Harris *et al.*, 2015).

Even though comparison of empirical patterns to model predictions suggests that real ecological communities are rarely neutral, Hubbell's neutral theory retains important merits (Alonso *et al.*, 2006). Indeed, it has been pointed out that all the processes of community ecology are underpinned by only four fundamental processes: natural selection, demographic drift, speciation, and dispersal (Vellend, 2010). Yet, the majority of ecological literature focuses on only one of them, natural selection, which underpins all niche differences between species and thus all deterministic ecological processes. In contrast, Hubbell's neutral theory focuses on the three remaining fundamental processes, which are inherently stochastic, and places them in a quantitative framework. In practice, neutral models are essential tools for two main uses (Rosindell *et al.*, 2012). Firstly, they may serve as a 'null model' against which empirical patterns can be contrasted, so as to identify cases where neutral processes are

sufficient to explain the data and cases where they are not. Secondly, they may serve as a parsimonious approximation to real systems, and as a foundation for incorporating relevant non-neutral mechanisms, such as niche differences (Chisholm & Pacala, 2010), environmental stochasticity (Kalyuzhny *et al.*, 2015), negative density-dependence (Du *et al.*, 2011), or a more realistic speciation dynamics (Rosindell *et al.*, 2010).

4. Spatial and temporal scales

Community assembly involves a range of temporal and spatial scales spanning many orders of magnitude - from the evolutionary timescale to the behaviour of individual organisms, and from the global scale to the scale of microorganisms (Chave, 2013). The continental scale is the realm of biogeography, where species distribution reflects the geological and evolutionary history of continents (Cox *et al.*, 2016), as well the latitudinal gradient of diversity (Hillebrand, 2004). At the opposite end, most studies on species interactions focus on a limited number of individuals. Community ecology is concerned with the intermediate scales (Lawton, 1999): namely, within a biogeographic unit (Morrone, 2015), but encompassing a number of individuals large enough for statistical patterns to emerge. The scale at which statistical patterns start emerging depends on the type of organisms considered, and will differ by many orders of magnitude between plants and bacteria.

Niche and neutral processes might alternately dominate at different spatial and temporal scales. Firstly, locally observed species interactions do not preclude random species assembly over larger spatial and temporal scales. Indeed, the majority of interspecific interactions are opportunistic and vary across space and time (Holt, 1996; Poisot *et al.*, 2014), despite much-studied instances of specialized interspecific interactions such as plant-pollinator mutualisms (Rønsted *et al.*, 2005). Secondly, species dynamically adapt their niche to the local competitive context, either through plasticity or through natural selection. For instance, closely related species with mostly disjoint geographical distributions are known to display greater phenotypic differences

(such as a difference in size) wherever they co-occur, a process known as 'character displacement' (Brown & Wilson, 1956). Natural selection has been found to have measureable effects on phenotype over timescales as short as a few generations when species are confronted with a sudden change in their biotic or abiotic environment, thus questioning the legitimacy of the traditional separation between the timescales of evolutionary and community assembly processes (Ghalambor *et al.*, 2015).

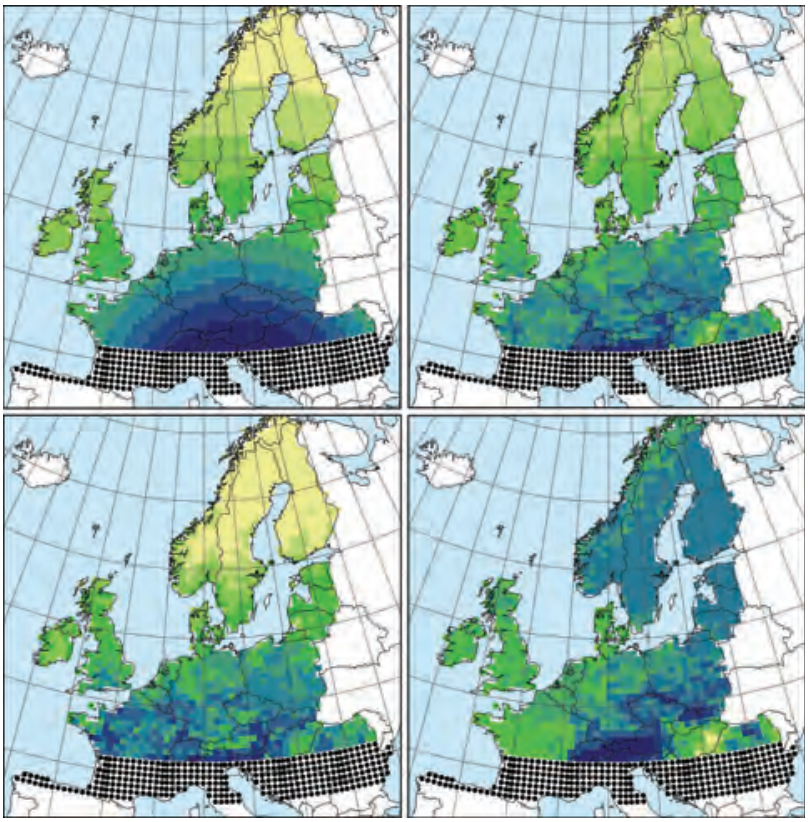

Figure 4. Community assembly processes depend on the spatial and temporal scales considered: current geographical patterns of tree diversity in Europe might reflect on-going dispersal from ice age tree refugia, which started 14,000 years ago. **Top right, bottom left and bottom right**: geographical distribution of tree diversity (increasing from yellow to blue) for all 60 European tree species, the 45 temperate species and the 15 boreal species, respectively. **Top left**: accessibility through dispersal from ice age tree refugia (black dots). Adapted from Svenning & Skov (2007).

Another key aspect of community assembly is how fast community composition responds to abiotic change, relative to the pace of the abiotic change itself. Indeed, if abiotic change is fast enough relative to community response, the community may never reach equilibrium, thus leading to an apparently random dynamics. This

phenomenon may be more pervasive than it seems: for instance, it has been shown that the dispersal of tree species in Europe following the end of the last ice age is still an on-going process (cf. Fig. 4; Svenning & Skov, 2007). In contrast, organisms with short generation time and high dispersal ability are able to track environmental changes more efficiently. Additionally, if several local communities are connected by a permanent and strong enough dispersal flux, they may never reach the optimal composition that would be expected based on local abiotic conditions (Gravel *et al.*, 2006). A local community will also be more prone to demographic stochasticity if it hosts a smaller population size (Fisher & Mehta, 2014). These observations have led to the development in the last decade of 'metacommunity theory', a family of mathematical models aiming at reconciling neutral and niche processes by explicitly accounting for spatial and temporal dynamics (Leibold *et al.*, 2004). However, unlike simpler neutral models, these models do not provide predictions that are easily amenable to statistical comparison with empirical data.

Lastly, most of the existing knowledge on community assembly comes from the study of plants and vertebrates, and the extension of community ecology to microorganisms is comparatively very recent (Curtis & Sloan, 2005; Martiny *et al.*, 2006; see section II.2). While the fundamental processes of community assembly apply to all living organisms, they operate over very different scales for microorganisms, and their relative importance is likely to differ (Hanson *et al.*, 2012). It has long been considered that microorganisms had effectively infinite dispersal capacity, and that abiotic filtering was the dominant process of community assembly (Baas Becking, 1934). Microbial communities have indeed been found to be very sensitive to local abiotic conditions and dominated by specialist taxa (Ramirez *et al.*, 2014; Mariadassou *et al.*, 2015). Nevertheless, this view has now been nuanced, and dispersal limitation has been shown to play a role as well (Ofiteru *et al.*, 2010; Martiny *et al.*, 2011; Roguet *et al.*, 2015). While microorganisms tend to be more cosmopolitan than larger organisms, biogeographic patterns do exist (Hanson *et al.*, 2012; Livermore & Jones, 2015). Microorganisms have also been found able of complex interactions beyond competition (Cordero *et al.*, 2012).

II. DNA-based biodiversity patterns

Most of ecological knowledge comes from studies performed at the level of individual species, and from this perspective, the singularity of each species and sometimes of each individual is striking. Thus, ecologists have long wondered whether general laws were hiding behind the collection of idiosyncrasies (Lawton, 1999). Integrative data on species richness, abundance and spatial occurrence have been gathered with the hope that they would yield insight into the general mechanisms of community assembly (Brown, 1995). The underlying idea is that, as in statistical physics, informative statistical properties might emerge from the observation of a large enough number of individuals and species irrespective of the details of species identities.

In this section, I first introduce two types of integrative patterns that have been widely studied in community ecology: the distribution of species abundances, and spatial patterns. I then discuss why the emergence of automated data collection is opening new horizons for the study of these patterns. Lastly, I briefly present the ecosystem that this thesis more specifically focuses on, the tropical forests of French Guiana.

1. Integrative biodiversity patterns

a. Species relative abundances

The distribution of species abundances in a random sample of individuals takes two forms in the ecological literature: the 'rank-abundance distribution' (RAD), or 'Whittaker's plot', consists of the abundances n_i of all S species in the sample ranked by decreasing abundance, while the 'species abundance distribution' (SAD), or 'Preston's

plot', is the distribution of the number Φ_n of species having abundance n for all the possible n values in $\{n_1, \dots, n_S\}$ (cf. Fig. 5; Preston, 1948; Whittaker, 1965). To accommodate the limited amount of data, the SAD is usually binned into abundance categories. This binning step leads to a loss of information, thus the RAD is more informative than the SAD. Nevertheless, the SAD has often been the preferred distribution because it is easier to handle mathematically and to derive from theoretical models. This is linked to the fact that it can be interpreted upon normalization as the probability distribution for the abundance of a randomly chosen species in the sample. Because of the wide range of abundances typically observed in empirical data, abundances are often log-transformed in SAD and RAD – in SAD, this amounts to binning species into abundance classes of exponentially increasing width from the lowest abundance class (one individual) to the highest, following the example of Preston (1948).

It was noticed early on that the distribution of species abundances tended to be similar in species-rich communities. Indeed, within a single trophic level, there are usually a few common species and a long tail of rare species – simply put, 'most species are rare' (cf. Fig. 5). This spurred attempts at finding a general explanation for this pattern. Fisher *et al.* (1943) and Preston (1948) were the first to propose statistical distributions to fit the distribution of species abundances.

Fisher assumed that the sampled species abundances followed a negative-binomial distribution without the zero-abundance class, and derived a SAD of the form $\mathbb{E}[\Phi_n] = \alpha x^n / n$, where α is a constant parameter, x is a function of α and of sample size N (with $0 < x < 1$), and $\mathbb{E}[\Phi_n]$ is the statistically expected value of Φ_n (cf. section III.3.b; Chave, 2004). Since $\sum_{n=1}^{\infty} \mathbb{E}[\Phi_n] = -\alpha \ln(1 - x)$, this distribution is called the 'log-series'. A remarkable property of this model is that the expected number of species $\mathbb{E}[S]$ in the sample is given as a function of the number of sampled individuals N by $\mathbb{E}[S] = \alpha \ln(1 + N/\alpha)$. Hence, the parameter α is sufficient to predict the observed species richness as a function of the sampling effort. It can thus be used as a sampling-independent measure of the community's diversity. The value of α can be easily

visualized in the RAD representation, since the log-transformed abundances are expected to decrease linearly with slope $-1/\alpha$ as a function of species rank (cf. Fig. 5).

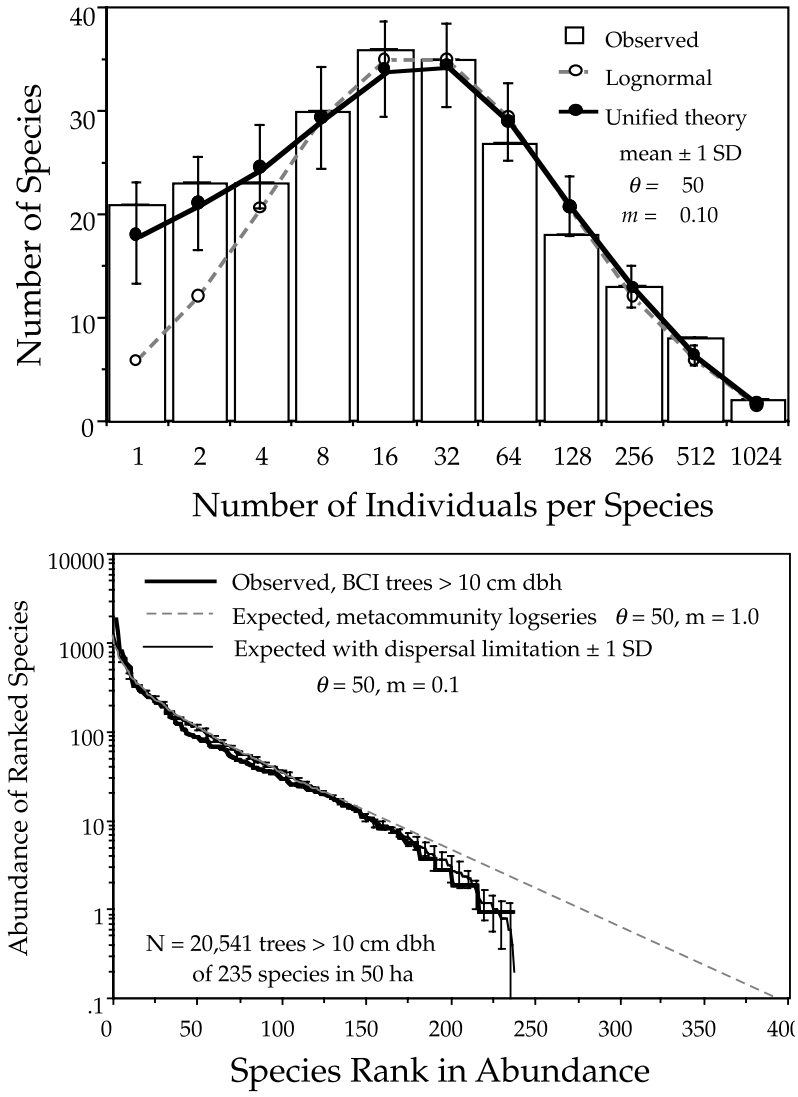

Figure 5: Species Abundance Distribution (**top**) and Rank Abundance Distibution (**bottom**) for mature trees in the 50-ha Barro Colorado Island (BCI) monitored forest plot (Panama). Mature trees are defined as stems with diameter larger than 10 cm at breast height (or '> 10 cm dbh'). The dispersal-limited Hubbell's model is fitted to the data ($\theta = 50$, $m = 0.1$), and is compared with the log-normal SAD (**top**; dashed line), and with the RAD of Fisher's model (**bottom**; dashed line). Fisher's model is equivalent to Hubbell's model without dispersal limitation (i.e., case $m = 1$) for large sample size. Error bars indicate ± 1 standard deviation.

Preston (1948) argued in contrast that a log-normal SAD best fitted empirical data, i.e. $\mathbb{E}[\Phi_n] \propto e^{-(\ln n - \mu)/(2\sigma^2)}$ with μ and σ constant parameters. A notable difference between the two SADs is that the log-normal distribution exhibits a mode (i.e., the abundance class with the most species is not the lowest abundance class), while Fisher's log-series does not. Preston explained the fact that both situations could be encountered in empirical data by the effect of sampling: a community in which the 'true' SAD (i.e., for

an infinite number of individuals) is log-normal can lose its mode if under-sampled, and be mistaken for a log-series. It has since then been acknowledged that the effect of sampling is indeed paramount in our ability to distinguish between differently-shaped SAD by curve-fitting (Sloan *et al.*, 2007). In the RAD representation with log-transformed abundances, a log-normal SAD takes the form of an S-shaped curve, the common species being commoner and the rare species rarer than in Fisher's log-series.

Later models have focused on finding a mechanistic justification for the proposed distributions. MacArthur (1957) proposed that species relative abundances resulted from the random partitioning of the niche space between the different species of the community. A number of more sophisticated niche partitioning models' were subsequently proposed (Tokeshi, 1996; McGill *et al.*, 2007). However, Hubbell's neutral model is the mechanistic model that has been the most successful at fitting empirical SADs (Hubbell, 2001; cf. section I.3 and III.4). Indeed, the metacommunity SAD converges toward Fisher's log-series for a large enough sample size and is characterized by a 'fundamental biodiversity number' θ that converges toward Fisher's α (Chave, 2004). In the absence of dispersal limitation, the local community is a random sample from the regional metacommunity, and hence also exhibits a log-series-like SAD. In the presence of dispersal limitation however, the depletion of rare species and the increase in abundance of locally common species lead to a log-normal-like SAD (cf. Fig. 5). Thus, Hubbell's neutral model can approximate both the log-series and the log-normal SADs, while providing a mechanistic justification for them and fully accounting for sampling effects.

Nevertheless, it has been shown that many types of non-neutral processes could yield SADs similar to neutral ones (Chave *et al.*, 2002; Pueyo *et al.*, 2007; Chisholm & Pacala, 2010). It has also been argued that the log-normal distribution fits empirical SADs at least as well as Hubbell's local community SAD (McGill, 2003). The log-normal is still the most popular choice when it comes to choosing a realistically-shaped SAD for modelling purposes irrespective of the underlying mechanisms (Connolly *et al.*, 2017). A log-normal SAD is not in itself very informative on the mechanisms of community assembly. Indeed, the log-normal distribution is the limiting probability distribution for

any product of sufficiently many random variables, as a consequence of the central limit theorem (cf. section III.2 and III.3.a), thus a log-normal SAD could arise as the result of any type of multiplicative process. More generally, it has been suggested that the range of empirically observed SADs could simply result from the iterative spatial aggregation of smaller-scale SADs, a phenomenon described as a 'spatial analogy of central limit theorem' (Sizling *et al.*, 2009). As a consequence, it has been called for, on the one hand, more statistically powerful tests than simple curve-fitting (Chave *et al.*, 2006; Al Hammal *et al.*, 2015), and on the other hand, testing multiple predicted patterns simultaneously instead of solely the SAD (McGill *et al.*, 2007).

b. Spatial patterns

Spatial patterns form a second family of integrative patterns in ecology. The relationship between the sampled area and the number of sampled species is the oldest such pattern to have been studied (Watson, 1859). This curve was first regarded as a mean to assess whether a community had been adequately sampled, i.e. to ensure that only a marginal number of new species would appear in the sample if the sampled area were to be increased. It was soon realized that the species-area relationship (SAR) might also contain valuable information regarding spatial community structure. Indeed, at the regional scale, the number of species S was found to consistently follow a power law $S \propto A^z$ as a function of area A, where the exponent z takes values between 0.15 and 0.40 (Arrhenius, 1921; Williamson, 1988). This 'law' has later been observed to break down at the extremes, either for areas that are below approximately 1 km^2 (for plants or vertebrates), or conversely for areas that exceed the boundaries of a single biogeographic unit (cf. Fig. 6; Preston, 1960; Shmida & Wilson, 1985). The resulting curve exhibits an 'S' shape on a log-log scale, with a linear domain in the central part corresponding to the power-law behaviour described above, and steeper slopes at both ends.

The three domains of the SAR reflect different processes at play. At the local scale, the SAR directly results from sampling the local species abundance distribution: the number of detected species first increases linearly with area and then progressively slows down as only the rarer species remain to be sampled. At the global scale, the SAR approaches linearity again as species with distinct evolutionary history are sampled in different biogeographic zones. At intermediate scales, the power-law regime reflects a slow increase in species richness with area once the local species richness has been fully sampled. This increase corresponds to a shift in species composition with distance, referred to as 'beta-diversity' by Whittaker (1960), i.e. the link between 'alpha-diversity', the number of species in the local community, and 'gamma-diversity', the number of species at the regional scale.

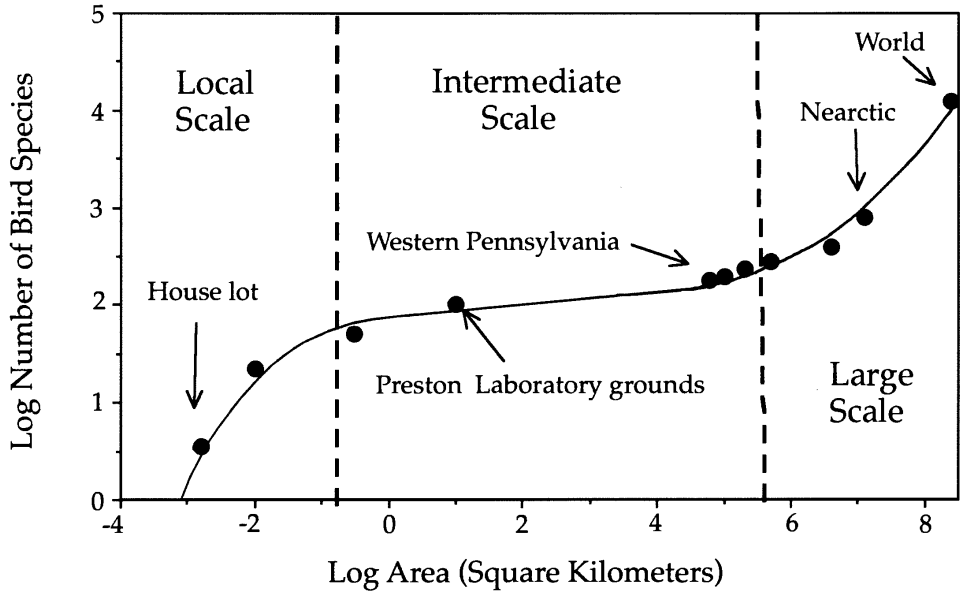

Figure 6: Number of bird species as a function of area; data from Preston (1960). The S-shaped Species-Area Relationship introduces two characteristic spatial scales for a given taxonomic group (vertical dashed lines), separating 'local', 'intermediate', and 'large' scales. The study of beta diversity mostly focuses on 'intermediate' scales, while biogeography is mostly concerned with 'large' scales. Adapted from Hubbell (2001).

Conceptually, beta-diversity is the variation in taxonomic composition among sites within a region of interest. However, several quantitative definitions coexist. One approach is to consider beta-diversity as a quantity β that links the mean local diversity

α to the regional diversity γ through $\gamma = \alpha\beta$, so that the regional diversity can be partitioned into independent within-community and among-community components (Whittaker, 1960; Jost, 2007). The spatial scale that separates alpha- and beta-diversity may be defined as the scale witnessing the regime shift in the SAR. Another approach is to measure beta-diversity independently of alpha- and gamma-diversity as the mean taxonomic similarity between sites or as the variance of the community matrix (Legendre & De Caceres, 2013). The community matrix is the matrix describing the number of individuals per species and per sites, taking usually species as columns and sites as rows. A wealth of similarity metrics can be used to compare sites to each other, depending for instance on the weight given to rare species, on whether the sampling effort is homogeneous among sites or not, and on whether abundance information or only species occurrence should be taken into account (Legendre & De Caceres, 2013).

Taxonomic similarity is well known to decrease with distance, a general pattern of ecology that is related to the monotonous increase of diversity with area (Soininen *et al.*, 2007). Depending on the mechanisms of community assembly, this 'distance-decay of similarity' can be interpreted either as the result of dispersal limitation, or as the consequence of new habitats and community types being encountered. A major motivation for the study of beta-diversity lies in the fact that it is an indirect means to investigate the drivers of community assembly. Indeed, taxonomic similarity between sites can be compared to distance and to environmental similarity, so as to empirically assess the relative importance of dispersal and abiotic filtering in shaping community composition (Tuomisto *et al.*, 2003). This question may also be addressed by directly comparing taxonomic composition with quantitative environmental descriptors using multivariate statistical methods, an approach deemed more statistically powerful (Legendre *et al.*, 2005, 2008; cf. section III.2).

Formally, the distance-decay of similarity can be described using the pair-correlation function of statistical physics, i.e. the probability for two individuals at a given distance to belong to the same species (Chave & Leigh, 2002; Zillio *et al.*, 2005; Houchmandzadeh, 2009). Predictions for both the SAR and the distance-decay of similarity can be obtained from a spatially explicit version of Hubbell's neutral model.

The neutral predictions are in qualitative agreement with observations, including the tri-phasic SAR (Hubbell, 2001; Condit *et al.*, 2002). Nevertheless, as in the case of species abundance distributions, this does not preclude other mechanisms from being involved.

2. Environmental DNA data

Collecting the large amounts of data required to study integrative patterns has long been a tedious and challenging task (Lawton *et al.*, 1998). Direct taxonomic identification relies on rare expert knowledge, and is prone to errors. Sampling protocols are difficult to standardize, and owing to the amount of work involved, data collection may spread over long periods of time – sometimes years – which may introduce biases. Last but not least, only a small fraction of biodiversity can be directly sampled and identified by a human observer, mostly vertebrates and plants. As a result, datasets available for the study of integrative biodiversity patterns have long been relatively rare and limited in their taxonomic extent. However, major technological advances have been made over the last decades that now allow for the automatic collection of ecological data. These advances are all related to the exponential increase in computer power that took place over the same period of time, and that has dramatically impacted all fields of science and industry.

For instance, remote sensing of ecological features over large spatial scales can be achieved using Lidar and hyperspectral imaging. Lidar is a small-wavelength equivalent of radar (either airborne or ground-based) that allows for fine-grain 3D imaging. Hyperspectral imaging consists in recording images (from a plane or a satellite) for a much larger spectrum of electromagnetic wavelengths than the human eye does: the additional information may for instance be used for the automated identification of tree species from their spectral signature, especially when combined with Lidar data (Alonzo *et al.*, 2014).

Arguably, the one recent technological innovation with the strongest impact on biology has been high-throughput DNA sequencing (Schuster, 2007). While DNA sequencing methods have existed since the 1970s (Sanger *et al.*, 1977), a breakthrough occurred around 2005 by which sequencing speed was multiplied by several orders of magnitude for a fraction of the cost of previous methods (cf. Fig. 7; Margulies *et al.*, 2005). Since 2011, the dominant high-throughput DNA sequencing method is Illumina sequencing, which consists in spreading and attaching the target DNA strands on a flat surface and synthesizing the complementary strands using four-colour fluorescent nucleotides (Bentley *et al.*, 2008). By recording the order of appearance of the different colours at the location of each DNA strand with a fast and high-resolution camera, millions of strands can be simultaneously sequenced with high accuracy. The main limitation of the method is on the length of the sequenced strands, which currently cannot exceed 150 or 300 base pairs, depending on the exact technology.

The idea of using DNA sequencing to study biodiversity predates high-throughput sequencing, and was introduced as a mean to study microorganisms (Giovannoni *et al.*, 1990). Indeed, most microorganisms can only be detected in the environment through their DNA, collected from soil or water samples (Pace, 1997). The idea was to identify a short DNA sequence satisfying two properties. First, it should have conserved extremities across the range of targeted taxa, so that it can be amplified by PCR using a single pair of primers from bulk DNA. Second, its central part should exhibit random mutations making the different taxa distinguishable, i.e. it should not be under strong evolutionary selection. Such a sequence is called a barcode, and the first that has been used is the 16S rRNA gene of prokaryotes, which codes for the RNA forming the small (16S) subunit of the prokaryotic ribosome (Giovannoni *et al.*, 1990; Pace, 1997). DNA barcodes have soon also been recognized as a mean to bypass the need for traditional taxonomic expertise in identifying larger organisms, for which DNA can be directly extracted from tissue (Hebert *et al.*, 2003). Nevertheless, barcode sequences can only be attributed to known taxa once a reference database has been established for the barcode. When no reference database is available for the organisms under study, molecular Operational Taxonomic Units (OTUs) defined based on

sequence similarity are substituted for species in analyses. Moreover, depending on the frequency at which mutations occur in the barcode sequence, the comparison of sequences across species may not be congruent with traditional species delineation, and two barcodes targeting the same taxonomic group may have widely differing levels of taxonomic resolution.

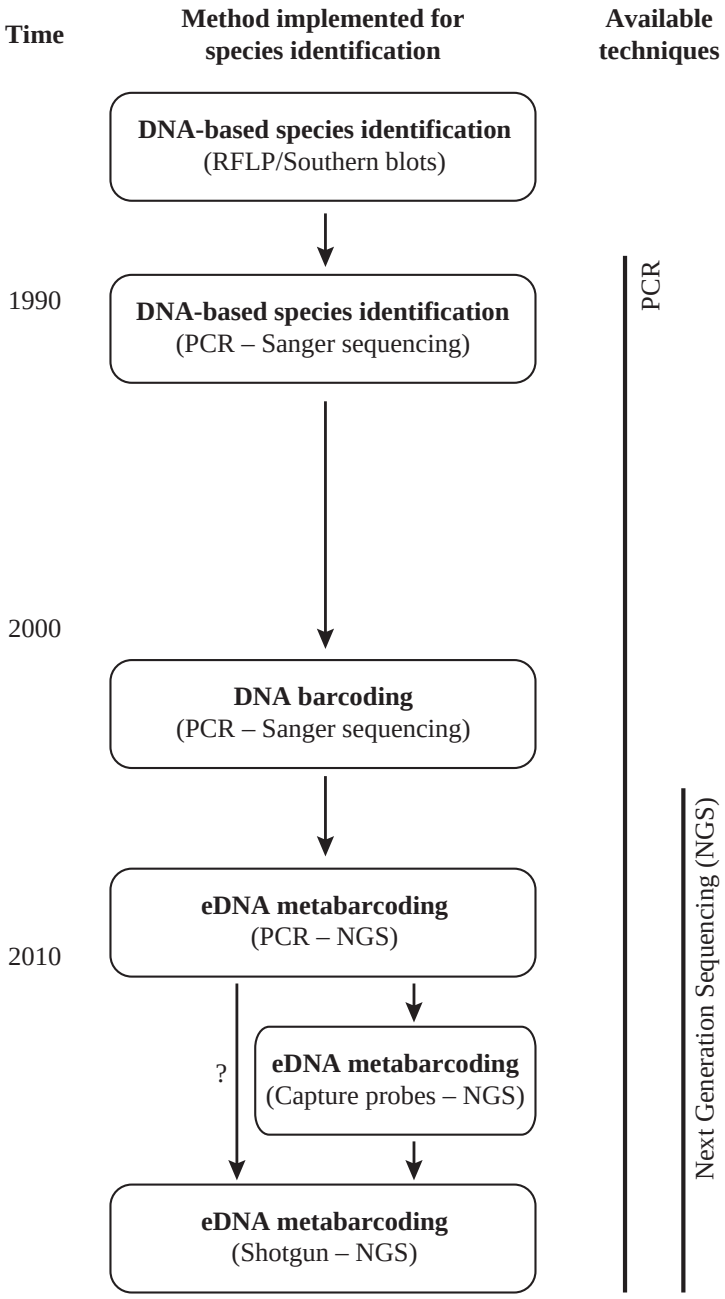

Figure 7: Evolution of DNA-based species identification over time. The successive introduction of Polymerase Chain Reaction (PCR) and High-throughput (or Next Generation) Sequencing to ecology have transformed the field, and automated data collection using molecular approaches is developing fast. Adapted from Taberlet *et al.* (2012b).

With the advent of high-throughput sequencing, several thousands to several millions of barcode sequences can now be readily sequenced from a single bulk DNA sample. As a consequence, the use of barcode sequencing to measure biodiversity from environmental DNA – or 'metabarcoding' – has boomed (Bik *et al.*, 2012; Taberlet *et al.*, 2012b,a; Bohmann *et al.*, 2014). The vast diversity of the microbial world only starts to be fully grasped, and whole new swathes of the tree of life are being discovered (Hug *et al.*, 2016). Ambitious projects aim at sampling microbial diversity across the globe, either on land (Gilbert *et al.*, 2014) or in the ocean (de Vargas *et al.*, 2015). In parallel, metabarcoding can be used as a fast and standardized means to gather information on macroscopic organisms, either using environmental DNA or, for small enough organisms, DNA extracted from a 'soup' of sampled specimen (Andersen *et al.*, 2012; Yu *et al.*, 2012; Gibson *et al.*, 2014). This wealth of data has led to a renewal of interest in the study of integrative biodiversity patterns and biogeography, which were until recently entirely unknown for microorganisms (Martiny *et al.*, 2006; Fuhrman, 2009; Hanson *et al.*, 2012). Since metabarcoding is but the simplest method to exploit the information contained in environmental DNA, and is being replaced by approaches making use of a larger fraction of the organisms' genome as sequencing capacity keeps increasing (Taberlet *et al.*, 2012b), the trend toward incorporating sequencing data into ecological studies is probably just starting.

3. The tropical forests of French Guiana

Tropical forests are estimated to concentrate half of global biodiversity, and are as such the archetypical 'hyperdiverse' ecosystem (Scheffers *et al.*, 2012). They have played an historical role in generating hypotheses in ecology and evolution, especially regarding the mechanisms of species coexistence (Wright, 2002). Indeed, like the phytoplanktonic communities at the origin of the 'paradox of the plankton' (Hutchinson, 1961), they harbour for many taxonomic groups a wide range of species competing for the same resources. Hubbell's neutral theory of biodiversity has been elaborated based primarily

on the observation of tropical forest tree communities (Hubbell, 2001), and much of the ensuing debate has initially focused on these communities as well (McGill, 2003; Ricklefs, 2003; Volkov *et al.*, 2003). In addition to their unparalleled biodiversity, tropical forests are also thought to harbour the majority of the non-microbial terrestrial taxa still unknown to science (Scheffers *et al.*, 2012). Hence, the automated measurement of integrative patterns is well suited to their study, and is in particular uniquely comprehensive compared to other possible approaches.

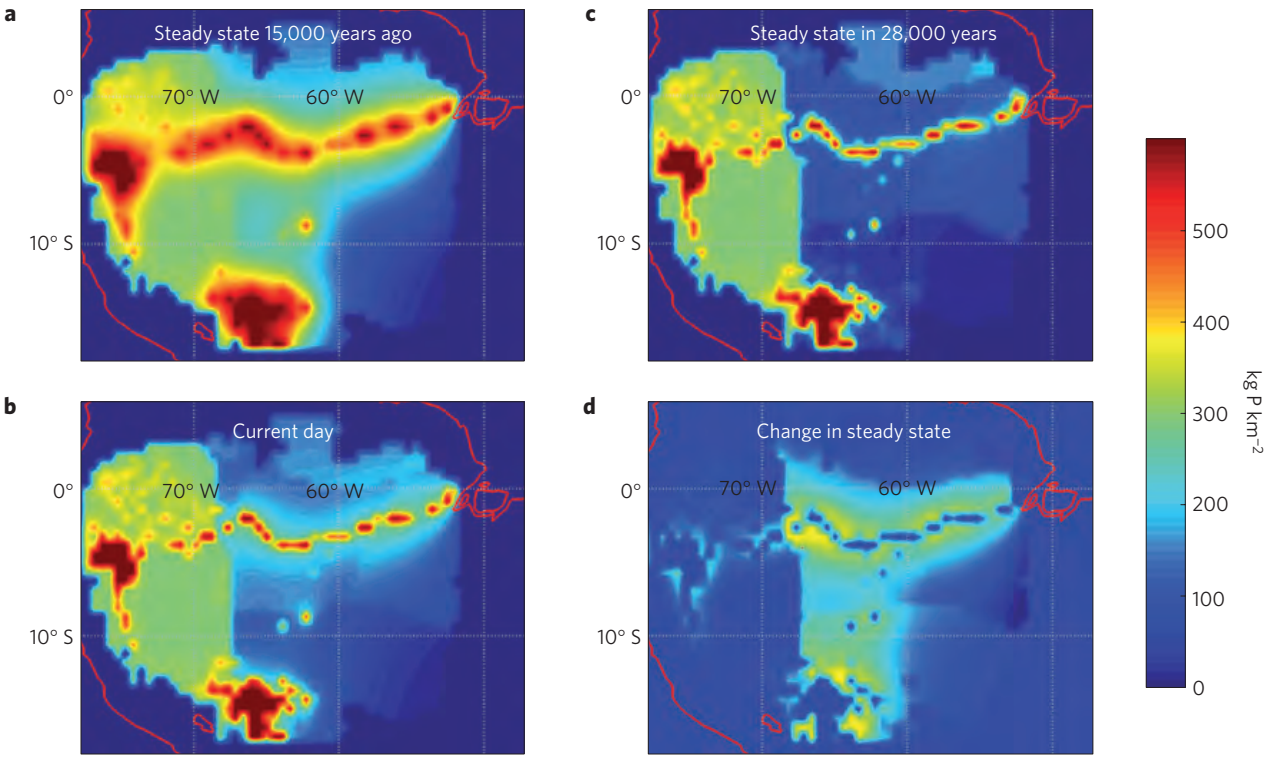

Figure 8. Whether ecosystems can be considered pristine depends on the temporal scale considered: map showing estimated changes in phosphorus (P) concentrations over time in South America following the sudden extinction of most large mammal species 12,000 years ago and the consecutive disruption of nutrient transport through dung, likely caused by human arrival. Adapted from Doughty *et al.* (2013).

Unlike most land ecosystems on Earth, a significant, if fast dwindling, fraction of tropical forests can still be considered to be in a pristine state, thus guaranteeing access

to natural processes unaffected by human activities. Amazonia represents the world's largest tropical forest, and within it, French Guiana counts among its least disturbed parts (Hansen *et al.*, 2013). Nevertheless, recent findings have challenged the idea of an entirely pristine Amazonian basin. Indeed, it appears that human population density was relatively high in places until the European conquest (Heckenberger *et al.*, 2008). Moreover, on a longer timescale, the Amazonian basin may still be in a transient state following the sudden disappearance of most large mammal species 12,000 years ago, which was likely caused by the arrival of human hunters and has had deep consequences on nutrient transport and seed dispersal (cf. Fig. 8; Doughty *et al.*, 2013).

Two research stations have been established in French Guiana in the 1980s for research on Amazonian biodiversity. This is where the data used in this thesis have been collected. The Nouragues research station is about 100 km inland, in the heart of the Nouragues natural reserve, and is devoted to the study of the undisturbed lowland forest as well as of the neighbouring inselberg. The Paracou research station, near the coast, is devoted to the study of the long-term effects of logging on biodiversity (Gourlet-Fleury *et al.*, 2004). In both stations, soils are acidic and nutrient-poor, as is typical in tropical forests, with a more sandy soil in Paracou and a more clayey soil in Nouragues. The mean rainfall is about 3,000 mm per year, with relatively strong seasonal variation, and temperature is around 26°C throughout the year.

III. Statistical approaches

In this section, I introduce the statistical approaches used in this thesis. I first briefly review the classical approaches of community ecology. I then introduce Hubbell's neutral model and Dirichlet mixture models, which are respectively the foci of the second and third chapters of this thesis, by emphasizing their common mathematical structure based on the Dirichlet distribution and its Dirichlet process extension.

1. Comparing models to data in ecology

In physics, empirical data can often be satisfyingly characterized by a one-dimensional mathematical function fitted to the data ('curve fitting'). This is not the case in ecology, because observations do not as a rule tightly follow the prediction of a theoretical model, and because data points are always relatively scarce and costly to acquire. To take full advantage of the available data, it is hence essential to account for the statistical distribution of the observations around the fitted model, and often for the statistical dependence between observations. In the absence of a theoretical prediction, deterministic trends in the relationship between variables are conversely assumed to be very simple (e.g., linear). Thus, ecological models aiming at comparison with data need to be expressed in probabilistic terms, and model fitting heavily relies on likelihood-based inference (Fisher, 1925; Pawitan, 2001).

The likelihood function of a model is given by the probability distribution $p(X|\theta)$ for the data X to be observed conditional on the model's parameters θ. The model is fitted to data by maximizing the likelihood function $L(\theta|X) = p(X|\theta)$, which is a means of simultaneously estimating the model's parameters as $\hat{\theta}(X) = \text{argmax}_\theta[L(\theta|X)]$ and measuring the goodness-of-fit as $\hat{L}(X) = \max_\theta[L(\theta|X)]$. In practice, the logarithm of the likelihood is maximized, and the normalization factor in the likelihood expression is discarded. Depending on the situation, the focus may be on measuring goodness-of-fit

or on estimating and interpreting model parameters. If several alternative models are to be compared to each other, this can be achieved by comparing the Akaike Information Criterion for each model, equal to $2K - 2\ln\hat{L}(X)$, where K is the number of parameters in the model (Akaike, 1974; Burnham & Anderson, 2002). If only one model is to be compared to the data, the most popular approach is to assess how likely the data at hand would be to be observed if they were to be generated by the probabilistic model under consideration. To this end, the value taken by a 'test statistics' - for instance the log-likelihood $\ln\hat{L}(X)$ - is compared to its theoretical distribution given the model. The threshold for rejecting the model with reasonable confidence is traditionally set at 5% probability, following the example of Fisher (1925).

Another approach to likelihood-based inference consists in estimating the full probability distribution of the model's parameters conditional on the data instead of only their most likely value (Gelman *et al.*, 2014). This approach is called Bayesian inference, in contrast to maximum-likelihood inference, since the full probability distribution of the model's parameters θ is given by Bayes' equation $p(\theta|X) = p(X|\theta)p(\theta)/p(X)$ (Bayes & Price, 1763). Another distinction between both approaches is that maximum-likelihood inference assumes that $\mathrm{argmax}_\theta[p(\theta|X)] = \mathrm{argmax}_\theta[p(X|\theta)]$, and thus implicitly that $p(\theta)$ is a uniform distribution. In contrast, $p(\theta)$ is often used to express prior belief on parameter values in Bayesian inference. The normalization factor $p(X) = \int_\theta p(X|\theta)p(\theta)$, or marginal likelihood, can then be used as a measure of goodness-of-fit accounting for all possible parameter choices. Because it is less analytically tractable than maximum-likelihood inference, Bayesian inference has been less employed historically. However, it can now be performed numerically, and even though it is usually more computationally demanding than maximum-likelihood inference, it has become increasingly popular with the steady increase in computer power. One of the reasons of its success is that it can accommodate complex models in which the likelihood is difficult to maximize.

2. The statistical tools of community ecology

Univariate models such as simple linear regression, where observations are regarded as realizations of a singe random variable, can be distinguished from multivariate models where observations result from several non-independent random variables. The analysis of community matrices relies on multivariate statistical methods, where the abundance, or the occurrence, of each of the p taxa is regarded as a random variable with a realization at each of the n sampling sites. Not all of the many multivariate methods classically used in community ecology are explicitly model-based: they typically combine multivariate linear regression, eigenvalue decomposition and the use of (dis)similarity metrics (Legendre & Legendre, 2012). Their results are often interpreted within the framework of the 'analysis of variance' (ANOVA), which consists in partitioning the variance of the observed variables into components corresponding to different sources of variation.

The multivariate methods that include an eigenvalue decomposition step (or a generalized version of it) are called 'ordination' methods. A cornerstone of multivariate analysis is Principal Component Analysis (PCA), a simple ordination method of widespread use well beyond ecology. It consists in rotating p observed variables around their mean so as to obtain p uncorrelated variables ordered by decreasing variance. Namely, the n-by-p matrix T containing the p new variables is obtained as the matrix product $T = XW$, where X is the n-by-p matrix containing the centred original variables, and W the p-by-p matrix formed by the eigenvectors of the covariance matrix $1/(n-1)\, X^T X$ ordered by decreasing eigenvalues. The first use of PCA is to decorrelate the data. It may also be used for reducing data dimensionality by discarding the independent variables accounting for the least variance. Thus, PCA allows for conveniently representing the data by projecting them on the two or three axes that account for the most variance. To investigate the dependence of a community matrix on a set of explanatory variables, such as environmental variables measured at the sampling sites, a classical method is to perform a multivariate linear regression of the community matrix on the explanatory variables, followed by a PCA on the matrix of

fitted coefficients, a method known as Canonical Redundancy Analysis (RDA). Using partial linear regression, RDA can be extended into 'partial RDA' to compare the effect of several sets of explanatory variables on the community matrix.

Clustering methods constitute another family of extensively used statistical methods in ecology (Legendre & Legendre, 2012), as well as more generally in data mining and machine learning (Bishop, 2006; Jain, 2010). They aim at partitioning the data into 'natural' clusters of observations, by searching for structure in the matrix of pairwise similarity between observations. As such, their scope overlaps to some extent with that of exploratory ordination methods such as PCA. In the terminology of machine learning, clustering algorithms are 'unsupervised' algorithms, i.e. they aim at discovering patterns without being provided any prior information, in contrast to 'supervised' algorithms aiming at classifying patterns based on pre-existing criteria.

The most popular clustering algorithms in ecology are 'hierarchical' ones. They consist in recursively splitting the data into clusters of observations starting from the whole dataset – or conversely, recursively agglomerating clusters of observations starting from the individual observations – by maximizing between-cluster dissimilarity at each step. Dissimilarity between two clusters is most commonly measured as the mean pairwise dissimilarity between the observations of each cluster, a method called UPGMA ('Unweighted Pair Group Method with Arithmetic Mean'). The pairwise dissimilarity between observations can be measured using any dissimilarity metrics, which is often an advantage in ecology owing to the wide range of dissimilarity metrics in use (Legendre & De Caceres, 2013; cf. section II.1.b). Another advantage of hierarchical clustering is that the result can be displayed as a tree of hierarchically nested clusters (or 'dendrogram'): in addition to visualizing data structure, this helps choose the number of clusters according to the desired level of similarity within clusters. Hierarchical clustering is however computationally intensive for large datasets. Moreover, because splits – or merges – decided at each hierarchical step cannot be undone and have a strong impact on the subsequent steps, the algorithm may be easily trapped in suboptimal solutions for large and noisy datasets.

'Partitional' algorithms, which consist in searching for the optimal partition of the data into a predefined number of clusters, form a second family of algorithms that are better adapted to large datasets (Jain, 2010). The most widespread partitional algorithm is *k*-means clustering, which formally consists in finding the *k* clusters that minimize within-cluster variance in the Euclidian space of observations, with *k* a fixed parameter. Unlike hierarchical clustering, which is purely heuristic, the problem of *k*-means clustering can be reframed as the fit of a multivariate statistical model to the data (specifically, a 'Gaussian mixture model'). This is however achieved using heuristic algorithms, which may converge to suboptimal solutions. The most common algorithm consists in randomly setting the position of the *k* cluster centres in the space of observations, delineating the clusters by assigning each observation to the closest cluster centre based on Euclidian distance, and then iteratively reshaping the clusters using their mean in the previous step as their new centre, until convergence. Lastly, 'network science' provides a range of clustering algorithms that are based on a graph representation of the similarity matrix (Rosvall *et al.*, 2009; Fortunato, 2010). These methods that are well adapted to large datasets have recently enjoyed a rise in popularity in ecology (Vilhena & Antonelli, 2015; Bloomfield *et al.*, 2017; Wang *et al.*, 2017).

A pervasive assumption in classical statistical models is that observations are normally distributed – i.e., follow Gaussian probability distributions. This assumption may be explicit, or sometimes implicit. For instance, model fitting by least-square regression amounts to maximizing the log-likelihood of independent identically distributed normal variables centred on the fitted model. Likewise, the assumption in PCA that the observed variables can be entirely characterized by their mean and variance implies that they are normally distributed, since this property is unique to the Gaussian distribution. A justification for the normality assumption is that an observation on a sample can typically be regarded as the sum, or the mean outcome, of many random draws, yet the central limit theorem states that the mean of a sufficiently large number of random variables is always normally distributed. Thank to the many convenient mathematical properties of the Gaussian distribution, exact analytical

expressions have been obtained for maximum-likelihood estimators and for the theoretical distribution of test statistics. Prior to the advent of computers, such analytical results were an essential condition for the practical usefulness of statistical models. This is however not the case anymore, and the exploration of models that are not based on the Gaussian distribution is now possible.

3. The Dirichlet distribution and its Dirichlet process extension

a. The Dirichlet distribution

Not all natural processes are additive, and as a consequence, not all quantities can be assumed to be normally distributed as the sum of a large number of random draws. Some processes are multiplicative, and a direct consequence of the central limit theorem is that the product of a large number of random draws will follow a log-normal distribution. Indeed, for N random variables X_i, $\ln(\prod_{i=1}^{N} X_i) = \sum_{i=1}^{N} \ln X_i$. Hence, the central limit theorem states that $\ln(\prod_{i=1}^{N} X_i)$ is normally distributed for large N. It follows from the definition of the log-normal distribution that $\prod_{i=1}^{N} X_i$ is log-normally distributed. As mentioned in section II.1.a, this is a possible explanation for the often-observed log-normal distribution of species abundances. Indeed, if the abundances of species are independent of each other, a species' change in abundance through time may take the form of a random multiplicative factor applied to its reproductive output at each generation, depending for instance on environmental fluctuations.

However, if changes in species abundance are rather driven by demographic drift, as assumed in a neutral framework, relative species abundances are better described by the following process: starting from abundances $(a_1, \dots, a_S)$, where a_i is the number of individuals in species *i*, one of the S species is picked at each time step with probability equal to its relative abundance (or equivalently, one individual is picked at random in the population), and its abundance is increased by one individual. If this sampling scheme, called a Pólya urn, is repeated indefinitely, the distribution of

species relative abundances $(x_1, \dots, x_S)$ will follow the Dirichlet distribution of parameters $(a_1, \dots, a_S)$, which may be regarded as a distribution over distributions (Blackwell & MacQueen, 1973):

$$\begin{cases} p(x_1, \dots, x_{S-1} | a_1, \dots, a_S) = \dfrac{\Gamma\left(\sum_{i=1}^{S} a_i\right)}{\prod_{i=1}^{S} \Gamma(a_i)} \prod_{i=1}^{S} x_i^{a_i - 1} \\ x_S = 1 - \sum_{i=1}^{S-1} x_i \end{cases}$$

Γ is the gamma function generalizing the factorial to real numbers and taking value $\Gamma(a) = (a-1)!$ when a is a positive integer. Note that the description of the Pólya urn originally involves drawing balls of different colours from an urn instead of individuals of different species from a community.

If there is no a priori reason to assume differences between the S species, parsimony leads to setting all initial abundances a_i to the same value a ('symmetric' Dirichlet distribution). In that case, some species will randomly emerge as more abundant than others over time in the Pólya urn sampling scheme, since any above-average abundance tends to be amplified. The shape of the limiting distribution after an infinite number of time steps is heavily influenced by the 'concentration parameter' a, which can formally take any positive real value. If a is much smaller than 1, the first species to be picked by the sampling scheme will have its abundance updated to $a + 1$, and will have a disproportionately higher probability to be picked again at the next time step. Conversely, if a is much larger than 1, the fact that a species' abundance is increased by 1 has little influence on its subsequent probability to be picked. Thus, depending on the value of a relative to 1, the symmetric Dirichlet distribution can either describe a species abundance distribution with a few dominant species and many rare one ($a \ll 1$), reminiscent of the structure observed in species-rich communities, or in contrast a very even species abundance distribution ($a \gg 1$). In the general case, any set of parameters $(a_1, \dots, a_S)$ can be rewritten as $(\theta p_1, \dots, \theta p_S)$, with $\sum_{i=1}^{S} a_i = \theta$ and $p_i = a_i / \theta$, so that $\sum_{i=1}^{S} p_i = 1$. The Dirichlet distribution with asymmetric parameters behaves similarly to the symmetric case, except that the relative abundance x_i of

species i has mean p_i over all possible draws from the Dirichlet distribution, while its variance is determined by the value of θ/S.

The symmetric Dirichlet distribution is the distribution that Fisher (1943) implicitly assumed for relative species abundances to derive the log-series SAD, defined as $\mathbb{E}[\Phi_n] = \alpha x^n/n$ (cf. section II.1.a). He assumed that the number of sampled individuals per species followed a negative-binomial distribution of parameters $(\alpha/S\,, x)$ (without the zero-abundance class, because the latter cannot be observed), as the result of Poisson sampling from a large number S of Gamma-distributed species abundances with shape parameter α/S and rate parameter $(1-x)/x$. The negative-binomial distribution P_{NB} can indeed be obtained as $P_{NB}(k|\,\alpha/S\,, x) = \int_0^\infty P_P(k|\lambda) p_G(\lambda|\,\alpha/S\,, (1-x)/x)\,d\lambda$, where P_P and p_G denote the Poisson and Gamma distributions. Yet, if S species have abundances n_i identically distributed as $\text{Gamma}(\alpha/S\,, \theta)$, their relative abundances n_i/N, where $N = \sum_{i=1}^{S} n_i$, follow a symmetric Dirichlet distribution with concentration parameter α/S (Devroye, 1986). Since Fisher assumed $\alpha/S \ll 1$ to obtain the log-series, this indeed corresponds to the regime of very uneven relative species abundances.

b. The Dirichlet process and the Ewens sampling formula

As it is apparent in the case of Fisher's log-series, a limitation of the Dirichlet distribution as a mean to describe species relative abundances is that it requires the number S of species to be fixed in advance. It is hence appealing to generalize the Dirichlet distribution by making S tend toward infinity. Let us consider the time step $N+1$ of the Pólya urn sampling scheme with symmetric concentration parameter a, where N individuals have already been added to the original Sa individuals. The probability to pick species i is $(n_i + a)/(N + Sa)$, where n_i is the number of times species i has already been picked. Hence, the probability to pick one of the S_N species that have already been picked at least once is $(N + S_N a)/(N + Sa)$, while the probability to pick one of the $S - S_N$ species that have never been picked is

$(S - S_N)a/(N + Sa)$. If we simultaneously make S tend toward infinity and a tend toward 0, keeping the product Sa equal to a constant θ, we obtain an infinite-dimensional version of the Pólya urn, called the Hoppe urn, where the probability to pick an existing species i at time step $N + 1$ is $n_i/(N + \theta)$ and the probability to pick a new species is $\theta/(N + \theta)$ (Hoppe, 1984). After an infinite number of time steps, species relative abundances are distributed according to a Dirichlet process of concentration parameter θ and uniform base distribution, which can be regarded as the limit of the S-dimensional symmetric Dirichlet distribution of concentration parameter θ/S when S tends toward infinity (Ferguson, 1973; Teh *et al.*, 2006).

More generally, a Dirichlet process of concentration parameter θ and base distribution $\boldsymbol{p} = (p_i)_{i \in \mathbb{N}^*}$ can be regarded as the limit of the S-dimensional Dirichlet distribution of concentration parameters $(\theta p_1, \dots, \theta p_S)$, where $\sum_{i=1}^{S} p_i = 1$, when S tends toward infinity (Ferguson, 1973). The infinite base distribution $\boldsymbol{p}$ is the distribution from which new species are sampled during the Hoppe urn scheme of parameter θ: each new species is sampled from an infinite number of possible species labels with probability weights $\boldsymbol{p}$. If the base distribution is uniform, as assumed in the previous paragraph, a never-encountered label is simply assigned to each new species.

The Dirichlet process is most intuitively understood by sampling from it. If N individuals are sampled from relative species abundances described by a Dirichlet process of parameter θ and uniform base distribution, their partition $(\Phi_1, \dots, \Phi_N)$ into S species, where Φ_n is the number of species with abundance n, obeys the 'Ewens sampling formula' of parameters (θ, N) (Ewens, 1972):

$$P(\Phi_1, \dots, \Phi_N | \theta, N) = \frac{N!}{(\theta)_N} \prod_{n=1}^{N} \frac{1}{\Phi_n!} \left(\frac{\theta}{n}\right)^{\Phi_n}$$

where $(\theta)_N = \Gamma(\theta + N)/\Gamma(\theta)$. This formula also describes the partition of N individuals into S species obtained by stopping a Hoppe urn scheme of parameter θ at step N, thus the Dirichlet process does not need to be explicitly defined for the Ewens formula to emerge from the Hoppe urn scheme. For a large enough sample, the Φ_n are approximately drawn from independent Poisson random variables with parameter θ/n

(Crane, 2016). A remarkable property of the Ewens formula is that it yields a sampling-invariant description of relative species abundances characterized by the parameter θ: indeed, any random subsample of $N_1 < N$ individuals taken from the initial sample obeys the Ewens sampling formula of parameters (θ, N_1). Moreover, the probability of observing S species in a sample of N individuals does not depend on the exact partition but only on θ and N, as $P(S|\theta, N) = s(N, S)\,\theta^S/(\theta)_N$, where the function s denotes the absolute value of the Stirling numbers of the first kind (Ewens, 1972). Thus, θ can be regarded as a sampling-invariant measure of diversity in a species pool described by Ewens sampling formula, irrespective of whether this species pool is finite or infinite.

4. Neutral models

The Ewens formula was first discovered by Ewens (1972) in the context of population genetics. Indeed, it arises as the stationary distribution of allele frequency in the Wright-Fisher and Moran models, which describe the neutral dynamics of alleles in a population (Fisher, 1930; Wright, 1931; Moran, 1958; Wakeley, 2009). More importantly for ecologists, the Ewens formula is also the stationary distribution of species frequency in Hubbell's neutral model of biodiversity, which was directly inspired by population genetics (Hubbell, 2001). These models all bear some resemblance to the Hoppe urn sampling scheme, except that they account for the death of individuals, so that the total number of individuals remains constant over time. The Wright-Fisher model assumes that all N individuals die at each time step and are replaced by a new generation of N new individuals. The alleles of these new individuals are sampled (with replacement) from the alleles in the previous generation, except for a small probability in each new individual of mutating into a never-encountered allele. This translates into a demographic drift of allele frequency through time, and, over longer time scales, by a turnover in the pool of alleles through random mutation and extinction events. The Moran model is similar but assumes than individuals die and are replaced one at a time, which allows for overlapping generations. Hubbell's model is almost identical to the Moran model, except that alleles are reinterpreted as species and

mutation as speciation, and that dying individuals cannot be replaced by their own offspring. Even though all three models have the same stationary abundance distribution described by Ewens formula, the exact expression of θ depends on the model's dynamics (Etienne & Alonso, 2007). In Hubbell's model, $\theta = (N-1)\,\nu/(1-\nu)$, where ν is the speciation probability at each time step, i.e. the probability that the dying individual is replaced by a new species.

The key innovation of Hubbell's model compared to population genetics models is that it also includes the description of a dispersal-limited local community connected to the regional metacommunity through immigration. The dynamics of the local community is identical to that of the metacommunity, except that new species arise through immigration instead of speciation: at each time step, an individual dies and there is probability m that the replacing individual results from immigration from the metacommunity instead of from local reproduction (if $m = 1$, there is no limitation to dispersal). The difference is that unlike individuals arising through speciation, an immigrating individual may belong to a species that is already present in the local community. Thus, the stationary distribution of species frequency in a local community of size N obeys a Ewens sampling formula of parameter $I = (N-1)\,m/(1-m)$, modified to account for the fact that the immigrating ancestors to the current local community are sampled from the Ewens formula of parameter θ (Etienne & Olff, 2004). The resulting two-layer sampling formula was derived by Etienne (2005). The 'Etienne sampling formula' can also be regarded as the result of 'dispersal-limited sampling' from Ewens formula, which can be defined as a type of skewed sampling (Etienne & Alonso, 2005). Importantly, Etienne formula still satisfies the sampling-invariance property of Ewens formula, i.e. any random subsample of $N_1 < N$ individuals will follow the Etienne formula of parameters (θ, I, N_1).

Ewens and Etienne sampling formula allow for likelihood-based inference of the neutral parameters θ and I, as well as for rigorous statistical tests of model fit (cf. Fig. 9; Etienne & Olff, 2005; Etienne, 2007; Al Hammal *et al.*, 2015). In practice, the metacommunity cannot be directly observed and is usually regarded as infinite, while the local community is equated with the observed sample of individuals. Thus, θ and m

are often chosen as model parameters instead of θ and I, reflecting the fact that the number of individuals is known in the local community but not in the metacommunity. I can be interpreted as the effective number of individuals in the metacommunity that are in direct competition with the local community for reproduction. Furthermore, data usually consist of samples from several local communities. This considerably increases statistical power, since statistical inference does not only rely on the shape of the local abundance distributions, but also on the taxonomic overlap between local communities. Exact sampling formulas have been derived both for the case where all local communities have the same immigration parameter m (Etienne, 2007) and for the case where they do not (Etienne, 2009).

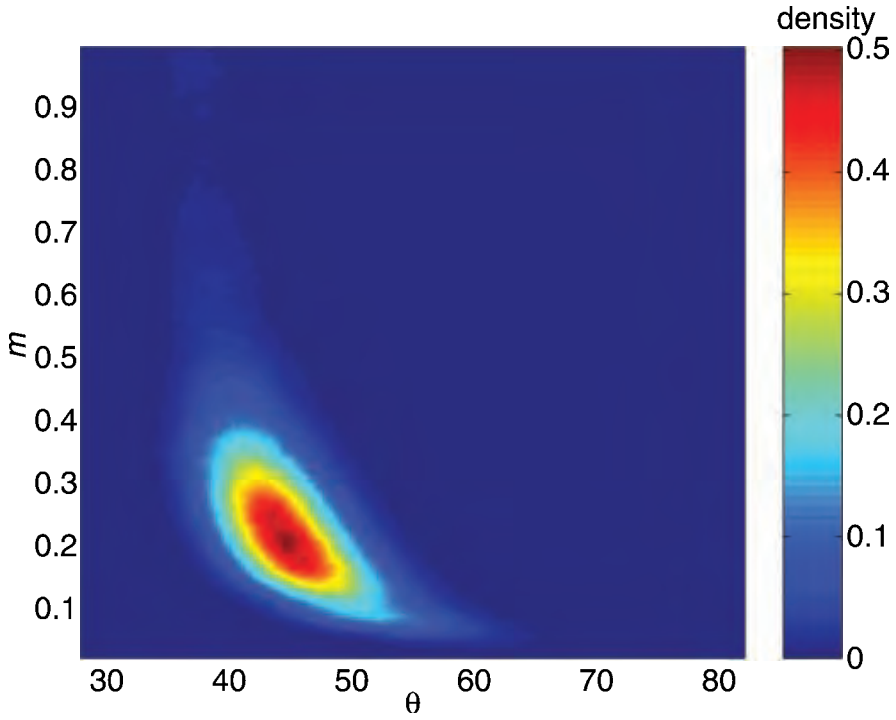

Figure 9. Bayesian inference of neutral parameters based on the 'Etienne sampling formula': map showing the joint posterior probability density of θ and m for the tree abundance data (>10 cm dbh) of the 50-ha Barro Colorado Island monitored plot. Adapted from Etienne & Olff (2004).

Despite the interest of exact sampling formulas for statistical inference, approximate approaches have proved more practical in some instances. When the number of samples is large enough, the metacommunity composition may be simply approximated as the sum of all samples, instead of being explicitly modelled. In so doing, one can avoid making any assumption on the metacommunity when estimating immigration rates, or when testing the assumption of dispersal-limited neutral community assembly (Sloan *et al.*, 2006; Jabot *et al.*, 2008; Harris *et al.*, 2015). Furthermore, when abundance data are unavailable or unreliable, the immigration rate from the metacommunity may be estimated solely based on the occurrence of species across samples (Sloan *et al.*, 2006). A limitation of exact sampling formulas is that their computation is numerically demanding when the number of individuals becomes large. An alternative approach is then to represent the sample as continuous species relative abundances rather than in a fully discrete way. The species relative abundances $(x_1, \dots, x_S)$ in a large dispersal-limited sample containing S species may be approximated as following the Dirichlet distribution of parameters $(Ip_1, \dots, Ip_S)$, where $(p_1, \dots, p_S)$ are the relative abundances of those S species in the metacommunity (Sloan *et al.*, 2007). In turn, $(p_1, \dots, p_S)$ can be approximated as following the symmetric Dirichlet distribution of parameter θ/S (Woodcock *et al.*, 2007). As is apparent from section III.b, these continuous approximations may be extended to the case of an infinite number of species S by modelling the relative abundances $\boldsymbol{x}$ in the local community as a Dirichlet process of parameter I and base distribution $\boldsymbol{p} = (p_i)_{i \in \mathbb{N}^*}$, and the relative abundances $\boldsymbol{p}$ in the metacommunity as a Dirichlet process of parameter θ and uniform base distribution (Harris *et al.*, 2015). Such a model is referred to as a 'hierarchical Dirichlet process' in the language of machine learning.

While multivariate likelihood expressions are powerful tools for statistical inference, they are difficult to visualize, and one-dimensional SADs may be better suited for intuitively understanding the model's behaviour. For instance, the non dispersal-limited SAD in Hubbell's model is equal to (Moran, 1958; Vallade & Houchmandzadeh, 2003):

$$\mathbb{E}[\Phi_n|\theta, N] = \frac{\theta}{n}\frac{(N+1-n)_n}{(N+\theta-n)_n}$$

and converges toward Fisher's log-series with $\theta = \alpha$ for a large enough number N of individuals (Chave, 2004). In general, the SAD can be regarded as the first moment of the multivariate sampling formula, since it is obtained as:

$$\mathbb{E}[\Phi_n|\theta, N] = \sum\nolimits_{\{\Phi_1,\ldots,\Phi_N|\sum_{i=1}^{N} i\Phi_i = N\}} \Phi_n P(\Phi_1, \ldots, \Phi_N|\theta, N)$$

A more straightforward approach to deriving this quantity is to express Hubbell's dynamical model through the approximate conditional transition probabilities $P(n+1|n,\theta)$, $P(n-1|n,\theta)$ and $P(n|n,\theta)$ that a given species with current abundance n will have abundances $n+1$, $n-1$, or n at the next time step, respectively. The stationary probability distribution of this 'master equation' then provides an estimate of $\mathbb{E}[\Phi_n|\theta, N]$, once multiplied by the observed number S of species in the sample (Volkov *et al.*, 2003; Alonso & McKane, 2004; McKane *et al.*, 2004; O'Dwyer *et al.*, 2009). Unlike the exact 'genealogical' approach described above, this approach typical of statistical physics does not explicitly account for the interdependence between species, induced by the constraint of a fixed total number of individuals through time ('mean field' approach). While this constraint was originally deemed a key element of the model since it accounts for competition between species (Hubbell, 2001), both the genealogical and the master equation approaches have been found to yield the same SAD expression for a large enough sample (Etienne *et al.*, 2007).

5. Categorical mixture models

Let us assume that the relative abundances $\boldsymbol{x} = (x_1, \ldots, x_S)$ of S species, with $\sum_{i=1}^{S} x_i = 1$, follow a Dirichlet distribution of parameters $\boldsymbol{a} = (a_1, \ldots, a_S)$. The categorical distribution describes the choice of one out of S species (or categories) with probability weights $\boldsymbol{x}$. It can be regarded as a special case of the multinomial distribution, defined as $P(\boldsymbol{n}|N, \boldsymbol{x}) = N!/(n_1! \ldots n_N!)\, x_1^{n_1} \ldots x_S^{n_S}$, which describes more generally the outcome

of N successive categorical draws with probability weights $\boldsymbol{x}$. A remarkable property of the Dirichlet distribution is that it is the conjugate prior of the categorical and multinomial distributions. Namely, if a multinomial sample $\boldsymbol{n} = (n_1, \dots, n_S)$ is observed from $\boldsymbol{x}$, with $\sum_{i=1}^{S} n_i = N$, then the posterior distribution of $\boldsymbol{x}$ given the observations $\boldsymbol{n}$ still follows a Dirichlet distribution, but with parameters updated to $\boldsymbol{a} + \boldsymbol{n} = (a_1 + n_1, \dots, a_S + n_S)$ to account for the observations. Fundamentally, this means that the Dirichlet distribution is the "*natural distribution occurring when the probability that a forthcoming observation is of certain class only depends on the number of times this class has already been observed and on the total number of observations made so far*" (Crane, 2016).

The posterior distribution of $\boldsymbol{x}$ given the observations $\boldsymbol{n}$ is defined as $p(\boldsymbol{x}|\boldsymbol{n}, N, \boldsymbol{a}) = P(\boldsymbol{n}|N, \boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{a})/P(\boldsymbol{n}|N, \boldsymbol{a})$. The marginal likelihood $P(\boldsymbol{n}|N, \boldsymbol{a}) = \int_{\boldsymbol{x}} P(\boldsymbol{n}|N, \boldsymbol{x})p(\boldsymbol{x}|\boldsymbol{a})\, d\boldsymbol{x}$ is the 'Dirichlet-multinomial' distribution of parameters $(N, \boldsymbol{a})$, i.e. the distribution of a N-individual multinomial sample with Dirichlet-distributed probability weights of parameters $\boldsymbol{a}$. The Dirichlet-multinomial distribution can be regarded as a finite-dimensional version of the Ewens formula (Crane, 2016).

Because the Dirichlet distribution is the conjugate prior of the categorical and multinomial distributions, it is the natural prior in any probabilistic model involving categorical or multinomial sampling from discrete classes. This is the case of 'categorical mixture models', which describe observations as sampled from a mixture of K classes, with 'mixture weights' $\boldsymbol{\theta_k} = (\theta_k)_{k \in [\![1,K]\!]}$, verifying $\sum_{k=1}^{K} \theta_k = 1$. The different classes are typically not directly observable: instead, each is characterized by a probability distribution of parameters $\boldsymbol{\phi^k}$ from which all the observations assigned to class k are sampled. The probability distribution associated with each class may be for instance Gaussian if observations are continuous, or categorical if observations are discrete. If the goal of statistical inference is to capture data structure, the focus will be on estimating the mixture weights $\boldsymbol{\theta_k}$ of the different classes given the observations, as well as the parameters $\boldsymbol{\phi^k}$ of the probability distribution associated with each class. If the goal is to cluster the observations (or to classify them, if inference is conducted in a supervised way), the focus will be on assigning to each observation its most likely class

k. In a fully Bayesian setting, the parameters $\boldsymbol{\phi^k}$ of the probability distribution associated with class *k* may also be given a prior distribution, such as a Dirichlet prior in the case of categorical observations. In this latter case, the model may be referred to as a 'Dirichlet mixture model', since it describes observations as categorical or multinomial samples from a mixture of Dirichlet-distributed classes.

This family of models has found applications in many fields. In particular, Holmes *et al.* (2012) applied such a model to investigate the structure of microbial communities sampled by environmental DNA sequencing. They assume that each sample is a local community belonging to one of *K* possible classes, which they interpret as 'metacommunities'. To each of these metacommunities is assigned a mixture weight θ_k, which is the probability for a sample to originate from it. A metacommunity *k* is defined by probability weights $\boldsymbol{\phi^k} = \left(\phi_i^k\right)_{i\in[\![1,S]\!]}$ over the *S* OTUs observed in the dataset, from which local OTU abundances are sampled. These probability weights are themselves Dirichlet-distributed with parameters $\boldsymbol{a^k} = \left(a_i^k\right)_{i\in[\![1,S]\!]}$. For practical purposes, the parameters a_i^k may be further assumed to follow a 'hyperprior' distribution parameterized by 'hyperparameters', so as to reduce the number of fixed parameters to estimate.

A version of this model was introduced earlier in population genetics, and implemented in the software Structure (Pritchard *et al.*, 2000). In the context of population genetics, each sample is an individual, each class is a population, and observations consist in the alleles found at a number of loci in each individual. As a consequence, the model exhibits a few minor differences compared to that of Holmes *et al.* (2012). Since there are *L* observed loci per individual, each class *k* is not defined by one distribution, but by *L* distributions $\boldsymbol{\phi^{k,l}} = \left(\phi_i^{k,l}\right)_{i\in[\![1,S_l]\!]}$ over the S_l possible alleles at locus *l*, each of these distributions having Dirichlet prior. Moreover, only one categorical draw from $\boldsymbol{\phi^{k,l}}$ is observed at each locus in each individual, instead of a multinomial sample.

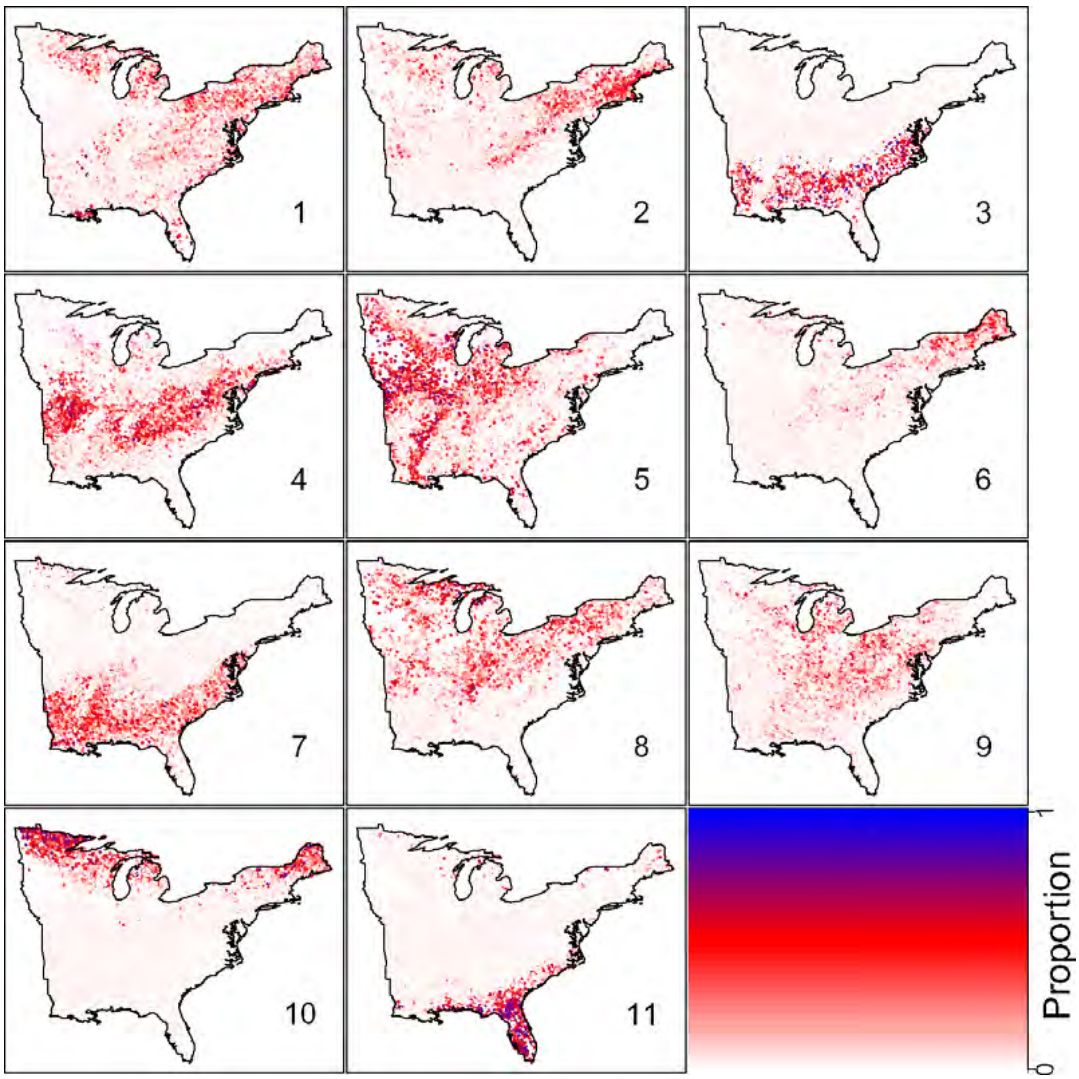

Figure 10. Valle *et al.* (2014) applied Latent Dirichlet Allocation to identify forest tree assemblages in the Eastern United States, based on tree census data from 34,174 forest plots. Maps show the relative proportion of each of the $K = 11$ LDA classes in each forest plot. Adapted from Valle *et al.* (2014).

In the same paper, Pritchard *et al.* (2000) proposed a second slightly more sophisticated model, which includes the possibility of admixture between populations. This is achieved by relaxing the assumption that each individual m originates from a single population, and by assuming instead that it originates from a mixture of K populations with individual-specific weights $\boldsymbol{\theta^m} = (\theta_k^m)_{k\in[\![1,K]\!]}$. As in the model without admixture, the K populations are each defined by a single set of L distributions $\boldsymbol{\phi^{k,l}}$ across the dataset. Thus, each observed allele is the result of a categorical draw from the

individual-specific weights $\boldsymbol{\theta^m}$, followed by a second categorical draw from the population-specific and locus-specific weights $\boldsymbol{\phi^{k,l}}$. Another version of this model with admixture was independently proposed under the name 'Latent Dirichlet Allocation' (LDA) by Blei *et al.* (2003) in the field of natural language processing, a subfield of machine learning, to address the problem of 'topic modelling'. In this context, the aim of the model is to decompose text documents into topics based on their word content. Each class or topic k is defined by its distribution $\boldsymbol{\phi^k} = \left(\phi_i^k\right)_{i \in [\![1,S]\!]}$ over the S distinct words observed in the whole text corpus, and each document m is a mixture, with document-specific weights $\boldsymbol{\theta^m}$, of multinomial samples from these distributions.

This model with admixture has proved very successful and has been subsequently extended, both in its population genetics version (Falush *et al.*, 2003, 2007; Hubisz *et al.*, 2009) and in its topic modelling version (Griffiths & Steyvers, 2004; Rosen-Zvi *et al.*, 2004; Teh *et al.*, 2006; Blei, 2012). The latter (LDA) has been applied to a wide range of domains pertaining to machine learning where its ability to handle large and complex datasets has been praised, including satellite image processing (Vaduva *et al.*, 2013), bioinformatics (Liu *et al.*, 2010), fraud detection in telecommunications (Olszewski, 2012) and social sciences (Mauch *et al.*, 2015). In particular, it has been recently applied to spatially and temporally explicit forest tree composition data in ecology, where its ability to decompose samples into classes learnt over the whole dataset allows for capturing smooth spatial and temporal gradients across the samples (cf. Fig. 10; Valle *et al.*, 2014). Related models have also been applied to the detection of different source environments in microbial community samples, with a focus on supervised inference: Knights *et al.* (2011) applied this approach to the detection of contamination in a medical environment, while Shafiei *et al.* (2015) proposed a more sophisticated two-layer model, where each class is itself a mixture of higher-level classes.

As in the case of neutral models, a limitation of Dirichlet-multinomial models is that the number of classes must be specified in advance. A number of methods have been used to help select the number of classes (Airoldi *et al.*, 2010). Nevertheless, the most rigorous approach is to design a model with a potentially infinite number of

classes, an approach referred to as 'nonparametric Bayesian', since the size of the model is not fixed in advance by a parameter. This can be achieved by setting a Dirichlet process prior over the mixture weights, since the Dirichlet process is, like the Dirichlet distribution, conjugate to the categorical and multinomial distributions (Crane, 2016). This amounts to making the number K of classes tend toward infinity.

In the infinite-dimensional extension of the model without admixture, the mixture weights $\boldsymbol{\theta} = (\theta_k)_{k\in\mathbb{N}^*}$ over classes follow a Dirichlet process of uniform base distribution over class labels, while each class k is defined as in the finite-dimensional case by its distribution $\boldsymbol{\phi^k} = \left(\phi_i^k\right)_{i\in[\![1,S]\!]}$ over the S possible observations (Teh *et al.*, 2006). In the model with admixture however, a hierarchical Dirichlet process needs to be defined. Indeed, if an independent Dirichlet process of uniform base distribution were to be assigned in each sample m as a prior to the mixture weights $\boldsymbol{\theta^m} = (\theta_k^m)_{k\in\mathbb{N}^*}$, two documents would not have any class in common. Thus, in the infinite-dimensional extension of the model with admixture, the mixture weights $\boldsymbol{\theta^m}$ in each sample m originate from a Dirichlet process of base distribution $\boldsymbol{\beta}$ over classes, while the distribution $\boldsymbol{\beta}$ follows itself a Dirichlet process of uniform base distribution over class labels (Teh *et al.*, 2006). Likewise, two local communities in the infinite-dimensional approximation of Hubbell's neutral model would not have any species in common if not for the hierarchical Dirichlet process construction (cf. section III.4).

IV. Objectives and outline

1. Objectives

Most of Earth's biodiversity is concentrated in a few hyperdiverse ecosystems, such as tropical forests. Yet, the mechanisms that permit the coexistence of such a large number of species are not fully understood. In particular, the relative influence of deterministic niche processes and stochastic dispersal limitation has long been debated. One approach to address this question is through the study of integrative biodiversity patterns, such as the distribution of species abundances and the turnover of species composition through space. At a time when human activities threaten both biodiversity and the associated ecosystems, a better understanding of these patterns and of the underlying mechanisms is much needed.

A major obstacle lies in the difficulty to measure biodiversity. Indeed, it has long relied on direct human observation. However, recent technological advances now make automated data collection possible, which could alleviate this problem. Environmental DNA sequencing is especially promising for improving our understanding of biodiversity patterns. Indeed, it eases and standardizes the measurement of biodiversity, increases the amount of available data by orders of magnitude, and dramatically expands the range of accessible taxa. In particular, it allows for taking into account microbial diversity, arguably the 'hidden part of the biodiversity iceberg'.

Nevertheless, taking advantage of this new type of data is challenging. First, the range of information types that can be collected is restricted, in that no complementary measurements, such as size for instance, can be made on organisms. In most cases, even taxonomic information is relatively imprecise owing to the lack of reference database for the retrieved DNA sequences. Thus, inference is mostly based on patterns of unidentified OTUs. Second, because observations are indirect and noisy, their interpretation is not as straightforward as in the case of direct censuses of individual

organisms. Third, the high diversity of microbial communities makes for large and sparse datasets, to which existing statistical approaches are not well suited.

The overarching goal of this thesis was to investigate how environmental DNA sequencing, and more generally the automated collection of ecological data, could contribute to our understanding of biodiversity patterns and of their underlying mechanisms. This work was motivated by two observations. First, theoretical models in ecology are for the most part not oriented toward comparison with data, and when they are, as in the case of Hubbell's neutral model, they are centred on individual organisms, which hampers their comparison to environmental DNA data. Second, existing statistical methods in ecology have limitations in their ability to tackle such data. Thus, this work has an important methodological component. A second goal of this thesis was to apply the developed approaches to soil DNA data collected in the forests of French Guiana, so as to better understand community assembly in tropical forests. This includes a dataset that was collected as part of this thesis.

2. Outline

The first chapter addresses the issue of measuring beta diversity patterns from environmental DNA data, and of using these patterns to disentangle dispersal-limited and niche-based processes across the different domains of life. To this end, a soil DNA dataset was collected in French Guiana, in forest plots that are approximately regularly spaced on a logarithmic scale. A range of soil properties was also measured from the soil samples. Three approaches are compared: distance-based analyses using dissimilarity metrics, raw-data analyses using multivariate ordination, and fitting the neutral prediction for the decay of taxonomic similarity with distance. These approaches are typical of those used to analyse classical biodiversity census data. In addition, the effect on human disturbance through logging is assessed, based on a more limited number of plots presenting a gradient of logging intensities.

The second chapter focuses on species abundance distributions measured from environmental DNA data, and addresses the problem of comparing this pattern to the prediction of Hubbell's neutral model. Indeed, it was unknown to what extent this pattern may remain informative in spite of the potential noise. Simulation results are presented, that quantify how the estimates of the neutral diversity and dispersal parameters are biased when inferred from environmental DNA data. A benchmark dataset of limited extent is used to assess the level of noise that is to be expected in real data.

Like the first chapter, the third chapter discusses spatial patterns in environmental DNA data, but it proposes an approach differing from those classically followed in ecology. It investigates the potential of a model-based statistical method, Latent Dirichlet Allocation, to decompose the data into assemblages of spatially co-occurring OTUs. In addition, a method is proposed to measure the stability of the decomposition. The approach is tested through simulations, and by applying it to a large soil DNA dataset. This dataset follows a regular spatial sampling scheme over a forest plot, and was collected in French Guiana before the start of this thesis. The insights on soil community structure provided by the approach are discussed, making use of Lidar measurements of environmental features.

Finally, the discussion provides a synthesis of the results, and discusses the perspectives arising from this thesis.

References

Airoldi, E.M., Erosheva, E.A., Fienberg, S.E., Joutard, C., Love, T. & Shringarpure, S. (2010) Reconceptualizing the classification of PNAS articles. *PNAS*, **107**, 20899–20904.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, 716–723.

Al Hammal, O., Alonso, D., Etienne, R.S. & Cornell, S.J. (2015) When Can Species Abundance Data Reveal Non-neutrality? *Plos Computational Biology*, **11**, 23.

Alonso, D., Etienne, R.S. & McKane, A.J. (2006) The merits of neutral theory. *Trends in Ecology & Evolution*, **21**, 451–457.

Alonso, D. & McKane, A.J. (2004) Sampling Hubbell's neutral theory of biodiversity. *Ecology Letters*, **7**, 901–910.

Alonzo, M., Bookhagen, B. & Roberts, D.A. (2014) Urban tree species mapping using hyperspectral and lidar data fusion. *Remote Sensing of Environment*, **148**, 70–83.

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K.H., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Aristotle (IVth cent. BC) History of Animals.

Armstrong, R.A. & McGehee, R. (1980) Competitive exclusion. *The American Naturalist*, **115**, 151–170.

Arnoldi, J.-F., Loreau, M. & Haegeman, B. (2016) Resilience, reactivity and variability: A mathematical comparison of ecological stability measures. *Journal of Theoretical Biology*, **389**, 47–59.

Arrhenius, O. (1921) Species and area. *Journal of Ecology*, **9**, 95–99.

Baas Becking, L.G.M. (1934) *Geobiologie of inleiding tot de milieukunde.*, W.P. Van Stockum & Zoon, The Hague, the Netherlands.

Bayes, M. & Price, M. (1763) An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs. *Philosophical Transactions (1683-1775)*, 370–418.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L. & Bignell, H.R. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, **456**, 53.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer. Michael Jordan, Jon Kleinberg, Bernhard Schölkopf.

Blackwell, D. & MacQueen, J.B. (1973) Ferguson distributions via Pólya urn schemes. *The annals of statistics*, 353–355.

Blei, D. (2012) Probabilstic Topic Models. *Communication of the Association for Computing Machinery*, **55**, 77–84.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Bloomfield, N.J., Knerr, N. & Encinas-Viso, F. (2017) A comparison of network and clustering methods to detect biogeographical regions. *Ecography*.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M., Yu, D.W. & de Bruyn, M. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, **29**, 358–367.

Braun-Blanquet & Pavillard (1922) Vocabulaire de sociologie végétale.

Brook, B.W., Ellis, E.C., Perring, M.P., Mackay, A.W. & Blomqvist, L. (2013) Does the terrestrial biosphere have planetary tipping points? *Trends in Ecology & Evolution*, **28**, 396–401.

Brown, J.H. (1995) *Macroecology*, University of Chicago Press.

Brown, J.H. (1978) The theory of insular biogeography and the distribution of boreal birds and mammals. *Great Basin Naturalist Memoirs*, 209–227.

Brown, J.H. & Kodric-Brown, A. (1977) Turnover Rates in Insular Biogeography: Effect of Immigration on Extinction. *Ecology*, **58**, 445–449.

Brown, W.L. & Wilson, E.O. (1956) Character displacement. *Systematic zoology*, **5**, 49–64.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference*, Springer, New York.

Cabeza, M. & Moilanen, A. (2001) Design of reserve networks and the persistence of biodiversity. *Trends in Ecology & Evolution*, **16**, 242–248.

Carpenter, S.R., Cole, J.J., Pace, M.L., Batt, R., Brock, W.A., Cline, T., Coloso, J., Hodgson, J.R., Kitchell, J.F., Seekell, D.A., Smith, L. & Weidel, B. (2011) Early Warnings of Regime Shifts: A Whole-Ecosystem Experiment. *Science*, **332**, 1079.

Caswell, H. (1976) Community structure - neutral model analysis. *Ecological Monographs*, **46**, 327–354.

Chase, J.M. & Leibold, M.A. (2003) *Ecological niches: linking classical and contemporary approaches*, University of Chicago Press.

Chave, J. (2004) Neutral theory and community ecology. *Ecology Letters*, **7**, 241–253.

Chave, J. (2013) The problem of pattern and scale in ecology: what have we learned in 20years? *Ecology Letters*, **16**, 4–16.

Chave, J., Alonso, D. & Etienne, R.S. (2006) Theoretical biology - Comparing models of species abundance. *Nature*, **441**, E1–E1.

Chave, J. & Leigh, E.G. (2002) A spatially explicit neutral model of beta-diversity in tropical forests. *Theoretical Population Biology*, **62**, 153–168.

Chave, J., Muller-Landau, H.C. & Levin, S.A. (2002) Comparing classical community models: Theoretical consequences for patterns of diversity. *American Naturalist*, **159**, 1–23.

Chesson, P. (2000) Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, **31**, 343–+.

Chisholm, R.A. & Pacala, S.W. (2010) Niche and neutral models predict asymptotically equivalent species abundance distributions in high-diversity ecological communities. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15821–15825.

Clements, F.E. (1916) *Plant succession: an analysis of the development of vegetation*, Carnegie Institution of Washington.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E. & Hubbell, S.P. (2002) Beta-diversity

in tropical forest trees. *Science*, **295**, 666–669.

Connell, J.H. (1983) On the prevalence and relative importance of interspecific competition: evidence from field experiments. *The American Naturalist*, **122**, 661–696.

Connell, J.H. (1970) On the role of natural enemies in preventing competitive exclusion in some marine animals and in rain forest trees. *Dynamics of populations*.

Connolly, S.R., Hughes, T.P. & Bellwood, D.R. (2017) A unified model explains commonness and rarity on coral reefs. *Ecology Letters*.

Cordero, O.X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Le Roux, F., Mincer, T. & Polz, M.F. (2012) Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science*, **337**, 1228.

Cox, C.B., Moore, P.D. & Ladle, R. (2016) *Biogeography: an ecological and evolutionary approach*, John Wiley & Sons.

Crane, H. (2016) The Ubiquitous Ewens Sampling Formula. *Statistical Science*, **31**, 1–19.

Curtis, T.P. & Sloan, W.T. (2005) Exploring microbial diversity - A vast below. *Science*, **309**, 1331–1333.

Daily, G. (1997) *Nature's services: societal dependence on natural ecosystems*, Island Press.

Darwin, C. (1859) On the Origin of Species by Means of Natural Selection.

Davies, T.J., Allen, A.P., Borda-de-Agua, L., Regetz, J. & Melian, C.J. (2011) Neutral biodiversity theory can explain the imbalance of phylogenetic trees but not the tempo of their diversification. *Evolution*.

Devroye, L. (1986) *Sample-based non-uniform random variate generation*. *Proceedings of the 18th conference on Winter simulation*, pp. 260–265. ACM.

Diamond, J.M. (1975) *Assembly of species communities*. *Ecology and Evolution of Communities*, pp. 342–444. Cody, M.L. & Diamond, J.M., Cambridge, MA.

Doughty, C.E., Wolf, A. & Malhi, Y. (2013) The legacy of the Pleistocene megafauna extinctions on nutrient availability in Amazonia. *Nature Geoscience*, **6**, 761–764.

Du, X., Zhou, S. & Etienne, R.S. (2011) Negative density dependence can offset the effect of species competitive asymmetry: A niche-based mechanism for neutral-like patterns. *Journal of Theoretical Biology*, **278**, 127–134.

Etienne, R.S. (2007) A neutral sampling formula for multiple samples and an "exact" test of neutrality. *Ecology Letters*, **10**, 608–618.

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Etienne, R.S. (2009) Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Journal of Theoretical Biology*, **257**, 510–514.

Etienne, R.S. & Alonso, D. (2005) A dispersal-limited sampling theory for species and alleles. *Ecology Letters*, **8**, 1147–1156.

Etienne, R.S. & Alonso, D. (2007) Neutral community theory: How stochasticity and dispersal-limitation can explain species coexistence. *Journal of Statistical Physics*, **128**, 485–510.

Etienne, R.S., Alonso, D. & McKane, A.J. (2007) The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, **248**, 522–536.

Etienne, R.S. & Olff, H. (2004) A novel genealogical approach to neutral biodiversity theory. *Ecology Letters*, **7**, 170–175.

Etienne, R.S. & Olff, H. (2005) Confronting different models of community structure to species-abundance data: a Bayesian model comparison. *Ecology Letters*, **8**, 493–504.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical population biology*, **3**, 87–112.

Falush, D., Stephens, M. & Pritchard, J.K. (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, **7**, 574–578.

Falush, D., Stephens, M. & Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Ferguson, T.S. (1973) A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 209–230.

Fisher, C.K. & Mehta, P. (2014) The transition between the niche and neutral regimes in ecology. *Proceedings of the National Academy of Sciences of the United States of America*, **111**, 13111–13116.

Fisher, J.B., Huntzinger, D.N., Schwalm, C.R. & Sitch, S. (2014) Modeling the Terrestrial Biosphere. *Annual Review of Environment and Resources*, **39**, 91–123.

Fisher, R.A. (1925) *Statistical methods for research workers*, Genesis Publishing Pvt Ltd.

Fisher, R.A. (1930) *The genetical theory of natural selection: a complete variorum edition*, Oxford University Press.

Fisher, R.A., Corbet, A.S. & Williams, C.B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 42–58.

Fortunato, S. (2010) Community detection in graphs. *Physics Reports*.

Fuhrman, J.A. (2009) Microbial community structure and its functional implications. *Nature*, **459**, 193–199.

Gause, G.F. (1932) Experimental studies on the struggle for existence: 1. Mixed population of two species of yeast. *Journal of Experimental Biology*, **9**, 389–402.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2014) *Bayesian data analysis*, Chapman & Hall/CRC Boca Raton, FL, USA.

Ghalambor, C.K., Hoke, K.L., Ruell, E.W., Fischer, E.K., Reznick, D.N. & Hughes, K.A. (2015) Non-adaptive plasticity potentiates rapid adaptive evolution of gene expression in nature. *Nature*, **525**, 372–375.

Gibson, J., Shokralla, S., Porter, T.M., King, I., van Konynenburg, S., Janzen, D.H., 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.

Gilbert, J.A., Jansson, J.K. & Knight, R. (2014) The Earth Microbiome project: successes and aspirations. *Bmc Biology*, **12**.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, **345**, 60–63.

Gleason, H.A. (1926) The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club*, 7–26.

Gourlet-Fleury, S., Ferry, B., Molino, J.-F., Petronelli, P. & Schmitt, L. (2004) *Experimental plots:*

key features, Elsevier.

Gravel, D., Canham, C.D., Beaudet, M. & Messier, C. (2006) Reconciling niche and neutrality: the continuum hypothesis. *Ecology Letters*, **9**, 399–409.

Griffiths, T. & Steyvers, M. (2004) Collapsed Gibbs Sampling for LDA. **101**, 5228–5235.

Grinnell, J. (1917) The niche-relationships of the California Thrasher. *The Auk*, **34**, 427–433.

Guisan, A. & Thuiller, W. (2005) Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, **8**, 993–1009.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O. & Townshend, J.R.G. (2013) High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, **342**, 850–853.

Hanson, C.A., Fuhrman, J.A., Horner-Devine, M.C. & Martiny, J.B.H. (2012) Beyond biogeographic patterns: processes shaping the microbial landscape. *Nature Reviews Microbiology*.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2015) Linking statistical and ecological theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE*, **PP**, 1–14.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 313–321.

Heckenberger, M.J., Russell, J.C., Fausto, C., Toney, J.R., Schmidt, M.J., Pereira, E., Franchetto, B. & Kuikuro, A. (2008) Pre-Columbian Urbanism, Anthropogenic Landscapes, and the Future of the Amazon. *Science*, **321**, 1214.

Hillebrand, H. (2004) On the generality of the latitudinal diversity gradient. *American Naturalist*, **163**, 192–211.

Holmes, I., Harris, K. & Quince, C. (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *Plos One*, **7**.

Holt, R.D. (2006) Emergent neutrality. *Trends in Ecology & Evolution*, **21**, 531–533.

Holt, R.D. (1996) *Food Webs in Space: An Island Biogeographic Perspective*. *Food Webs: Integration of Patterns & Dynamics* (ed. by G.A. Polis) and K.O. Winemiller), pp. 313–323. Springer US, Boston, MA.

Hoppe, F.M. (1984) Polya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, **20**, 91–94.

Houchmandzadeh, B. (2009) Theory of neutral clustering for growing populations. *Physical Review E*, **80**, 8.

Hubbell, S.P. (1997) A unified theory of biogeography and relative species abundance and its application to tropical rain forests and coral reefs. *Coral Reefs*, **16**, S9–S21.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Hubbell, S.P. (1979) Tree dispersion, abundance, and diversity in a tropical dry forest. *Science*, **203**, 1299–1309.

Hubisz, M.J., Falush, D., Stephens, M. & Pritchard, J.K. (2009) Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, **9**, 1322–1332.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M.,

Amundson, R., Thomas, B.C. & Banfield, J.F. (2016) A new view of the tree of life. *Nature Microbiology*, **1**, 16048.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415–427.

Hutchinson, G.E. (1961) The paradox of the plankton. *The American Naturalist*, **95**, 137–145.

Jabot, F., Etienne, R.S. & Chave, J. (2008) Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos*, **117**, 1308–1320.

Jain, A.K. (2010) Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, **31**, 651–666.

Janzen, D.H. (1970) Herbivores and the number of tree species in tropical forests. *The American Naturalist*, **104**, 501–528.

Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.

Kalyuzhny, M., Kadmon, R. & Shnerb, N.M. (2015) A neutral theory with environmental stochasticity explains static and dynamic properties of ecological communities. *Ecology Letters*, **18**, 572–580.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, **8**, 761-U107.

Kraft, N.J.B., Adler, P.B., Godoy, O., James, E.C., Fuller, S. & Levine, J.M. (2015) Community assembly, coexistence and the environmental filtering metaphor. *Functional Ecology*, **29**, 592–599.

Lawton, J.H. (1999) Are There General Laws in Ecology? *Oikos*, **84**, 177.

Lawton, J.H., Bignell, D.E., Bolton, B., Bloemers, G.F., Eggleton, P., Hammond, P.M., Hodda, M., Holt, R.D., Larsen, T.B. & Mawdsley, N.A. (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature*, **391**, 72–76.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2008) Analyzing or explaining beta diversity? Comment. *Ecology*, **89**, 3238–3244.

Legendre, P. & De Caceres, M. (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters*, **16**, 951–963.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, Elsevier.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M. & Gonzalez, A. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**, 601–613.

Linnaeus, C. (1753) Species Plantarum.

Liu, B., Liu, L., Tsykin, A., Goodall, G.J., Green, J.E., Zhu, M., Kim, C.H. & Li, J. (2010) Identifying functional miRNA-mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics*, **26**, 3105–3111.

Livermore, J.A. & Jones, S.E. (2015) Local-global overlap in diversity informs mechanisms of bacterial biogeography. *ISME J*, **9**, 2413–2422.

Loreau, M. & de Mazancourt, C. (2013) Biodiversity and ecosystem stability: a synthesis of underlying mechanisms. *Ecology Letters*, **16**, 106–115.

MacArthur, R.H. (1972) *Geographical ecology: patterns in the distribution of species*, Princeton University Press.

MacArthur, R.H. (1957) On the relative abundance of bird species. *Proceedings of the National Academy of Sciences of the United States of America*, **43**, 293–5.

MacArthur, R.H. (1958) Population ecology of some warblers of northeastern coniferous forests. *Ecology*, **39**, 599–619.

MacArthur, R.H. & Wilson, E. 0. (1967) The theory of island biogeography. *Monographs in Population Biology*, **1**.

Maire, V., Gross, N., Börger, L., Proulx, R., Wirth, C., Pontes, L. da S., Soussana, J.-F. & Louault, F. (2012) Habitat filtering and niche differentiation jointly explain species relative abundance within grassland communities along fertility and disturbance gradients. *New Phytologist*, **196**, 497–509.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z. & others (2005) Genome sequencing in open microfabricated high density picoliter reactors. *Nature*, **437**, 376.

Mariadassou, M., Pichon, S. & Ebert, D. (2015) Microbial ecosystems are dominated by specialist taxa. *Ecology Letters*, **18**, 974–982.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R.K., Fuhrman, J.A., Green, J.L., Horner-Devine, M.C., Kane, M., Krumins, J.A., Kuske, C.R., Morin, P.J., Naeem, S., Øvreås, L., Reysenbach, A.-L., Smith, V.H. & Staley, J.T. (2006) Microbial biogeography: putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102–112.

Martiny, J.B.H., Eisen, J.A., Penn, K., Allison, S.D. & Horner-Devine, M.C. (2011) Drivers of bacterial beta-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7850–7854.

Mauch, M., MacCallum, R.M., Levy, M. & Leroi, A.M. (2015) The evolution of popular music: USA 1960-2010. *Royal Society open science*, **2**, 150081–150081.

McCann, K.S. (2000) The diversity-stability debate. *Nature*, **405**, 228.

McGill, B.J. (2003) A test of the unified neutral theory of biodiversity. *Nature*, **422**, 881–885.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F.L., Hurlbert, A.H., Magurran, A.E., Marquet, P.A., Maurer, B.A., Ostling, A., Soykan, C.U., Ugland, K.I. & White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, **10**, 995–1015.

McGill, B.J., Maurer, B.A. & Weiser, M.D. (2006) Empirical evaluation of neutral theory. *Ecology*, **87**, 1411–1423.

McKane, A.J., Alonso, D. & Sole, R.V. (2004) Analytic solution of Hubbell's model of local community dynamics. *Theoretical Population Biology*, **65**, 67–73.

Miller, J. (2010) Species Distribution Modeling. *Geography Compass*, **4**, 490–509.

Moran, P.A.P. (1958) *Random processes in genetics*. *Mathematical Proceedings of the Cambridge Philosophical Society*, pp. 60–71. Cambridge University Press.

Morrone, J.J. (2015) Biogeographical regionalisation of the world: a reappraisal. *Australian Systematic Botany*, **28**, 81.

Noble, I.R. & Slatyer, R.O. (1977) *Post-fire succession of plants in Mediterranean ecosystems*. *Symposium on Environmental Consequences of Fire and Fuel Management in*

Mediterranean Ecosystems, Palo Alto, CA, USA, pp. 27–36.

O'Dwyer, J.P., Lake, J.K., Ostling, A., Savage, V.M. & Green, J.L. (2009) An integrative framework for stochastic, size-structured community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6170–6175.

Ofiteru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C.A. & Sloan, W.T. (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15345–15350.

Olszewski, D. (2012) Employing Kullback-Leibler divergence and Latent Dirichlet Allocation for fraud detection in telecommunications. *Intelligent Data Analysis*, **16**, 467–485.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.

Paine, R.T. (1980) Food Webs: Linkage, Interaction Strength and Community Infrastructure. *The Journal of Animal Ecology*, **49**, 666.

Palmer, M.W. (1994) Variation in species richness: towards a unification of hypotheses. *Folia Geobotanica*, **29**, 511–530.

Pawitan, Y. (2001) *In all likelihood: statistical modelling and inference using likelihood*, Oxford University Press.

Poisot, T., Stouffer, D.B. & Gravel, D. (2014) Beyond species: why ecological interaction networks vary through space and time. *BioRxiv preprint*.

Polis, G.A., Sears, A.L., Huxel, G.R., Strong, D.R. & Maron, J. (2000) When is a trophic cascade a trophic cascade? *Trends in Ecology & Evolution*, **15**, 473–475.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Preston, F.W. (1960) Time and space and the variation of species. *Ecology*, **41**, 611–627.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Pueyo, S., He, F. & Zillio, T. (2007) The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecology Letters*, **10**, 1017–1028.

Ramirez, K.S., Leff, J.W., Barberan, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H. & Fierer, N. (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 9.

Ricklefs, R.E. (2003) A comment on Hubbell's zero-sum ecological drift model. *Oikos*, **100**, 185–192.

Ricklefs, R.E. (1987) Community diversity: relative roles of local and regional processes. *Science(Washington)*, **235**, 167–171.

Ricklefs, R.E. (2006) The unified neutral theory of biodiversity: Do the numbers add up? *Ecology*, **87**, 1424–1431.

Roguet, A., Laigle, G.S., Therial, C., Bressy, A., Soulignac, F., Catherine, A., Lacroix, G., Jardillier, L., Bonhomme, C., Lerch, T.Z. & Lucas, F.S. (2015) Neutral community model explains the bacterial community assembly in freshwater lakes. *Fems Microbiology Ecology*, **91**, 11.

Rønsted, N., Weiblen, G.D., Cook, J.M., Salamin, N., Machado, C.A. & Savolainen, V. (2005) 60 million years of co-divergence in the fig–wasp symbiosis. *Proceedings of the Royal*

Society B: Biological Sciences, **272**, 2593–2599.

Rosen-Zvi, M., Gri_ffiths, T., Steyvers, M. & Smyth, P. (2004) *The Author-Topic Model for Authors and Documents*.

Rosindell, J., Cornell, S.J., Hubbell, S.P. & Etienne, R.S. (2010) Protracted speciation revitalizes the neutral theory of biodiversity. *Ecology Letters*, **13**, 716–727.

Rosindell, J., Hubbell, S.P., He, F., Harmon, L.J. & Etienne, R.S. (2012) The case for ecological neutral theory. *Trends in Ecology & Evolution*, **27**, 203–208.

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2009) The map equation. *The European Physical Journal Special Topics*, **178**, 13–23.

Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, J.C., Hutchison, C.A., Slocombe, P.M. & Smith, M. (1977) Nucleotide sequence of bacteriophage φX174 DNA. *nature*, **265**, 687–695.

Scheffer, M., Carpenter, S.R., Lenton, T.M., Bascompte, J., Brock, W., Dakos, V., Van de Koppel, J., Van de Leemput, I.A., Levin, S.A., Van Nes, E.H. & others (2012) Anticipating critical transitions. *science*, **338**, 344–348.

Scheffer, M. & van Nes, E.H. (2006) Self-organized similarity, the evolutionary emergence of groups of similar species. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 6230–6235.

Scheffers, B.R., Joppa, L.N., Pimm, S.L. & Laurance, W.F. (2012) What we know and don't know about Earth's missing biodiversity. *Trends in Ecology & Evolution*, **27**, 501–510.

Schemske, D.W., Mittelbach, G.G., Cornell, H.V., Sobel, J.M. & Roy, K. (2009) Is There a Latitudinal Gradient in the Importance of Biotic Interactions? *Annual Review of Ecology, Evolution, and Systematics*, **40**, 245–269.

Schuster, S.C. (2007) Next-generation sequencing transforms today's biology. *Nature Methods*, **5**, 16–18.

Shafiei, M., Dunn, K.A., Boon, E., MacDonald, S.M., Walsh, D.A., Gu, H. & Bielawski, J.P. (2015) BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome*, **3**, 8.

Shmida, A.V.I. & Wilson, M.V. (1985) Biological determinants of species diversity. *Journal of biogeography*, 1–20.

Simberloff, D.S. & Wilson, E.O. (1969) Experimental Zoogeography of Islands: The Colonization of Empty Islands. *Ecology*, **50**, 278–296.

Sizling, A.L., Storch, D., Sizlingova, E., Reif, J. & Gaston, K.J. (2009) Species abundance distribution results from a spatial analogy of central limit theorem. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6691–6695.

Sloan, W.T., Lunn, M., Woodcock, S., Head, I.M., Nee, S. & Curtis, T.P. (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, **8**, 732–740.

Sloan, W.T., Woodcock, S., Lunn, M., Head, I.M. & Curtis, T.P. (2007) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microbial Ecology*.

Soininen, J., McDonald, R. & Hillebrand, H. (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

ter Steege, H., Nigel, C.A., Sabatier, D., Baraloto, C., Salomao, R.P., Guevara, J.E., Phillips, O.L.,

Castilho, C.V., Magnusson, W.E., Molino, J.F., Monteagudo, A., Vargas, P.N., Montero, J.C., Feldpausch, T.R., Coronado, E.N.H., Killeen, T.J., Mostacedo, B., Vasquez, R., Assis, R.L., Terborgh, J., Wittmann, F., Andrade, A., Laurance, W.F., Laurance, S.G.W., Marimon, B.S., Marimon, B.H., Vieira, I.C.G., Amaral, I.L., Brienen, R., Castellanos, H., Lopez, D.C., Duivenvoorden, J.F., Mogollon, H.F., Matos, F.D.D., Davila, N., Garcia-Villacorta, R., Diaz, P.R.S., Costa, F., Emilio, T., Levis, C., Schietti, J., Souza, P., Alonso, A., Dallmeier, F., Montoya, A.J.D., Piedade, M.T.F., Araujo-Murakami, A., Arroyo, L., Gribel, R., Fine, P.V.A., Peres, C.A., Toledo, M., Gerardo, A.A.C., Baker, T.R., Ceron, C., Engel, J., Henkel, T.W., Maas, P., Petronelli, P., Stropp, J., Zartman, C.E., Daly, D., Neill, D., Silveira, M., Paredes, M.R., Chave, J., Lima, D.D., Jorgensen, P.M., Fuentes, A., Schongart, J., Valverde, F.C., Di Fiore, A., Jimenez, E.M., Mora, M.C.P., Phillips, J.F., Rivas, G., van Andel, T.R., von Hildebrand, P., Hoffman, B., Zent, E.L., Malhi, Y., Prieto, A., Rudas, A., Ruschell, A.R., Silva, N., Vos, V., Zent, S., Oliveira, A.A., Schutz, A.C., Gonzales, T., Nascimento, M.T., Ramirez-Angulo, H., Sierra, R., Tirado, M., Medina, M.N.U., van der Heijden, G., Vela, C.I.A., Torre, E.V., Vriesendorp, C., Wang, O., Young, K.R., Baider, C., Balslev, H., Ferreira, C., Mesones, I., Torres-Lezama, A., Giraldo, L.E.U., Zagt, R., Alexiades, M.N., Hernandez, L., Huamantupa-Chuquimaco, I., Milliken, W., Cuenca, W.P., Pauletto, D., Sandoval, E.V., Gamarra, L.V., Dexter, K.G., Feeley, K., Lopez-Gonzalez, G. & Silman, M.R. (2013) Hyperdominance in the Amazonian Tree Flora. *Science*, **342**, 325–+.

Svenning, J. & Skov, F. (2007) Could the tree diversity pattern in Europe be generated by postglacial dispersal limitation? *Ecology letters*, **10**, 453–460.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

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

Tansley, A.G. (1935) The Use and Abuse of Vegetational Concepts and Terms. *Ecology*, **16**, 284–307.

Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Tilman, D. (1982) *Resource competition and community structure*, Princeton university press.

Tilman, D., May, R.M., Lehman, C.L. & Nowak, M.A. (1994) Habitat destruction and the extinction debt. *Nature*, **371**, 65–66.

Tilman, D., Reich, P.B. & Knops, J.M.H. (2006) Biodiversity and ecosystem stability in a decade-long grassland experiment. *Nature*, **441**, 629–632.

Tokeshi, M. (1996) Power Fraction: A New Explanation of Relative Abundance Patterns in Species-Rich Assemblages. *Oikos*, **75**, 543–550.

Tuomisto, H., Ruokolainen, K. & Yli-Halla, M. (2003) Dispersal, environment, and floristic variation of western Amazonian forests. *Science*, **299**, 241–244.

Vaduva, C., Gavat, I. & Datcu, M. (2013) Latent Dirichlet Allocation for Spatial Analysis of Satellite Images. *Ieee Transactions on Geoscience and Remote Sensing*, **51**, 2770–2786.

Vallade, M. & Houchmandzadeh, B. (2003) Analytical solution of a neutral model of biodiversity. *Physical Review E*, **68**, 5.

Valle, D., Baiser, B., Woodall, C.W. & Chazdon, R. (2014) Decomposing biodiversity data using the

Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, **17**, 1591–1601.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Lukes, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E. & Tara Oceans, C. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**.

Vellend, M. (2010) Conceptual synthesis in community ecology. *Quarterly Review of Biology*, **85**, 183–206.

Vilhena, D.A. & Antonelli, A. (2015) A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, **6**.

Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, **424**, 1035–1037.

Wakeley, J. (2009) *Coalescent Theory: An Introduction*, Roberts & Company Publishers, Greenwood Village.

Wallace, A.R. (1876) The Geographical Distribution of Animals: With a Study of the Relations of Living and Extinct Faunas as Elucidating the Past Changes of the Earth's Surface: In Two Volumes.

Wang, H.G., Wei, Z., Mei, L.J., Gu, J.X., Yin, S.S., Faust, K., Raes, J., Deng, Y., Wang, Y.L., Shen, Q.R. & Yin, S.X. (2017) Combined use of network inference tools identifies ecologically meaningful bacterial associations in a paddy soil. *Soil Biology & Biochemistry*, **105**, 227–235.

Watson, H.C. (1859) *Cybele Britannica*,.

Watterson, G.A. (1974) Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, **6**, 217–250.

Whittaker, R.H. (1965) Dominance and Diversity in Land Plant Communities. *Science*, **147**, 250–260.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Williamson, M.H. (1988) *Relationship of species number to area, distance and other variables. Á In: Myers, AA and Giller, P. S.(eds), Analytical biogeography, an integrated approach to the study of animal and plant distributions*, Chapman and Hall, pp. 91Á115.

Wisz, M.S., Pottier, J., Kissling, W.D., Pellissier, L., Lenoir, J., Damgaard, C.F., Dormann, C.F., Forchhammer, M.C., Grytnes, J.-A., Guisan, A., Heikkinen, R.K., Høye, T.T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M.-C., Normand, S., Öckinger, E., Schmidt, N.M., Termansen, M., Timmermann, A., Wardle, D.A., Aastrup, P. & Svenning, J.-C. (2013) The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, **88**, 15–30.

Woodcock, S., van der Gast, C.J., Bell, T., Lunn, M., Curtis, T.P., Head, I.M. & Sloan, W.T. (2007) Neutral assembly of bacterial communities. *Fems Microbiology Ecology*, **62**, 171–180.

Wright, J.P., Jones, C.G. & Flecker, A.S. (2002) An ecosystem engineer, the beaver, increases species richness at the landscape scale. *Oecologia*, **132**, 96–101.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright, S.J. (2002) Plant diversity in tropical forests: a review of mechanisms of species coexistence. *Oecologia*, **130**, 1–14.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zillio, T., Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2005) Spatial scaling in model plant communities. *Physical Review Letters*, **95**.

Chapter 1

Causes of variation in soil beta diversity across domains of life in the tropical forests of French Guiana

Guilhem Sommeria-Klein[1,2], Lucie Zinger[1,2], Amaia Iribar[1], Eliane Louisanna[3], Sophie Manzi[1], Vincent Schilling[1], Eric Coissac[4], Heidy Schimann[3], Pierre Taberlet[4], Jérôme Chave[1]

[1] Université Toulouse 3 Paul Sabatier, CNRS, IRD, UMR 5174 Laboratoire Evolution et Diversité Biologique (EDB), F-31062 Toulouse, France.

[2] Ecole Normale Supérieure, CNRS, UMR 8197 Institut de Biologie de l'ENS (IBENS), F-75005 Paris, France

[3] INRA, AgroParisTech, CIRAD, CNRS, Université des Antilles, Université de la Guyane, UMR Ecologie des Forets de Guyane (EcoFoG), F-97379 Kourou, France.

[4] Université Grenoble Alpes, CNRS, UMR Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France.

Chapter outline

Beta diversity patterns, i.e. how taxonomic composition shifts through space, have long been used to infer the mechanisms of community assembly. Indeed, depending on whether taxonomic composition covaries with environmental conditions or with geographical distance, it can be inferred whether community assembly is driven by deterministic niche processes or by neutral dispersal limitation. In this chapter, this reasoning is applied to a soil DNA dataset collected in various 1-ha forest plots in French Guiana, for a range of barcodes spanning most of the tree of life. To enable both types of processes to be distinguished, the sampled plots cover a range of soil types as well as a range of inter-plot distances. Inter-plot distances are approximately regularly spaced on a logarithmic scale, so as to better assess the effect of dispersal limitation on taxonomic composition. Indeed, neutral dispersal limitation is predicted to yield a linear decrease of taxonomic similarity with log-distance. As a side question, the effect of past logging activities on soil biodiversity is assessed based on a set of disturbed forest plots.

Abstract

Disentangling the processes that cause the assembly of ecological communities is a key challenge, and these include both stochastic (neutral) processes and deterministic niche filtering. Progress in biodiversity assessment using environmental DNA now streamlines the study of biodiversity patterns across domains of life. Using soil DNA samples, we quantified the causes of variation in beta diversity patterns across major taxonomic groups in the lowland tropical forest of French Guiana on a spatial scale ranging from 40 m to 140 km, for a range of soil physico-chemical properties. We quantified the respective influence of soil conditions, dispersal limitation, and human disturbances on beta diversity. In undisturbed forest plots, we found that the beta diversity of bacteria and protists was primarily driven by soil conditions, while the observed patterns in plants, and to a lesser extent in annelids, were best explained by dispersal limitation. Both factors had an effect on fungi, arthropods and insects, whereas we could not detect influence of either factor on nematodes and flat worms. This analysis was consistent with a comparison of our data to the similarity decay predicted by the neutral theory of biodiversity. These results suggest that spatial patterns of plant biodiversity across the Amazon do not necessarily extend to other taxonomic groups, and that environmental factors play a foremost role in explaining these patterns in tropical soils. Along the disturbance gradient, we found a significant shift in taxonomic composition in two functionally important groups, plants and annelids, a smaller effect on fungi, and no effect in the other groups.

Introduction

Beta diversity describes the turnover of taxonomic composition through geographical and environmental space, and yields insight into the mechanisms of community assembly (Whittaker, 1960, 1972; Rosenzweig, 1995; Gaston & Blackburn, 2008). As a measure of the spatial variability of taxonomic composition, it may be broadly defined as the difference or ratio between regional (gamma) diversity and local (alpha) diversity (Whittaker, 1960; Chao *et al.*, 2012). This has important practical implications for biodiversity estimates and conservation (Basset *et al.*, 2012; Hubbell, 2013; ter Steege *et al.*, 2013; Socolar *et al.*, 2017).

The extent of beta diversity and its causal mechanisms are dependent on the spatial scale at which taxonomic turnover is considered (Soininen *et al.*, 2007). Beta diversity is often quantified within a biogeographic region, so that it is not caused by a large climatic difference or a different biogeographic history between stations (Kreft & Jetz, 2010). Variation in beta diversity can be ascribed to two types of processes: niche-based processes, when abiotic and biotic environmental heterogeneity determines the spatial distribution of taxa based on their phenotypic differences, and neutral processes, when turnover in taxonomic composition results from demographic stochasticity combined with limited dispersal (Leibold *et al.*, 2004). However, because environmental differences tend to also be spatially structured, both types of processes are often difficult to disentangle (Gilbert & Lechowicz, 2004).

One frontier in the study of beta diversity is that it has most often been restricted to a single taxonomic group, and especially forest trees (Whittaker, 1960, 1972; Nekola & White, 1999; Condit *et al.*, 2002), amphibians (Baselga *et al.*, 2012), and arthropods (Harrison *et al.*, 1992; Novotny *et al.*, 2007; Hortal *et al.*, 2011), and freshwater taxa (Cottenie, 2005). Studies that have attempted to compare patterns of beta diversity across taxa are scarce (but see Harrison *et al.*, 1992). This is largely because the effort needed to coordinate inventories of biological diversity across taxa is enormous, and increases dramatically for smaller-bodied taxa (Lawton *et al.*, 1998). DNA-based methods have lifted this constraint and they have dramatically widened the range of

taxa for which diversity patterns can be measured. Instead of collecting organisms and assigning them a taxon label based on observation and on expert knowledge, identification is based on minute amounts of biological material and on the sequencing of universal DNA amplicons (DNA barcodes), a method first developed for microorganisms (Pace, 1997). This method has been extended to rapid taxonomic surveys: bulk DNA is extracted from environmental samples and DNA is amplified using universal primers, then sequenced (Taberlet *et al.*, 2012, Yu et al. 2012). This environmental DNA approach to biological diversity inventory aims at detecting the presence of cells or of extracellular DNA for a range of taxa in a sample. Such an approach is in principle applicable to any taxonomic group in the tree of life (Bahram *et al.*, 2013; Schuldt *et al.*, 2015; Siles & Margesin, 2016; Vincent *et al.*, 2016). Since it is possible to normalize the DNA extraction and sequencing procedures for many samples at once, such an approach is suited to the exploration of beta diversity patterns.

We expect that smaller organisms with short generation times display higher beta diversity at short spatial scale, i.e. over a few meters, than larger organisms, because they are locally filtered by environmental heterogeneity (Ramirez *et al.*, 2014; Mariadassou *et al.*, 2015). Conversely, the beta diversity of small organisms is predicted to be less dependent on distance compared to large organisms, owing to their higher dispersal ability (Soininen *et al.*, 2007). Thus, we expect the spatial distribution of small organisms to be primarily governed by niche effects, while we expect large organisms to better comply with distance-limited neutral dynamics (Hubbell, 2001; Martiny *et al.*, 2011). These predictions have direct implications for the maintenance of biodiversity in disturbed landscapes. Organisms with higher dispersal abilities should be found even in heavily disturbed habitats. On the other hand, slow dispersers should be more affected by disturbances, and would also take longer to recolonize habitats after abandonment.

In this study, we compare soil beta diversity patterns across domains of life in a lowland tropical rainforest. We collected soil samples at locations separated by a geographical distance ranging from 40 m to 140 km, and spanning a variety of soil types, which we quantified, as well as a range of human disturbance intensities. We targeted taxonomic groups using barcodes with different levels of taxonomic resolution, which allowed us to test the robustness of the observed patterns. We thus address the following questions: 1) What is the relative importance of dispersal

limitation and environmental filtering in explaining beta diversity across taxonomic groups? 2) How good a fit is the dispersal-limited neutral theory for the various taxonomic groups? 3) How does beta diversity depend on forest disturbance by logging activities? Finally, we explore the implications of our findings for community ecology and for the conservation of tropical forest ecosystems.

Methods

1. Sampling scheme

We sampled fifteen 1-ha plots in the undisturbed lowland rain forest of French Guiana, to which we added four 1-ha plots in disturbed habitats (see below). Geographical distances between plots in pristine forest are approximately regularly spaced on a logarithmic scale. This choice was motivated by the expectation of a linear relationship between taxonomic similarity and log-distance in a spatially explicit neutral model (Chave & Leigh, 2002). Twelve plots are located at the Nouragues research station (about 100 km inland; latitude 4° 5’ 17" N and longitude 52° 40’ 48" W; Bongers *et al.*, 2001), and three at the Paracou research station (near the coast; latitude 5° 18′ N and longitude 52° 53′ W; Gourlet-Fleury et al. 2004); see Fig. 1 for locations. All plots consist of *terra firme* forest, but cover a range of soil types (see below).

In addition to sampling plots in undisturbed forest, we also sampled areas that have undergone disturbances of different intensities. At Paracou, some plots have been experimentally logged at several logging intensities starting in 1986 (https://paracou.cirad.fr/experimental-design). In the two heaviest logging treatments (T2 and T3), 33-56% of the aboveground biomass was lost due to the felling operations. Eighteen years after logging, the impact of logging activities was still visible. We sampled two contiguous 1-ha plots in one of the most heavily impacted areas (P12 plot). We also sampled two contiguous 1-ha plots in a 25-ha area (Arbocel plot) 14 km away from Paracou, that was fully clear-cut in 1976 and left regenerating since then.

Within each 1-ha plot, we collected eighty soil samples of about 30 g each with an auger from the mineral soil horizon (~10 cm deep) along a square grid. To minimize sampling bias and coarsen the spatial grain, we pooled soil samples five by five following a cross-shaped pattern about 15 meters across, with one sample at the centre and four samples in the corners (Fig. 2). This resulted in sixteen pooled samples per plot. We extracted DNA from about 10 g of soil per pooled sample within a few hours after

sample collection, using the protocol described in Zinger *et al.* (2016). The remaining soil was dried for subsequent analyses of soil properties.

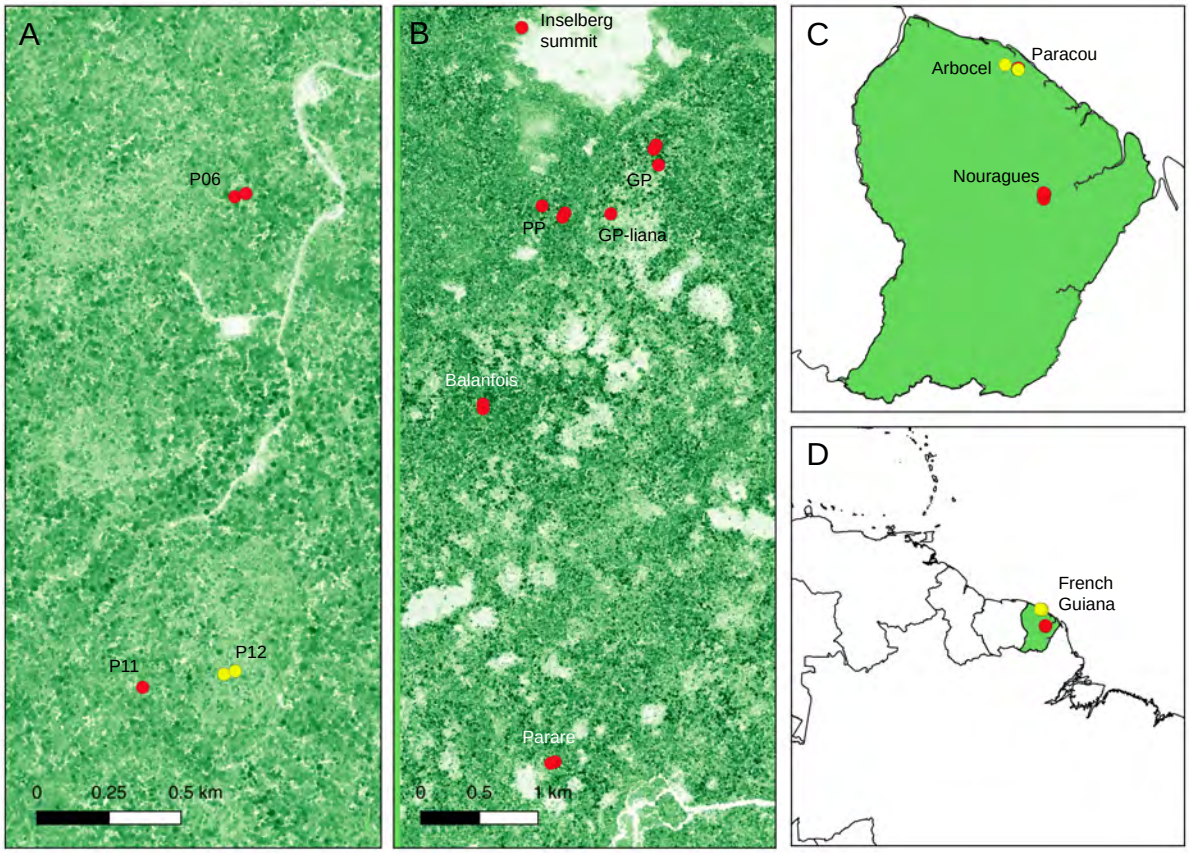

Figure 1: Sampling scheme. Relative position of all sampled 1-ha forest plots, in (A) Paracou, and (B) Nouragues; (C) relative position of the Paracou, Arbocel, and Nouragues sites. Undisturbed plots are in red and the four disturbed plots (two in Paracou and two in Arbocel) in yellow. In Nouragues, PP and GP denote respectively the Petit Plateau and Grand Plateau permanent monitored plots, and 'GP-liana' denotes the L18 subplot in Grand Plateau.

DNA amplification and sequencing yielded read counts for Operational Taxonomic Units (OTUs) at sixteen sites per plot (see below). We further pooled these samples four by four by averaging relative OTU abundances, so as to obtain one sampling point per 0.25-ha plot (Fig. 2). We defined the distance between two sampling points as the distance between the centres of the two sets of pooled samples. Some samples were removed from the dataset owing to insufficient PCR yields (see below); hence some sampling points have fewer than four samples or are missing.

Soil samples were also pooled four by four to obtain a single composite sample per 0.25-ha subplot. For each pooled soil sample, twelve measurements were made from about 60 g of dry soil. Granulometry distinguished the clay (0-2 μm), silt (2-63 μm) and sand fractions (63-2000 μm). The pH of soil in water solution was measured, as well as total carbon (C) and nitrogen (N) mass fractions. The mass fraction of plant-available phosphorus (P_2O_5) was measured using the Olsen extraction method. Lastly, a $BaCl_2$ extraction was performed and the concentration of major elements was measured using ICPMS (Ca, Mg, K, Fe, Mg, and Al).

2. Molecular and sequence analyses

We amplified five barcodes by PCR from soil samples, targeting bacteria (16S rRNA gene V5-V6 regions; Fliegerova *et al.*, 2014), eukaryotes (18S rRNA gene v7 region; Guardiola *et al.*, 2015), Viridiplantae (chloroplastic trnL-P6 loop; Taberlet *et al.*, 1991), fungi (ITS1) and insects (mitochondrial 16S rRNA; Clarke *et al.*, 2014). Each soil sample was amplified thrice independently by PCR, following the same protocol as in Zinger *et al.* (2017). Amplicons were labelled with a distinct nucleotide tag for each PCR, and six sequencing libraries, one per barcode, were prepared. Sequencing was carried out using paired-end Illumina sequencing (MiSeq 2x250 for 16S bacteria, 16S insects and ITS fungi; HiSeq 2x100 for 18S eukaryotes and trnL plants). Negative PCR controls were included in the protocol to help detect contaminants. The PCRs that yielded less than 1,000 reads were discarded from subsequent analyses.

Data analyses were conducted as in Zinger *et al.* (2017). Sequencing data were curated using the OBITools package (Boyer *et al.*, 2016): paired-end reads were assembled, dereplicated, and low-quality sequences were excluded. The resulting sequences were clustered into OTUs using the Infomap algorithm (Rosvall *et al.*, 2009), with a dissimilarity threshold of three mismatches and exponentially decreasing weights on edges. OTUs represented by a single sequence were removed, and the most abundant sequence in the cluster was taken to be the true sequence. Taxonomic identifications were assigned to OTUs using the ecotag program in the OBITools package based on Genbank and SILVA databases (Zinger *et al.*, 2017). OTUs with less than 75% similarity

to any reference sequence were removed, as well as those with a taxonomic identification outside of the taxonomic group targeted by the barcode. Further steps were taken to minimize the number of contaminant OTUs as described in Zinger *et al.* (2017). Rare OTUs were not removed, and only the relative OTU abundances in each sample were used for further analyses.

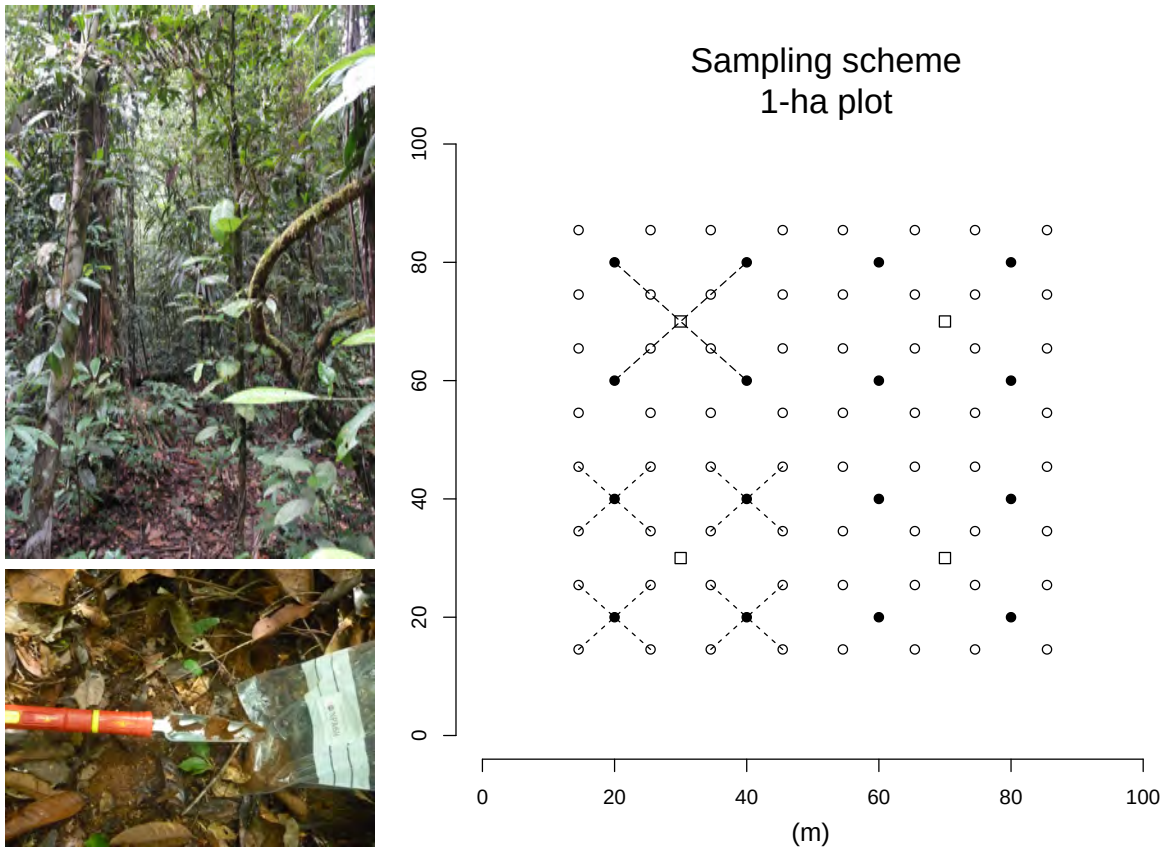

Figure 2: Sampling scheme in each 1-ha forest plot. In each of the nineteen plots (fifteen undisturbed and four disturbed), eighty soil samples were collected (open and full black circles), and were pooled five by five (small dashed crosses). After conducting the molecular and sequence analyses on the sixteen pooled samples (full black circles), results were pooled four by four (large dashed cross), and statistical analyses were performed on the resulting four effective sampling points (open squares). The sixteen pooled soil samples were also directly pooled four by four for conducting soil analyses.

Taxonomic identifications for the eukaryote 18S marker were used to assign OTUs to sub-clades (Table S1): arthropods, insects, annelids, nematodes, flat worms (Platyhelminthes), protists, fungi, and plants (Viridiplantae). The 18S marker was

compared with more specific markers for fungi, plants, and insects (ITS1, trnL, and 16S, respectively; Table S1). A rarefaction analysis was performed for each marker by sampling with replacement between 1 and 8,000 reads per sample (Fig. S1). For all markers, the number of OTUs reached near saturation in most samples.

3. Statistical analyses

We performed all statistical analyses in R using the 'vegan' package (version 2.4-2, available at https://cran.r-project.org/), and followed the guidelines of Legendre & Legendre (2012). Analyses were performed separately for each taxonomic group.

We performed a PCA on soil variables after centering and normalizing them (i.e., subtracting their mean and dividing them by their standard deviation over all sampling points). Since clay, silt and sand fractions sum to 1, they yield only two independent measurements; we chose to keep clay and silt fractions, as clay and sand fractions were almost perfectly anticorrelated (correlation coefficient of -0.97; see Results, Table S3). Before conducting the PCA, we lumped Ca, Mg, Mn and K concentrations together into a single 'exchangeable cations' variable. In all further analyses, we used the first four PCA axes as environmental variables.

We first studied the taxonomic dissimilarity among pairs of sampled locations ('distance-based' approach). We computed the Sorensen taxonomic dissimilarity index (number of non-shared OTUs divided by number of OTUs in both samples), which is one possible measure of occurrence-based beta diversity (Koleff *et al.*, 2003). The Sorensen index between pairs of sampling points was regressed against their environmental dissimilarity and against the logarithm of their geographical distance (measured in meters). The environmental dissimilarity between two sampling points was defined as their Euclidian distance with respect to the four soil PCA axes. To test the significance of regressions of the Sorensen index against environmental and geographical distances, we performed Mantel tests with 999 permutations using simple and partial Pearson's correlation coefficients as test statistics (functions 'mantel' and 'mantel.partial').

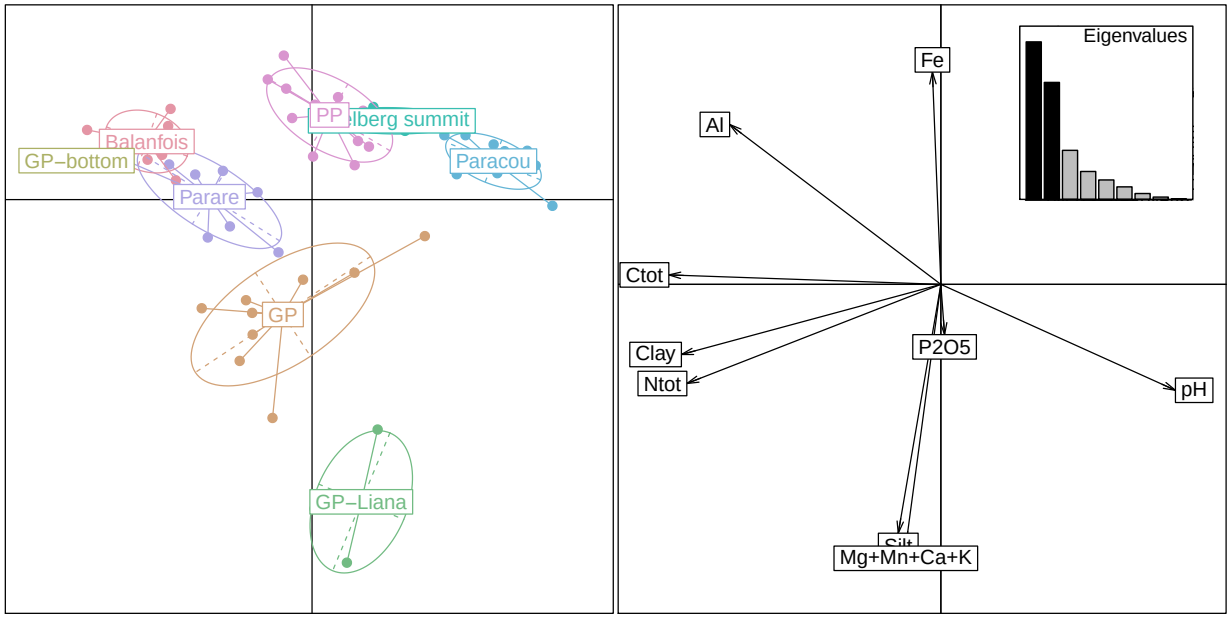

Figure 3. Principal Component Analysis of soil variables for the fifteen undisturbed plots, projected on the first two axes (40% and 30% of total variance). 'GP-bottom' corresponds to the lower half of the GP-O13 plot, which belongs to a bottomland.

We then directly compared the taxonomic composition of sampled locations using canonical ordination ('raw-data' approach; Legendre *et al.*, 2005). We regressed the OTU abundance data on environmental and spatial variables using Canonical Redundancy Analysis (RDA; function 'rda'). We first applied the Hellinger transformation to OTU abundance data (i.e., square-root of the relative OTU abundances at each sampling point) and centred them per OTU (i.e., subtracted the mean over sampling points). We used the six selected soil variables as explanatory environmental variables, after centring and normalization. We used Principal Coordinates of Neighbour Matrices (PCNM) as spatial explanatory variables representing different possible patterns of spatial autocorrelation in the data (Borcard & Legendre, 2002; Borcard *et al.*, 2004). Two separate PCNM decompositions were performed for the Nouragues and Paracou sites (function 'pcnm'; Borcard & Legendre, 2002), i.e. in each site we performed a Principal Coordinates Analysis of the distance matrix between sampling points, after setting all distances larger than a threshold distance to four times this threshold distance (chosen as the minimal distance required to connect all sampling points). We obtained seventeen PCNM variables with positive eigenvalues for Nouragues, and six for Paracou. PCNM variables from both sites were assembled into a

single staggered matrix. The two submatrices were connected by adding a 'dummy' variable distinguishing Nouragues and Paracou sites by two different values. At each site, we added UTM coordinates (northings and eastings) as two additional explanatory variables after centring and normalization, so as to account for linear spatial trends that cannot be captured by PCNM variables.

The total variance of taxonomic composition was partitioned between an environmental component and a spatial component (function 'varpart'; Borcard *et al.*, 1992; Legendre *et al.*, 2005). Two RDA-based forward selections of environmental variables and spatial variables were performed separately (function 'ordiR2step' with 0.05 threshold p-value for adding a variable to the model; Blanchet *et al.*, 2008), yielding two RDA-based linear models. We only proceeded with variable selection when the RDA conducted on all variables was significant ($p < 0.05$; Blanchet *et al.*, 2008); when it was not for either environmental or spatial variables, we did not partition the variance.

We then tested the predictions of the dispersal-limited neutral theory on the dataset. Neutral processes are predicted to yield a decay of taxonomic similarity with distance in the absence of dispersal barrier (Chave & Leigh, 2002). We used here $F_2(A,B) = \sum_{s=1}^{S} p_s^A p_s^B$ as a measure of taxonomic similarity between samples A and B, where p_s^A is the proportion of species s in sample A, p_s^B that in sample B, and S the total number of species. Chave and Leigh (2002) predicted that in a continuous spatially explicit dispersal-limited neutral model with spatial density of individuals ρ, dispersal parameterized by a Gaussian kernel with variance σ^2, and a rate of apparition of new species equal to ν, $F_2(A,B)$ depends only on the pairwise distance r between samples, and can be expressed as $F_2(r) = -a\ln(r) + b$, with $b/a = \ln\left(\sqrt{2\nu}/2\sigma\right) + \gamma$ (where γ is Euler's constant) and $1/a = \rho\pi\sigma^2 - \ln(\nu)/2$ (cf. Appendix). We measured F_2 among pairs of sampling points, regressed it against the log-transformed geographical distance $\ln(r)$, and assessed significance by Mantel test for 999 permutations, using Pearson's correlation coefficient as test statistics. The mean dispersal distance per generation $\sqrt{2}\sigma$ can be obtained provided that an estimate of ρ is available. For plants, we assumed that most of DNA retrieved came from tree species, and that the forest holds 500 mature trees (≥ 10 cm dbh) per hectare, i.e. $\rho = 0.05\ \mathrm{m}^{-2}$, which is close to observed densities

(see Condit *et al.* 2002). We also computed the quantity σ^2/v, which may be interpreted as the ratio between dispersal ability and diversification rate.

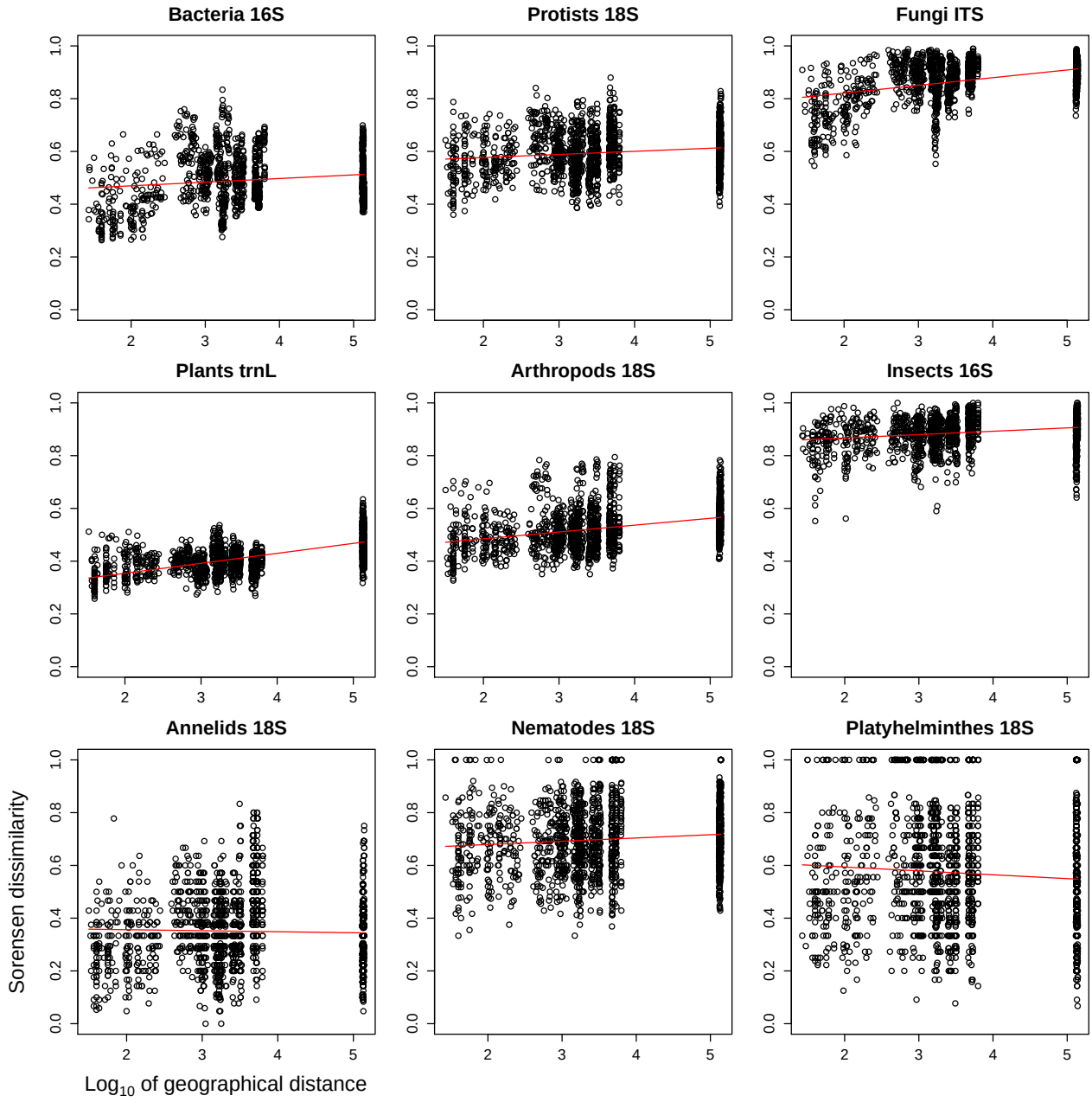

Figure 4: Occurrence-based (Sorensen) dissimilarity as a function of log-distance. The red line figures the linear regression.

Finally, we conducted a separate analysis to explore how beta diversity depends on logging activities. Because our sampling effort along this disturbance gradient was

limited, we simply investigated the relative effect of disturbance and soil conditions on the various taxonomic groups without accounting for spatial structure. We measured the Sorensen dissimilarity index among pairs of sampling points in all Paracou and Arbocel plots, both disturbed and undisturbed. We quantified logging intensity by a dummy variable taking value 0 in undisturbed locations (Paracou P6 and P11 areas), 1 in mildly disturbed ones (Paracou P12 area), and 2 in strongly disturbed ones (clear cutting; Arbocel). We then followed a similar approach as for the comparison between soil effects and spatial aggregation in the main dataset. We performed a multivariate linear regression (i.e., a one-dimensional RDA) of the OTU abundance data (Hellinger-transformed and OTU-centred) against the logging intensity variable (centred and normalized). When the linear regression was significant, we partitioned the total variance of taxonomic composition between a logging intensity component and a soil component. We obtained the soil component as previously: we performed a PCA on soil variables, kept the first four axes, and built a RDA-based model by forward variable selection.

Results

Chemical and physical soil properties varied across the samples (Table S2). The pH ranged from 3.8 to 5.5, C content from 1.9% to 4.2%, N content from 0.12 to 0.31%, and P content was very low (see also Grau *et al.*, 2017). Soils were also poor in terms of exchangeable cation content (K^+, Ca^{2+}, Mg^{2+}, Mn^{2+}), and varied significantly in terms of texture, with sandy (up to 80% sand) to clayey (up to 80% clay) soils. Paracou soils tended to be sandier and more nutrient-poor than Nouragues soils. This suggests that the Nouragues-Paracou comparison compounds geographical distance and environmental distance effects. The first PCA axis (40% of total variance) corresponds to organic matter (total carbon and nitrogen) and clay contents, which are correlated to aluminium concentration and anticorrelated to pH; the second PCA axis (30% of variance) corresponds to nutrient and silt contents, the third to phosphorus (13% of variance) and the fourth to iron (7%; Fig. 3).

		Geographical distance			Soil		
	Mean $D_{Sorensen}$	r_{dist}	$r_{dist,part}$	slope$_{dist}$	r_{soil}	$r_{soil,part}$	slope$_{soil}$
Plants trnL	0.42	0.65***	0.61***	0.038	0.29***	0.06	0.011
Bacteria 16S	0.49	0.16*	-0.02	0.014	0.46***	0.44***	0.028
Protists 18S	0.60	0.16**	0.05	0.012	0.30***	0.26***	0.015
Fungi ITS	0.87	0.43***	0.29***	0.029	0.54***	0.45***	0.025
Arthropods 18S	0.53	0.36***	0.29***	0.026	0.28***	0.17*	0.014
Insects 16S	0.89	0.23***	0.16**	0.013	0.25***	0.18**	0.010
Annelids 18S	0.35	-0.031	-0.08	-0.004	0.10	0.12	0.009
Nematodes 18S	0.70	0.11*	0.09	0.012	0.05	0.02	0.004
Platyhelminthes 18S	0.57	-0.079	-0.11*	-0.015	0.07	0.10	0.009

Table 1: Linear regression of taxonomic dissimilarity (Sorensen index) against soil and geographical distance. r_{dist}, r_{soil}, $r_{dist,part}$, $r_{soil,part}$ are the simple and partial Pearson's correlation coefficients. Significance was assessed using Mantel tests: *** for $p < 0.001$; ** for $0.001 < p < 0.01$; * for $0.01 < p < 0.05$.

Sorensen dissimilarity varied across taxonomic groups, and for the same group depending on the tested DNA barcode (Table 1; see Table S4, Fig. S2 and S3 for a comparison between barcodes within group). It was highest for insects 16S and fungi ITS (ca. 0.9 in average), and lowest for annelids and plants trnL (ca. 0.4 in average). When plotted against log-transformed geographical distance (Fig. 4), Sorensen dissimilarity showed a strongly significant correlation for plants, fungi, arthropods and insects, a weak correlation for protists, bacteria, and nematodes, and no correlation for annelids and flat worms (by decreasing order of correlation coefficient; Table 1). Sorensen dissimilarity was also regressed against soil dissimilarity (Fig. 5). We found a strong correlation to soil dissimilarity in fungi, bacteria, protists, plants, arthropods and insects, and no correlation in annelids, flat worms and nematodes (Table 1). To test a possible collinearity between soil dissimilarity and geographical distance, we finally computed the partial correlation $r_{dist,part}$ to log-distance conditional on soil dissimilarity. The partial correlation to log-distance was significant in plants, fungi, arthropods, and insects, but not in the other groups. Conversely, when computing the partial correlation $r_{soil,part}$ to soil dissimilarity conditional on log-distance, the correlation was retained in fungi, bacteria, protists, insects and arthropods, but lost in plants.

RDA-based partitioning of beta-diversity showed that environmental factors and spatial aggregation together explained a proportion of beta-diversity that ranged from 45% in bacteria to zero in flat worms (Fig. 6, Tables 2, S5). Within the fraction of beta diversity explained by soil effects, the first two soil PCA axes were the main explanatory factors, with the silt-nutrient axis playing a particularly important role in bacteria (Fig. S4). The relative contribution of spatial aggregation and soil properties varied across groups, with a major effect of spatial aggregation relative to soil in annelids and plants, while both effects were of the same magnitude for bacteria. While the collinearity between environmental and spatial variables introduced uncertainty as to their actual relative importance to beta diversity, pure spatial aggregation explained an equal or higher proportion of the variation compared to pure environmental factors in all groups. For bacteria and protists, this contrasts with the conclusions of distance-based analyses.

The fit of the neutral prediction for the decay of taxonomic similarity F_2 with geographical distance was statistically significant for plants, bacteria, protists, fungi, insects and annelids, but not for arthropods, nematodes and flat worms (Table S6, Fig.

S6). At a given geographical distance, the F_2 statistic tended to be more scattered than Sorensen dissimilarity and to exhibit outliers (Fig. S6). Assuming a density of one plant individual per 20 m², as measured for mature neotropical forest trees, we estimated a mean dispersal distance per generation of 43 m in plants. The dispersal to diversification ratio σ^2/ν was highest for fungi and insects, intermediate for plants and annelids, and smallest for protists and bacteria (Table S6).

Finally, we found that past logging activities had the strongest effect on plant composition (Table S7, Fig. S6). They also had an effect on annelids, which was larger than the effect of soil conditions, and a small but strongly significant effect on fungi. However, they had little to no detectable effect on other groups.

	Pure soil fraction	Mixed fraction	Pure spatial fraction	Total explained variance
Plants trnL	2.4***	7.8	11.0***	21.1***
Bacteria 16S	12.7***	18.5	14.0***	45.2***
Protists 18S	2.2**	8.7	10.0***	20.8***
Fungi ITS	3.8***	4.9	5.9***	14.5***
Arthropods 18S	1.5*	2.8	2.4**	6.7***
Insects 16S	0.1	1.3	1.5**	2.9***
Annelids 18S	5.5**	5.5	15.3***	26.2***
Nematodes 18S	1.4**	1.4	2.4***	5.2***
Platyhelminthes 18S	NA	NA	NA	NA

Table 2: Fractions of variance (adjusted R², in %) explained by Canonical Redundancy Analysis for environment-only and spatial-only models. Significance: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

Discussion

We have explored the patterns of soil beta diversity in the tropical forests of French Guiana based on fifteen undisturbed 1-ha plots, as well as four disturbed plots. Distance-based analyses using Sorensen dissimilarity suggest that at our study scale, plant beta diversity is driven predominantly by geographical distance, bacteria and protist beta diversity by soil properties, while fungi, arthropod and insect beta diversity depends on both types of factors. Finally, annelid, nematode and flat worm beta diversity did not correlate with any of these factors. The observation that both geographical distance and environment play a role in explaining community assembly has already been reported for a range of taxonomic groups, either in eukaryotes or in bacteria (Cottenie, 2005; Thompson & Townsend, 2006; Martiny *et al.*, 2011). However, our results are one of the rare case studies where beta diversity has been quantified across the same sites over a broad range of taxonomic groups.

The dependence of plant beta diversity on geographical distance in tropical forests has been reported in the past, and has been presented as evidence for the importance of dispersal-limited neutral processes in shaping these ecological communities (Condit *et al.*, 2002). Likewise, the strong dependence of beta diversity on soil conditions in unicellular organisms (bacteria and protists) is in agreement with expectations (Soininen *et al.*, 2007; Ramirez *et al.*, 2014). While we could expect fungi to be primarily responsive to environmental conditions owing to their good dispersal abilities, widespread plant-fungi associations may be responsible for the observed dependence on both environmental conditions and geographical distance (Bahram *et al.*, 2013). Indeed, dispersal is hampered by host specificity, and the the distribution of host-specific fungal taxa reflects that of their plant hosts.

For insects, previous studies have reported a low betadiversity (Novotny *et al.*, 2007; Basset *et al.*, 2012). However, these studies have primarily focused on above-ground herbivores, which are known to have good dispersal ability. In contrast, we have sampled soil-dwelling insects, and thus our finding that these organisms have high beta diversity, influenced by both soil properties and dispersal limitation, does not

necessarily contradicts the results of previous publications. However, our finding is significant because it shows that spatial patterns of biodiversity in insects cannot be easily generalized across ecosystem compartments. Finally, annelids, nematodes and flatworms are represented by a limited number of OTUs (Table S1), and the lack of patterns in these groups might be due to a lack of statistical power.

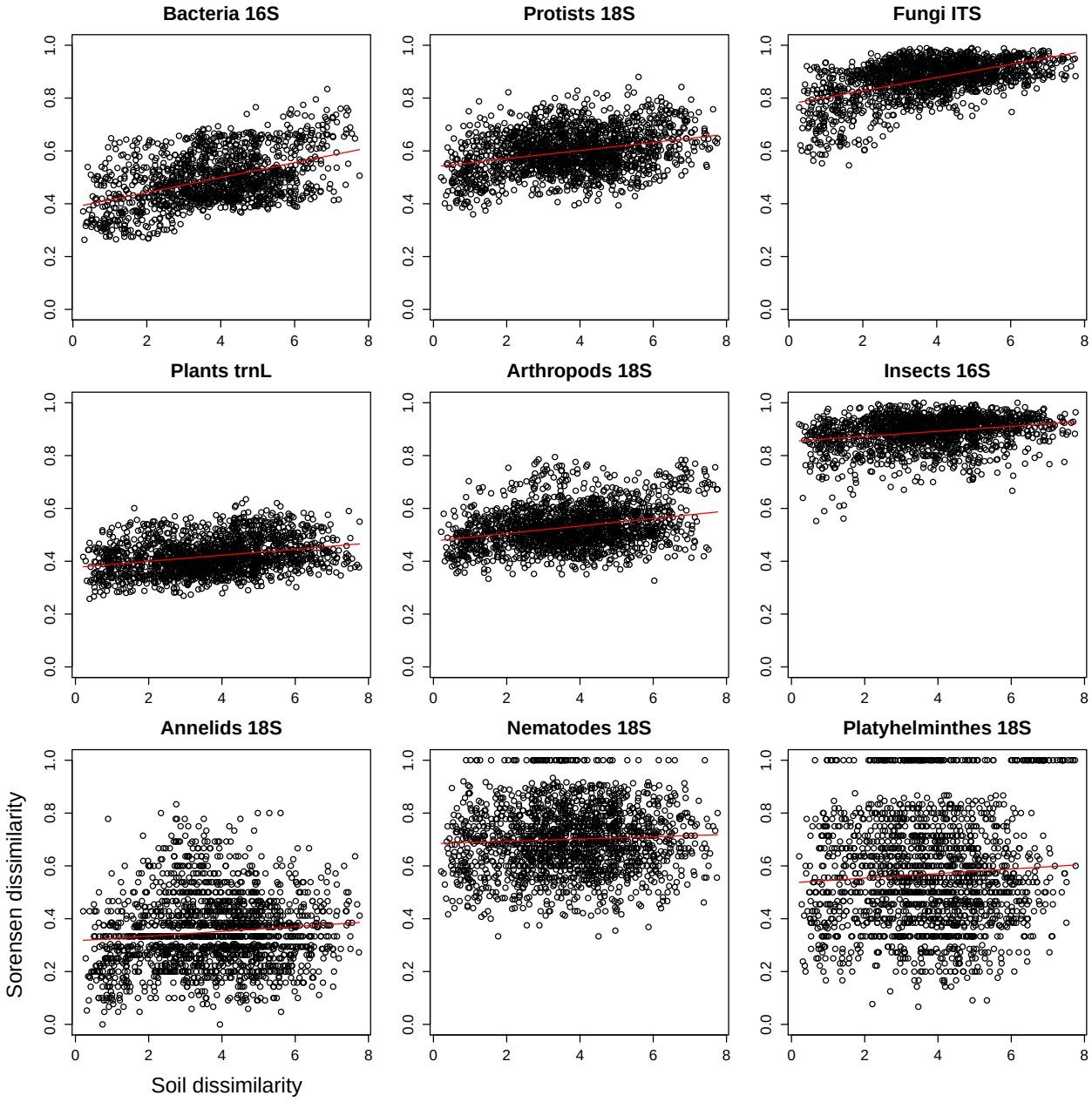

Figure 5: Occurrence-based (Sorensen) dissimilarity as a function of soil dissimilarity. Soil dissimilarity is computed as the Euclidian distance between the first four PCA axes of the measured soil variables. The red line figures the linear regression.

The RDA ('raw data') approach led to slightly different conclusions than Mantel-based correlations, and brought additional insight. A significant fraction of annelid beta diversity could be explained by spatial aggregation, while distance-based analyses using Sorensen dissimilarity did not detect any signal in this group. This is in line with the limited dispersal abilities reported for annelids in this area (Decaëns *et al.*, 2016). Spatial aggregation was also found to be an important factor explaining the spatial distribution of protists and bacteria in addition to soil properties. In contrast, we found little explanatory power for insects and arthropods. Overall, this is in line with the higher sensitivity to spatial structure reported in the literature for 'raw data' analyses (Legendre *et al.*, 2005), even though the interpretation of this spatial structure as being indicative of neutral processes is not straightforward (Smith & Lundholm, 2010). However, a potential problem in our study design is that the logarithmic geographic sampling scheme is not ideally suited to the description by PCNM variables. Because of the challenging nature of extracting DNA onsite to minimize contaminations, we could not multiply the number of sampling points, but we hope to address the issue of the sampling design for DNA-based beta diversity analyses in a forthcoming contribution.

The fit of the neutral prediction for the decay of similarity with distance was significant for all taxonomic groups except arthropods, nematodes and flat worms, but was poorer than the fit of Sorensen dissimilarity to log-distance. A possible confounding factor is that unlike the Sorensen index, the F_2 similarity measure is sensible to noise in OTU abundances, and may also be biased by uneven sampling effort among samples in DNA-based data. Overall, a decay of F_2 similarity with distance was detected in the groups for which raw-data analyses showed an effect of spatial aggregation, which is consistent with the fact that both types of analysis rely on abundance information. In particular, a decay of F_2 similarity with distance was found in annelids while none was detected using Sorensen dissimilarity, which suggests that in this group, differences between samples lie in the abundance pattern of OTUs rather in their occurrence pattern.

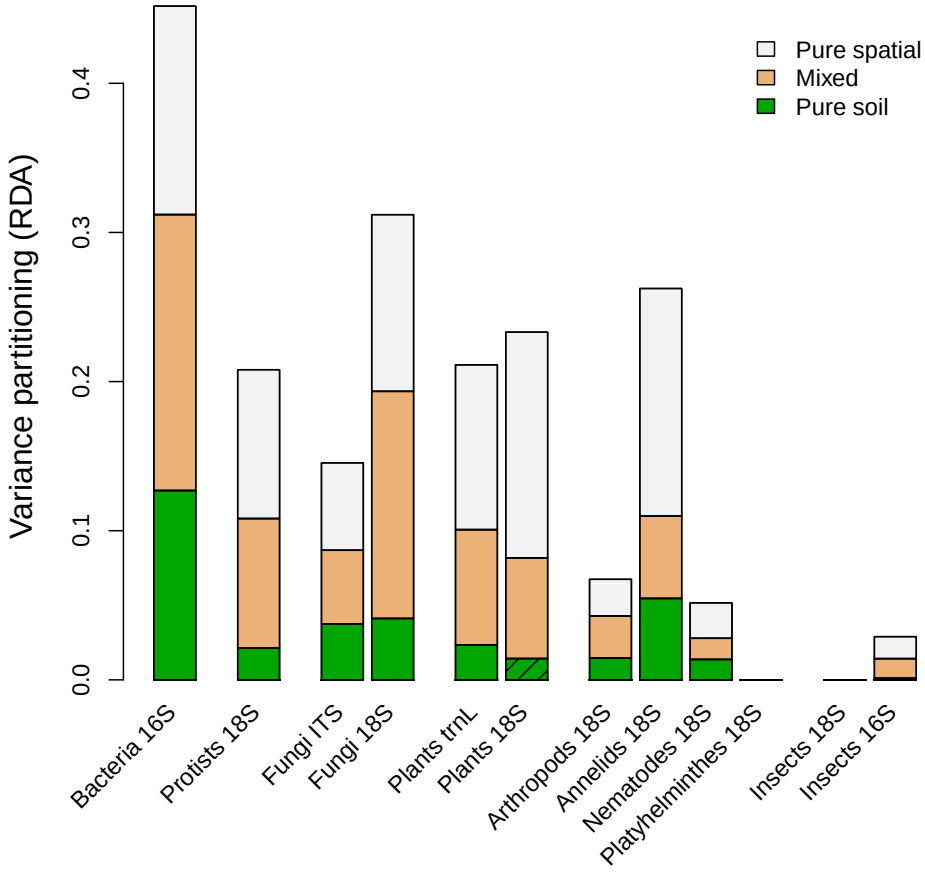

Figure 6: Variance partitioning between soil PCA axes and spatial structure (PCNM decomposition). The spatial model is the reunion of two independent PCNM decompositions, one for the Nouragues sampling sites and one for the Paracou sampling sites, plus the UTM coordinates in both groups of sites. The two PCNM decompositions are connected by a dummy variable that takes one value in Nouragues and another in Paracou. Forward variable selection is performed on soil and spatial variables before variance partitioning. Hatching indicates non-significant pure fractions.

Our estimate of 43 m for the mean dispersal distance per generation in plants was close to that measured empirically for neotropical trees (39 m; Condit *et al.*, 2002), and to that estimated by fitting the neutral similarity distance-decay prediction to tree census data (between 40 and 73 m; Condit *et al.*, 2002). Because an important part of the retrieved plant DNA originates from the tree root system, conflating the density of plant individuals with that measured for trees may be a reasonable assumption, however such estimates for the density of individuals are difficult to obtain in the other

taxonomic groups. The dispersal to diversification ratio σ^2/ν is directly measured as the ratio between the intercept b and the slope a of the linear regression of F_2 against log-distance (cf. Appendix). Lower σ^2/ν in fungi and insects reflects the low mean level of similarity between samples in these groups, which is, under a neutral model, indicative of faster diversification than dispersal, while the reverse would hold true in higher-σ^2/ν bacteria and protists (Table S6).

The challenge of measuring beta diversity is critical in conservation biology (Koleff *et al.*, 2003; Socolar *et al.*, 2017), and today the vast majority of the lowland tropical landscapes are partly deforested or at least degraded by human activities, with direct and measurable impact on biological diversity (Barlow *et al.*, 2016). The tropical forests of French Guiana have experienced low rates of forest clearance over the past decade (Hansen *et al.*, 2013) and our sampling sites can therefore be considered as undisturbed, and a baseline for the many studies focused on disturbed landscapes. Hence, in our study, the processes shaping community assembly are unlikely to be ascribed to human factors. We acknowledge that humans may have had previously unnoticed impacts on biodiversity especially on cultivated plants (Heckenberger *et al.*, 2008) or earthworms (Marichal *et al.*, 2010), however the great majority of our undisturbed sites are located far from present or historical locations of disturbances and we are therefore fairly confident that the patterns we have uncovered are contingent on natural processes. However, to better quantify the possible magnitude of human disturbances, we also studied how beta diversity is altered by intensive logging and by clear-cutting, at sites where the forest has had at least 18 years to recover. Differences in vegetation are easily noticeable on the field, and are indeed reflected in our DNA-based study. This analysis, although limited in the number of samples, also shows an effect of past logging activities on annelids and to a lesser extent on fungi; however little impact on the other components of soil biodiversity can be detected.

The current study is predicated on our assumption that DNA-based metrics of beta diversity do capture the same ecological processes as classic ones. We did find that our data capture most of the diversity present in our soil samples, as indicated by rarefaction analyses (Fig. S1). We also tested whether our results were dependent on the choice of the DNA barcode, by comparing the results obtained for the same taxonomic group with two distinct DNA barcodes (Tables S4, S5, Fig. S4, S5). In most

cases, the results appear robust to the choice of the DNA barcode, even though we detected, as expected, more signal in the specific barcodes for plants, insects and fungi than in the generic 18S barcode of lower taxonomic resolution. Overall, although we emphasize that current DNA-based inventories do not always capture the same taxonomic grain as classic surveys, this approach has the advantage of being scalable, and it should thus be appropriate for rapid biodiversity inventories, especially in fragile, or threatened ecosystems.

Acknowledgements

We thank Maxime Réjou-Méchain for fruitful discussions. This work has benefited from "*Investissement d'Avenir*" grants managed by the French *Agence Nationale de la Recherche* (CEBA, ref. ANR-10-LABX-25-01 and TULIP, ref. ANR-10-LABX-0041; ANAEE-France: ANR-11-INBS-0001), an additional ANR grant (METABAR project; PI P. Taberlet), and funds from CNRS. Work has been carried out at the CNRS Nouragues Research Station, within the Nouragues Natural Reserve, and at the CIRAD Paracou Research Station. We thank the managers of both research stations.

References

Bahram, M., Koljalg, U., Courty, P.-E., Diedhiou, A.G., Kjoller, R., Polme, S., Ryberg, M., Veldre, V. & Tedersoo, L. (2013) The distance decay of similarity in communities of ectomycorrhizal fungi in different ecosystems and scales. *Journal of Ecology*, **101**, 1335–1344.

Barlow, J., Lennox, G.D., Ferreira, J., Berenguer, E., Lees, A.C., Nally, R.M., Thomson, J.R., Ferraz, S.F. de B., Louzada, J., Oliveira, V.H.F., Parry, L., Solar, R.R. de C., Vieira, I.C.G., Aragão, L.E.O.C., Begotti, R.A., Braga, R.F., Cardoso, T.M., Jr, R.C. de O., Jr, C.M.S., Moura, N.G., Nunes, S.S., Siqueira, J.V., Pardini, R., Silveira, J.M., Vaz-de-Mello, F.Z., Veiga, R.C.S., Venturieri, A. & Gardner, T.A. (2016) Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. *Nature*.

Baselga, A., Gómez-Rodríguez, C. & Lobo, J.M. (2012) Historical legacies in world amphibian diversity revealed by the turnover and nestedness components of beta diversity. *PLoS One*, **7**, e32341.

Basset, Y., Cizek, L., Cuénoud, P., Didham, R.K., Guilhaumon, F., Missa, O., Novotny, V., Ødegaard, F., Roslin, T., Schmidl, J., Tishechkin, A.K., Winchester, N.N., Roubik, D.W., Aberlenc, H.-P., Bail, J., Barrios, H., Bridle, J.R., Castaño-Meneses, G., Corbara, B., Curletti, G., Duarte da Rocha, W., De Bakker, D., Delabie, J.H.C., Dejean, A., Fagan, L.L., Floren, A., Kitching, R.L., Medianero, E., Miller, S.E., Gama de Oliveira, E., Orivel, J., Pollet, M., Rapp, M., Ribeiro, S.P., Roisin, Y., Schmidt, J.B., Sørensen, L. & Leponce, M. (2012) Arthropod Diversity in a Tropical Forest. *Science*, **338**, 1481–1484.

Blanchet, F.G., Legendre, P. & Borcard, D. (2008) Forward selection of explanatory variables. *Ecology*, **89**, 2623–2632.

Bongers, F., Charles-Dominique, P., Forget, P.-M. & Théry, M. (2001) *Nouragues: dynamics and plant-animal interactions in a Neotropical rainforest*, Springer Science & Business Media.

Borcard, D. & Legendre, P. (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, **153**, 51–68.

Borcard, D., Legendre, P., Avois-Jacquet, C. & Tuomisto, H. (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology*, **85**, 1826–1832.

Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, **73**, 1045–1055.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

Chao, A., Chiu, C.-H. & Hsieh, T.C. (2012) Proposing a resolution to debates on diversity partitioning. *Ecology*, **93**, 2037–2051.

Chave, J. & Leigh, E.G. (2002) A spatially explicit neutral model of beta-diversity in tropical forests. *Theoretical Population Biology*, **62**, 153–168.

Clarke, L.J., Soubrier, J., Weyrich, L.S. & Cooper, A. (2014) Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. *Molecular Ecology Resources*, **14**, 1160–1170.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E. & Hubbell, S.P. (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

Cottenie, K. (2005) Integrating environmental and spatial processes in ecological community

dynamics. *Ecology Letters*, **8**, 1175–1182.

Decaëns, T., Porco, D., James, S.W., Brown, G.G., Chassany, V., Dubs, F., Dupont, L., Lapied, E., Rougerie, R. & Rossi, J.-P. (2016) DNA barcoding reveals diversity patterns of earthworm communities in remote tropical forests of French Guiana. *Soil Biology and Biochemistry*, **92**, 171–183.

Fliegerova, K., Tapio, I., Bonin, A., Mrazek, J., Callegari, M.L., Bani, P., Bayat, A., Vilkki, J., Kopečný, J. & Shingfield, K.J. (2014) Effect of DNA extraction and sample preservation method on rumen bacterial population. *Anaerobe*, **29**, 80–84.

Gaston, K. & Blackburn, T. (2008) *Pattern and process in macroecology*, John Wiley & Sons.

Gilbert, B. & Lechowicz, M.J. (2004) Neutrality, niches, and dispersal in a temperate forest understory. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 7651–7656.

Gourlet-Fleury, S., Ferry, B., Molino, J.-F., Petronelli, P. & Schmitt, L. (2004) *Experimental plots: key features*, Elsevier.

Grau, O., Peñuelas, J., Ferry, B., Freycon, V., Blanc, L., Desprez, M., Baraloto, C., Chave, J., Descroix, L. & Dourdain, A. (2017) Nutrient-cycling mechanisms other than the direct absorption from soil may control forest structure and dynamics in poor Amazonian soils. *Scientific Reports*, **7**, 45017.

Guardiola, M., Uriz, M.J., Taberlet, P., Coissac, E., Wangensteen, O.S. & Turon, X. (2015) Deep-sea, deep-sequencing: metabarcoding extracellular DNA from sediments of marine canyons. *PloS one*, **10**, e0139633.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O. & Townshend, J.R.G. (2013) High-Resolution Global Maps of 21st-Century Forest Cover Change. *Science*, **342**, 850–853.

Harrison, S., Ross, S.J. & Lawton, J.H. (1992) Beta Diversity on Geographic Gradients in Britain. *Journal of Animal Ecology*, **61**, 151–158.

Heckenberger, M.J., Russell, J.C., Fausto, C., Toney, J.R., Schmidt, M.J., Pereira, E., Franchetto, B. & Kuikuro, A. (2008) Pre-Columbian Urbanism, Anthropogenic Landscapes, and the Future of the Amazon. *Science*, **321**, 1214.

Hortal, J., Diniz-Filho, J.A.F., Bini, L.M., Rodríguez, M.Á., Baselga, A., Nogués-Bravo, D., Rangel, T.F., Hawkins, B.A. & Lobo, J.M. (2011) Ice age climate, evolutionary constraints and diversity patterns of European dung beetles. *Ecology Letters*, **14**, 741–748.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Hubbell, S.P. (2013) Tropical rain forest conservation and the twin challenges of diversity and rarity. *Ecology and Evolution*, **3**, 3263–3274.

Koleff, P., Gaston, K.J. & Lennon, J.J. (2003) Measuring beta diversity for presence–absence data. *Journal of Animal Ecology*, **72**, 367–382.

Kreft, H. & Jetz, W. (2010) A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, **37**, 2029–2053.

Lawton, J.H., Bignell, D.E., Bolton, B., Bloemers, G.F., Eggleton, P., Hammond, P.M., Hodda, M., Holt, R.D., Larsen, T.B. & Mawdsley, N.A. (1998) Biodiversity inventories, indicator taxa and effects of habitat modification in tropical forest. *Nature*, **391**, 72–76.

Legendre, P., Borcard, D. & Peres-Neto, P.R. (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecological Monographs*.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, Elsevier.

Leibold, M.A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J.M., Hoopes, M.F., Holt, R.D., Shurin, J.B., Law, R., Tilman, D., Loreau, M. & Gonzalez, A. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecology Letters*, **7**, 601–613.

Mariadassou, M., Pichon, S. & Ebert, D. (2015) Microbial ecosystems are dominated by specialist taxa. *Ecology Letters*, **18**, 974–982.

Marichal, R., Martinez, A.F., Praxedes, C., Ruiz, D., Carvajal, A.F., Oszwald, J., del Pilar Hurtado, M., Brown, G.G., Grimaldi, M., Desjardins, T. & others (2010) Invasion of Pontoscolex corethrurus (Glossoscolecidae, Oligochaeta) in landscapes of the Amazonian deforestation arc. *Applied Soil Ecology*, **46**, 443–449.

Martiny, J.B.H., Eisen, J.A., Penn, K., Allison, S.D. & Horner-Devine, M.C. (2011) Drivers of bacterial beta-diversity depend on spatial scale. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 7850–7854.

Nekola, J.C. & White, P.S. (1999) The distance decay of similarity in biogeography and ecology. *Journal of Biogeography*, **26**, 867–878.

Novotny, V., Miller, S.E., Hulcr, J., Drew, R.A.I., Basset, Y., Janda, M., Setliff, G.P., Darrow, K., Stewart, A.J.A., Auga, J., Isua, B., Molem, K., Manumbor, M., Tamtiai, E., Mogia, M. & Weiblen, G.D. (2007) Low beta diversity of herbivorous insects in tropical forests. *Nature*, **448**, 692–695.

Pace, N.R. (1997) A molecular view of microbial diversity and the biosphere. *Science*, **276**, 734–740.

Ramirez, K.S., Leff, J.W., Barberan, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H. & Fierer, N. (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 9.

Rosenzweig, M.L. (1995) *Species diversity in space and time*, Cambridge University Press.

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2009) The map equation. *The European Physical Journal Special Topics*, **178**, 13–23.

Schuldt, A., Wubet, T., Buscot, F., Staab, M., Assmann, T., Böhnke-Kammerlander, M., Both, S., Erfmeier, A., Klein, A.-M., Ma, K., Pietsch, K., Schultze, S., Wirth, C., Zhang, J., Zumstein, P. & Bruelheide, H. (2015) Multitrophic diversity in a biodiverse forest is highly nonlinear across spatial scales. *Nature Communications*.

Siles, J.A. & Margesin, R. (2016) Abundance and Diversity of Bacterial, Archaeal, and Fungal Communities Along an Altitudinal Gradient in Alpine Forest Soils: What Are the Driving Factors? *Microbial Ecology*, **72**, 207–220.

Smith, T.W. & Lundholm, J.T. (2010) Variation partitioning as a tool to distinguish between niche and neutral processes. *Ecography*, **33**, 648–655.

Socolar, J.B., Gilroy, J.J., Kunin, W.E. & Edwards, D.P. (2017) How Should Beta-Diversity Inform Biodiversity Conservation? *Trends in Ecology & Evolution*, **31**, 67–80.

Soininen, J., McDonald, R. & Hillebrand, H. (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

ter Steege, H., Nigel, C.A., Sabatier, D., Baraloto, C., Salomao, R.P., Guevara, J.E., Phillips, O.L., Castilho, C.V., Magnusson, W.E., Molino, J.F., Monteagudo, A., Vargas, P.N., Montero, J.C., Feldpausch, T.R., Coronado, E.N.H., Killeen, T.J., Mostacedo, B., Vasquez, R., Assis, R.L., Terborgh, J., Wittmann, F., Andrade, A., Laurance, W.F., Laurance, S.G.W., Marimon, B.S.,

Marimon, B.H., Vieira, I.C.G., Amaral, I.L., Brienen, R., Castellanos, H., Lopez, D.C., Duivenvoorden, J.F., Mogollon, H.F., Matos, F.D.D., Davila, N., Garcia-Villacorta, R., Diaz, P.R.S., Costa, F., Emilio, T., Levis, C., Schietti, J., Souza, P., Alonso, A., Dallmeier, F., Montoya, A.J.D., Piedade, M.T.F., Araujo-Murakami, A., Arroyo, L., Gribel, R., Fine, P.V.A., Peres, C.A., Toledo, M., Gerardo, A.A.C., Baker, T.R., Ceron, C., Engel, J., Henkel, T.W., Maas, P., Petronelli, P., Stropp, J., Zartman, C.E., Daly, D., Neill, D., Silveira, M., Paredes, M.R., Chave, J., Lima, D.D., Jorgensen, P.M., Fuentes, A., Schongart, J., Valverde, F.C., Di Fiore, A., Jimenez, E.M., Mora, M.C.P., Phillips, J.F., Rivas, G., van Andel, T.R., von Hildebrand, P., Hoffman, B., Zent, E.L., Malhi, Y., Prieto, A., Rudas, A., Ruschell, A.R., Silva, N., Vos, V., Zent, S., Oliveira, A.A., Schutz, A.C., Gonzales, T., Nascimento, M.T., Ramirez-Angulo, H., Sierra, R., Tirado, M., Medina, M.N.U., van der Heijden, G., Vela, C.I.A., Torre, E.V., Vriesendorp, C., Wang, O., Young, K.R., Baider, C., Balslev, H., Ferreira, C., Mesones, I., Torres-Lezama, A., Giraldo, L.E.U., Zagt, R., Alexiades, M.N., Hernandez, L., Huamantupa-Chuquimaco, I., Milliken, W., Cuenca, W.P., Pauletto, D., Sandoval, E.V., Gamarra, L.V., Dexter, K.G., Feeley, K., Lopez-Gonzalez, G. & Silman, M.R. (2013) Hyperdominance in the Amazonian Tree Flora. *Science*, **342**, 325–+.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant molecular biology*, **17**, 1105–1109.

Thompson, R. & Townsend, C. (2006) A truce with neutral theory: local deterministic factors, species traits and dispersal limitation together determine patterns of diversity in stream invertebrates. *Journal of Animal Ecology*, **75**, 476–484.

Vincent, J.B., Weiblen, G.D. & May, G. (2016) Host associations and beta diversity of fungal endophyte communities in New Guinea rainforest trees. *Molecular Ecology*, **25**, 825–841.

Whittaker, R.H. (1972) Evolution and Measurement of Species Diversity. *Taxon*, **21**, 213–251.

Whittaker, R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California. *Ecological Monographs*, **30**, 279–338.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., Schilling, V., Schimann, H., Sommeria-Klein, G. & Taberlet, P. (2016) Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology and Biochemistry*, **96**, 16–19.

Zinger, L., Taberlet, P., Schimann, H., Bonin, A., Boyer, F., De Barba, M., Gaucher, P., Gielly, L., Giguet-Covex, C., Iribar, A., Réjou-Méchain, M., Rayé, G., Rioux, D., Schilling, V., Tymen, B., Viers, J., Zouiten, C., Thuiller, W., Coissac, E. & Chave, J. (2017) Soil community assembly varies across body sizes in a tropical forest. *BioRxiv*.

Supplementary Information

	# OTUs	# Reads
Plants trnL	776	5,142,400
Plants 18S	71	366,646
Bacteria 16S	11,380	3,863,620
Protists 18S	295	240,223
Fungi ITS	4,312	2,151,746
Fungi 18S	386	832,153
Arthropods 18S	342	463,057
Insects 16S	3,497	1,331,880
Insects 18S	70	185,446
Annelids 18S	18	145,044
Nematodes 18S	81	10,672
Platyhelminthes 18S	32	15,619

Table S1: Number of OTUs and read count per taxonomic group.

	pH	C_{tot}	N_{tot}	P_2O_5	Clay	Silt	Sand	Al	Fe	Mg	Mn	K	Ca
Unit	None	(g/kg)		(mg /kg)	(%)	(%)	(%)	(cmol+/kg)					
Inselberg summit	4.9	30.5	1.8	< 5.0	27.5	4.6	67.9	2.1	0.081	0.27	0.011	0.097	0.08
PP-F21	4.9	29.1	1.9	5.8	33.6	4.0	62.6	2.6	0.068	0.13	0.011	0.088	0.09
PP-H20	4.3	34.3	2.1	7.3	48.1	4.5	47.4	2.5	0.144	0.46	0.018	0.127	0.26
PP-H21	4.8	31.1	2.2	6.0	51.0	4.3	44.7	2.3	0.063	0.37	0.017	0.082	0.25
GP-L11	5.0	35.7	3.0	5.3	73.3	12.8	13.9	1.2	0.006	0.67	0.125	0.113	1.48
GP-L12	4.6	37.0	3.1	7.3	71.8	16.8	11.4	1.6	0.006	0.43	0.215	0.112	0.75
GP-O13	4.3	32.6	2.1	6.8	71.7	12.8	15.6	3.5	0.030	0.51	0.070	0.114	0.75
GP-Liana	5.2	27.6	2.6	7.8	52.0	30.4	17.6	0.4	0.005	1.65	0.252	0.143	3.64
Balanfois-1	3.9	41.8	2.9	< 5.0	78.7	7.6	13.8	3.5	0.041	0.22	0.028	0.084	0.35
Balanfois-2	3.9	40.7	2.9	< 5.0	79.6	6.0	14.4	3.3	0.046	0.31	0.025	0.087	0.27
Parare-5	4.4	35.9	2.6	10.3	64.5	12.3	23.3	2.8	0.056	0.59	0.020	0.114	0.21
Parare-6	4.0	38.1	2.5	11.3	55.5	19.2	25.4	4.0	0.058	0.24	0.019	0.116	0.16
Paracou-06.3	5.0	19.2	1.2	6.5	13.4	7.2	79.4	1.2	0.043	0.22	<0.010	0.078	0.10
Paracou-06.4	4.9	20.1	1.3	8.3	12.2	7.0	80.9	1.2	0.035	0.16	<0.010	0.058	0.11
Paracou-11.1	5.0	20.1	1.2	6.8	16.3	8.0	75.7	1.6	0.053	0.23	<0.010	0.080	0.14
Paracou-12.1	*4.6*	*27.8*	*1.7*	*< 5.0*	*27.0*	*7.6*	*65.5*	*2.3*	*0.124*	*0.24*	*<0.010*	*0.080*	*0.07*
Paracou-12.2	*4.6*	*20.6*	*1.2*	*7.3*	*16.5*	*6.5*	*77.1*	*1.9*	*0.113*	*0.21*	*<0.010*	*0.079*	*0.16*
Arbocel-7.3	*4.6*	*30.7*	*1.9*	*7.0*	*22.0*	*10.0*	*68.0*	*1.6*	*0.143*	*0.32*	*<0.010*	*0.085*	*0.13*
Arbocel-7.4	*4.6*	*30.3*	*1.9*	*< 5.0*	*24.6*	*10.7*	*64.7*	*1.7*	*0.147*	*0.29*	*<0.010*	*0.076*	*0.16*

Table S2: Mean soil variables in all nineteen 1-ha plots. Each value is the average of four separate measurements, each made on twenty pooled soil samples. Al, Fe, Mg, Mn, K, and Ca concentrations are expressed in cmol of positive charges per kg. Values in italics correspond to disturbed plots.

	pH	C_{tot}	N_{tot}	P_2O_5	Cly	Silt	Al	Fe	Mg	Mn	K	Ca
pH	1	-0.59	-0.41	-0.02	-0.53	0.09	**-0.82**	-0.17	0.27	0.19	-0.01	0.32
C_{tot}		1	**0.91**	-0.04	**0.74**	0.11	0.60	-0.03	0.05	0.08	0.33	-0.01
N_{tot}			1	-0.01	**0.83**	0.36	0.35	-0.31	0.33	0.41	0.47	0.31
P_2O_5				1	-0.01	0.27	0.08	-0.13	0.12	0.14	0.30	0.07
Clay					1	0.24	0.48	-0.45	0.27	0.41	0.51	0.29
Silt						1	-0.22	-0.39	**0.76**	0.67	0.52	**0.76**
Al							1	0.15	-0.38	-0.38	0.04	-0.45
Fe								1	-0.34	-0.56	-0.23	-0.48
Mg									1	**0.70**	0.60	**0.91**
Mn										1	0.56	**0.81**
K											1	0.59
Ca												1
VIF	5.0	35.9	44.8	1.5	12.5	4.5	8.3	3.3	7.2	5.0	3.2	10.8

Table S3: Correlation coefficients between soil variables in the fifteen undisturbed plots. Bold font indicates correlation coefficients above 0.70. Variance Inflation Factors (VIF) are computed as the diagonal elements of the inverse correlation matrix.

	Mean $D_{Sorensen}$	Geographical distance			Soil		
		r_{dist}	$r_{dist,part}$	slope$_{dist}$	r_{soil}	$r_{soil,part}$	slope$_{soil}$
Plants trnL	0.42	0.65***	0.61***	0.038	0.29***	0.06	0.011
Plants 18S	0.50	0.22**	0.15*	0.022	0.24***	0.17*	0.016
Fungi ITS	0.87	0.43***	0.29***	0.029	0.54***	0.45***	0.025
Fungi 18S	0.45	0.31***	0.20*	0.022	0.39***	0.31***	0.019
Insects 16S	0.89	0.23***	0.16**	0.013	0.25***	0.18**	0.010
Insects 18S	0.57	0.07	0.05	0.008	0.06	0.03	0.005

Table S4: Linear regression of taxonomic dissimilarity (Sorensen index) against soil and geographical distance: comparison between barcodes within taxonomic groups (cf. Table 2). $r_{dist}, r_{soil}, r_{dist,part}, r_{soil,part}$ are the simple and partial Pearson's correlation coefficients. Significance was assessed using Mantel tests: *** for p < 0.001; ** for 0.001 < p < 0.01; * for 0.01 < p < 0.05.

	Pure soil fraction	Mixed fraction	Pure spatial fraction	Total explained variance
Plants trnL	2.4***	7.8	11.0***	21.1***
Plants 18S	1.4	6.8	15.1***	23.3***
Fungi ITS	3.8***	4.9	5.9***	14.5***
Fungi 18S	4.1***	15.2	11.8***	31.2***
Insects 16S	0.1	1.3	1.5**	2.9***
Insects 18S	NA	NA	NA	NA

Table S5: Fractions of variance (adjusted R^2, in %) explained by Canonical Redundancy Analysis for environment-only and spatial-only models: comparison between barcodes within taxonomic groups (cf. Table 3). Significance: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

	R	$1/a$ $(\times 10^3)$	b/a	$\log_{10}\sigma^2/\nu$
Plants trnL	0.26***	0.55	24	-20
Plants 18S	0.16**	0.24	33	-28
Bacteria 16S	0.22***	6.8	48	-42
Protists 18S	0.33***	0.075	46	-40
Fungi ITS	0.14***	2.3	13	-11
Fungi 18S	0.23***	0.76	31	-26
Arthropods 18S	0.07	0.71	43	-37
Insects 16S	0.08**	1.1	15	-13
Insects 18S	0.09	0.11	33	-28
Annelids 18S	0.25***	0.061	30	-26
Nematodes 18S	0.06	1.0	53	-46
Platyhelminthes 18S	0.10	0.13	34	-30

Table S6: Fitting the neutral prediction for the decay of taxonomic similarity with distance (Chave & Leigh, 2002). $F_2(A,B) = \sum_{s=1}^{S} p_s^A p_s^B$, where p_s^A is the proportion of species s in sample A and p_s^B that in sample B, is regressed against the log-transformed geographical distance r between samples (expressed in meters), as $F_2(r) = a \ln r + b$: R is the correlation coefficient; ***,** and * denote the significance assessed by Mantel test ($p < 0.001$, $p < 0.01$ and $p < 0.05$, respectively); and σ^2/ν (expressed in square meters) is the ratio between the variance σ^2 of the dispersal kernel and the neutral speciation probability ν.

	Logging only	Pure logging fraction	Mixed fraction	Pure soil fraction	Total explained variance
Plants trnL	12.9***	4.0***	9.0	3.5**	16.4***
Plants 18S	10.9***	3.1*	7.8	2.3	13.1***
Bacteria 16S	11.9***	1.9*	10.0	18.2***	30.1***
Protists 18S	4.8**	1.1	3.6	4.8**	9.6**
Fungi ITS	4.3***	1.6***	2.6	5.2***	9.4***
Fungi 18S	7.6***	4.7***	2.9	8.6***	16***
Arthropods 18S	1.7	NA	NA	NA	NA
Insects 16S	1.6*	NA	NA	NA	NA
Insects 18S	3.4	NA	NA	NA	NA
Annelids 18S	6.0*	6.4*	-0.3	5.4*	11.4**
Nematodes 18S	1.9*	NA	NA	NA	NA
Platyhelminthes 18S	0.6	NA	NA	NA	NA

Table S7: Fractions of variance (adjusted R^2, in %) explained by Canonical Redundancy Analysis for logging intensity and for soil conditions. Significance: *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

Taxonomic group	Selected spatial variables			
Bacteria 16S	UTMN.Nouragues	UTME.Nouragues	Nouragues.MEM.1	Nouragues.MEM.5
Protists 18S	UTMN.Nouragues	UTME.Nouragues	Nouragues.MEM.1	Paracou.MEM.1
Plants trnL	UTMN.Nouragues	UTME.Nouragues	Nouragues.MEM.1	Nouragues.MEM.5
	Paracou.Nouragues	UTME.Paracou		
Plants 18S	*UTMN.Nouragues*	*UTME.Nouragues*	*Nouragues.MEM.1*	*Nouragues.MEM.4*
	Nouragues.MEM.5	*Nouragues.MEM.8*	*Nouragues.MEM.12*	*Nouragues.MEM.15*
	Nouragues.MEM.16			
Fungi ITS	UTMN.Nouragues	UTME.Nouragues	Nouragues.MEM.1	
Fungi 18S	*UTMN.Nouragues*	*UTME.Nouragues*	*Nouragues.MEM.1*	
Arthropods 18S	UTMN.Nouragues	UTME.Nouragues		
Annelids 18S	UTMN.Nouragues	Nouragues.MEM.1	Paracou.MEM.3	
Nematodes 18S	UTMN.Nouragues			
Platyhelminthes 18S	No selected model			
Insects 16S	UTME.Nouragues	Paracou.Nouragues	Nouragues.MEM.2	
	Nouragues.MEM.11	Nouragues.MEM.15		
Insects 18S	*No selected model*			

Table S8: Selected spatial models after forward variable selection. Selection is applied on the following variables: UTM coordinates in Nouragues and Paracou ('UTMN.Nouragues', 'UTME. Nouragues', 'UTMN.Paracou', 'UTME.Paracou'), the dummy variable connecting Nouragues and Paracou sites ('Paracou.Nouragues'), and PCNM variables in Nouragues ('Nouragues.MEM.1' to 'Nouragues.MEM.17') and Paracou ('Paracou.MEM.1' to 'Paracou.MEM.7'), which represent different possible patterns of spatial autocorrelation.

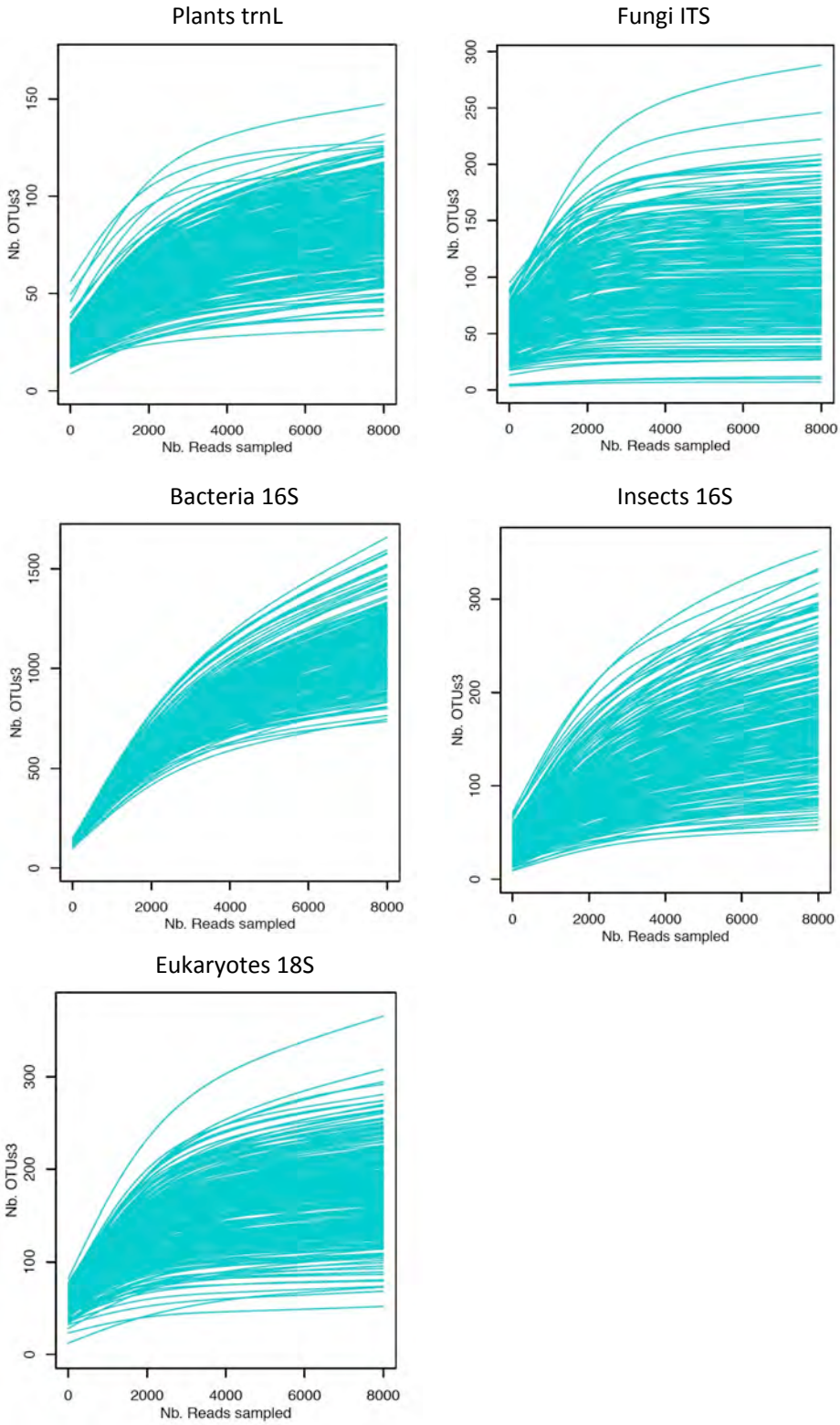

Figure S1: Rarefaction analyses. In each sample and for each barcode, we sampled with replacement between 1 and 8,000 reads, and plotted the corresponding number of OTUs (one curve per sample and per barcode).

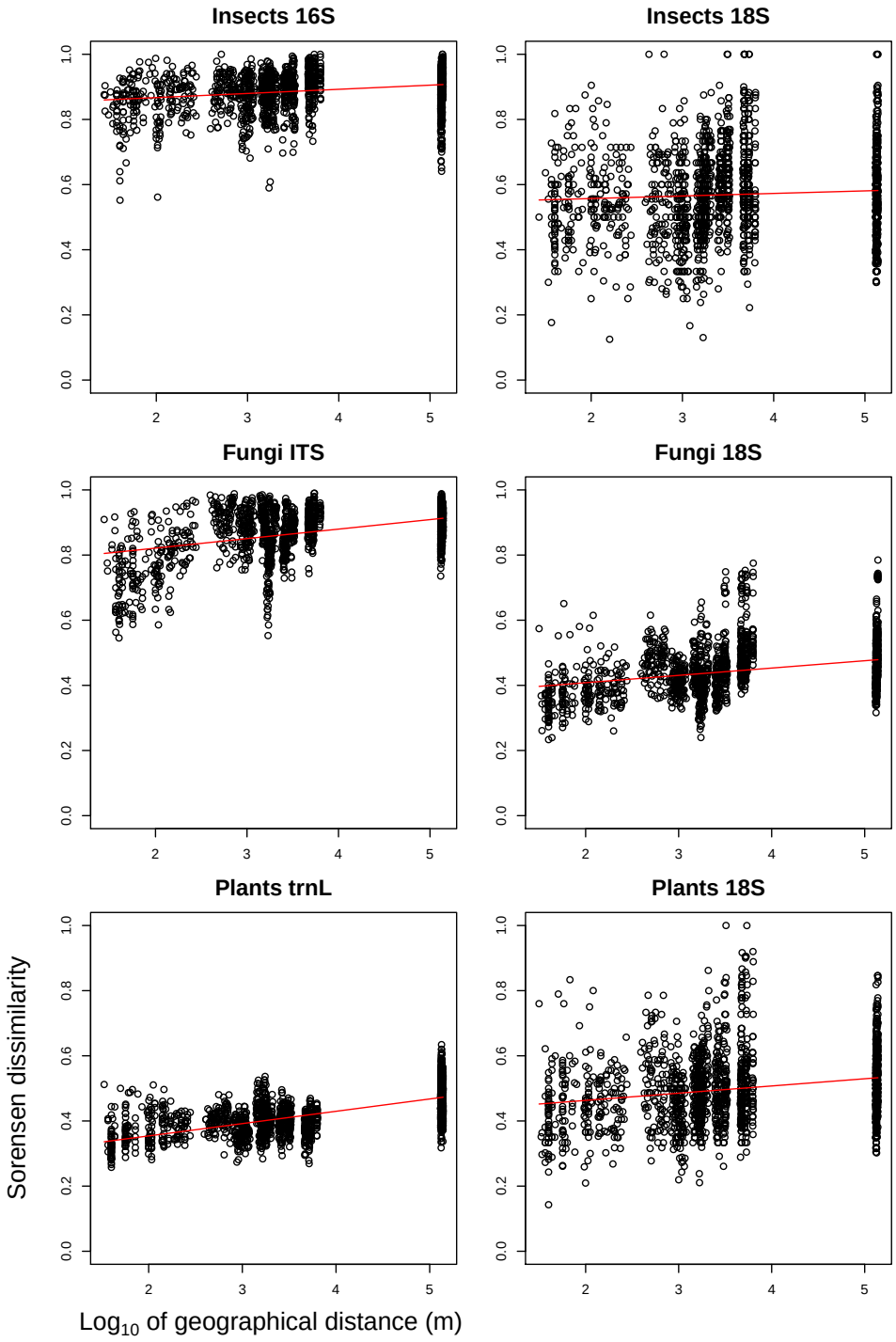

Figure S2: Occurrence-based (Sorensen) dissimilarity as a function of log-distance; comparison between barcodes within taxonomic groups (cf. Fig. 4). The red line figures the linear regression.

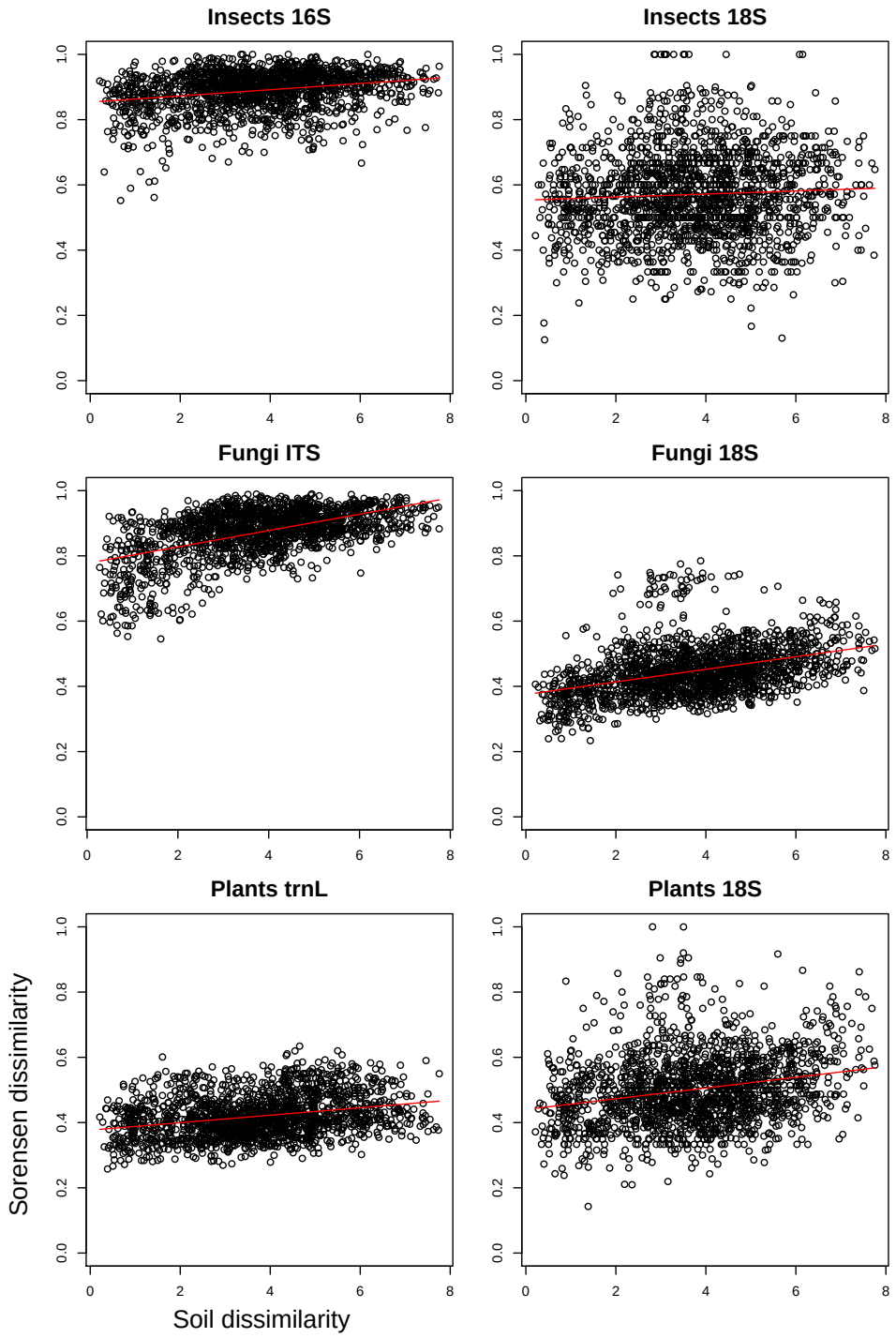

Figure S3: Occurrence-based (Sorensen) dissimilarity as a function of log-distance; comparison between barcodes within taxonomic groups (cf. Fig. 5)**.** The red line figures the linear regression.

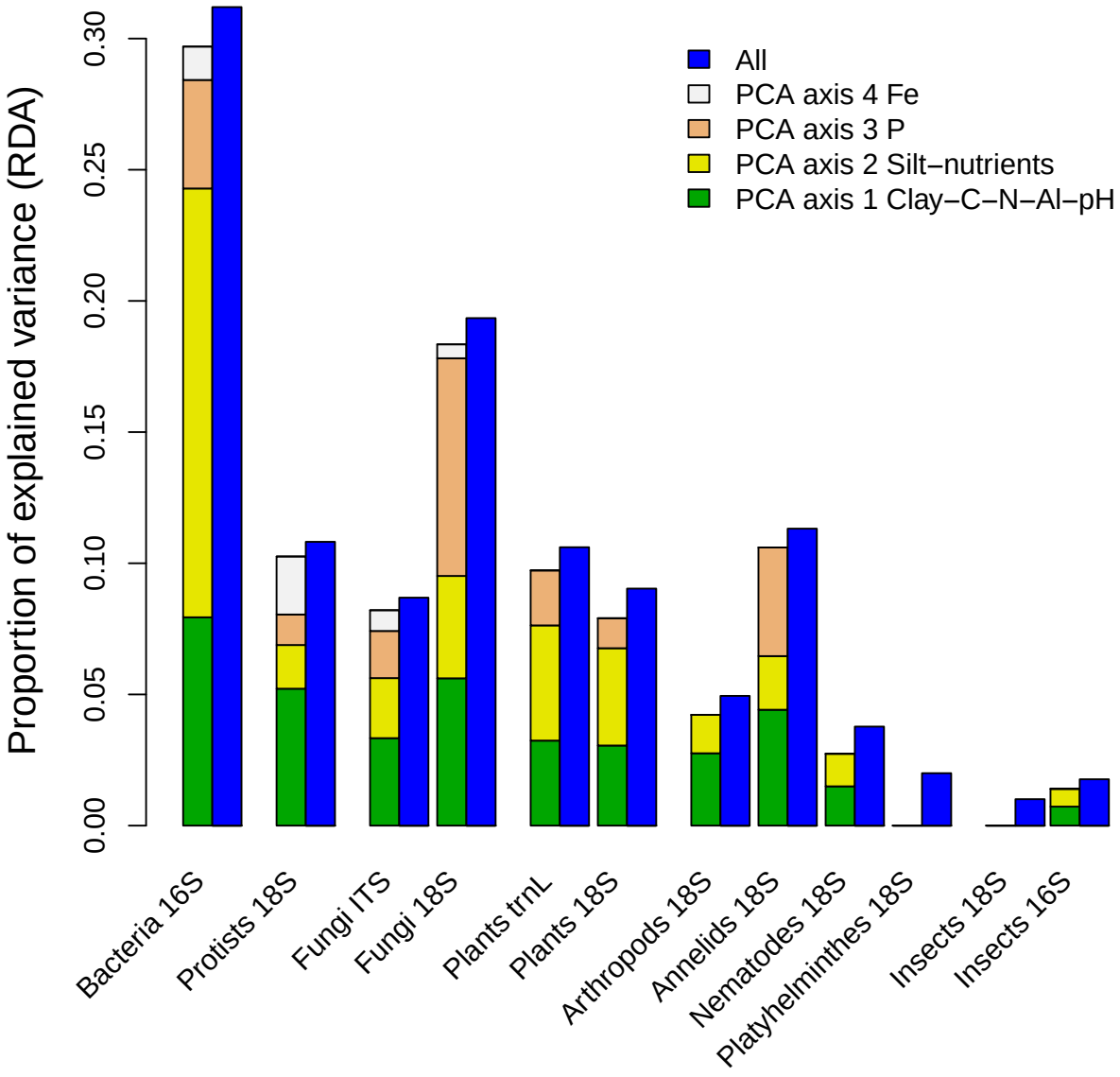

Figure S4: Proportion of variance explained by RDA for each soil variable. Only soil variables selected by forward variable selection (out of the six initial variables) are shown.

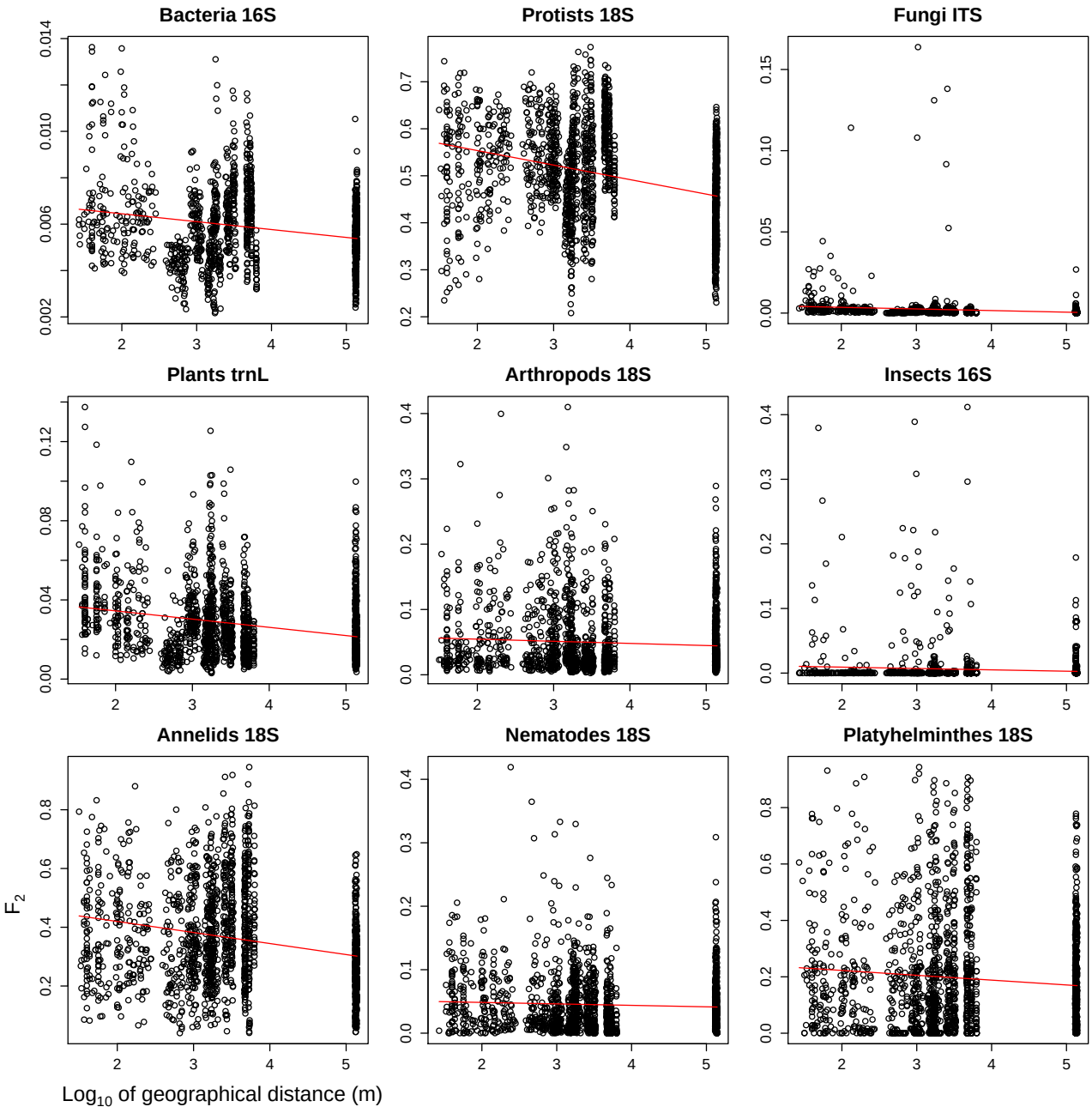

Figure S5: Testing the neutral prediction for the decay of taxonomic similarity with geographical distance: F_2 similarity as a function of log-distance. Red line denotes linear regression. Note that y-scale varies across taxonomic groups.

Appendix: Fitting the neutral prediction for the distance-decay of similarity

Chave & Leigh (2002) derived an analytical prediction for the decay of taxonomic similarity with distance in a continuous spatially explicit version of Hubbell's neutral model of biodiversity, where individuals have spatial density ρ and where dispersal follows a radially symmetric Gaussian probability density $P(r) = (1/2\pi\sigma^2)\exp(-r^2/2\sigma^2)$ as a function of distance r. They predict that the stationary probability $F_2(r)$ that two randomly selected individuals distant of r belong to the same species decreases as:

$$F_2(r) \simeq \frac{2K_0\left(\frac{r\sqrt{2\nu}}{2\sigma}\right)}{\ln\frac{1}{\nu} + 2\rho\pi\sigma^2}$$

provided that r is larger than σ. In our dataset, the minimal value taken by r is 40 m, which is approximately equal to the mean dispersal distance per generation for tropical trees (Condit *et al.*, 2002). Because the mean dispersal distance per generation is $\sqrt{2}\sigma$ in the model, and because trees are likely to be the organisms with the largest σ in our study, the assumption that r is larger than σ can be regarded here as reasonable.

The parameter ν it is the speciation probability in the underlying neutral dynamics, i.e. the probability for a newly born individual to belong to a new species. This parameter characterizes the regional species diversity for a given population size. The function $K_0(\tilde{r})$ is the modified Bessel function of the second kind and of zeroth order, that can be approximated as $K_0(\tilde{r}) \simeq -\ln(\tilde{r}/2) - \gamma$ if $\tilde{r} \ll 1$, where γ is Euler's constant. Because $\nu \ll 1$, this approximation can be regarded as valid in our case where $\tilde{r} = r\sqrt{2\nu}/\sigma$. The probability $F_2(r)$ then becomes:

$$F_2(r) \simeq -\frac{2\ln\left(\frac{r\sqrt{2\nu}}{2\sigma}\right) + 2\gamma}{\ln\frac{1}{\nu} + \rho\pi\sigma^2}$$

In empirical data, the probability $F_2(A,B)$ that a randomly selected individual in site A belongs to the same species as a randomly selected individual in site B can be measured as $F_2(A,B) = \sum_{s=1}^{S} p_s^A p_s^B$, where p_s^A is the proportion of species s in site A and p_s^B that in site B, and S is the total number of species in both sites (Chave & Leigh, 2002). We can thus compute the quantity $F_2(A,B)$ for every pair of sampling points and performed the linear regression $F_2 = -a\ln(r) + b$, where r is the distance between two sampling points (in meters). By identification with the model's prediction, we obtain:

$$\begin{cases} \frac{b}{a} = \ln\left(\frac{\sqrt{2\nu}}{2\sigma}\right) + \gamma \\ \frac{1}{a} = \rho\pi\sigma^2 + \frac{1}{2}\ln\frac{1}{\nu} \end{cases}$$

The first equation provides the value of σ as a function of ν, a and b as $\sigma^2/\nu = 1/\sqrt{2}\exp(-b/a + \gamma)$, while the sum of the two equations provides the value of σ as a function of ρ, a and b as the solution of: $\rho\pi\sigma^2 - \ln(2\sigma) + (b+1)/a + \gamma + \ln(2)/2$.

Chapter 2

Inferring neutral biodiversity parameters using environmental DNA data sets

Guilhem Sommeria-Klein[1], Lucie Zinger[1], Pierre Taberlet[2], Eric Coissac[2], Jérôme Chave[1]

As published in *Scientific Reports*,
Volume 6, 2016

[1] Université Toulouse 3 Paul Sabatier, CNRS, UMR 5174 Laboratoire Evolution et Diversité Biologique, F-31062 Toulouse, France.

[2] Université Grenoble Alpes, CNRS, UMR 5553 Laboratoire d'Ecologie Alpine, F-38000 Grenoble, France.

Chapter outline

The distribution of species abundances has been one of the most intensively studied patterns in ecology, and the use of environmental DNA could dramatically increase our ability to measure empirical species abundance distributions over a wide range of taxa. However, DNA-based abundance measurements are noisy and difficult to interpret compared to classical censuses of individual organisms. This chapter discusses to which extent and under which conditions the whole species abundance distribution may nevertheless remain informative. The bias on the estimates of Hubbell's neutral parameters is taken as a measure of this loss of information. Indeed, Hubbell's neutral theory has been the first to propose a realistic quantitative prediction for this pattern on mechanistic grounds. Even though the underlying assumptions have been much debated, this model remains fundamental as a null model against which non-neutral effects can be contrasted. It also provides a characterization of species abundance distributions based on two parameters, one measuring intrinsic diversity, and the other measuring the connectivity between local and regional communities through migration. The problem is addressed by simulating several plausible sources of bias, based on literature and on assumptions backed by a benchmark dataset.

Abstract

The DNA present in the environment is a unique and increasingly exploited source of information for conducting fast and standardized biodiversity assessments for any type of organisms. The datasets resulting from these surveys are however rarely compared to the quantitative predictions of biodiversity models. In this study, we simulate neutral taxa-abundance datasets, and add simulated noise typical of DNA-based biodiversity surveys. The resulting noisy taxa abundances are used to assess whether the two parameters of Hubbell's neutral theory of biodiversity can still be estimated. We find that parameters can be inferred provided that PCR noise on taxa abundances does not exceed a certain threshold. However, inference is seriously biased by the presence of artifactual taxa. The uneven contribution of organisms to environmental DNA owing to size differences and barcode copy number variability does not impede neutral parameter inference, provided that the number of sequence reads used for inference is smaller than the number of effectively sampled individuals. Hence, estimating neutral parameters from DNA-based taxa abundance patterns is possible but requires some caution. In studies that include empirical noise assessments, our comprehensive simulation benchmark provides objective criteria to evaluate the robustness of neutral parameter inference.

Introduction

The observation of biodiversity patterns such as the diversity, relative abundance and spatial distribution of organisms underpins much of ecological theory (Brown, 1995; Rosenzweig, 1995; Hubbell, 2001). Yet empirical measurements of these patterns are noisy. In all cases, some taxa are counted more effectively than others, and error is generated by misidentification. A major question is whether this noise is significant enough to undermine comparisons between empirical measurements and models (Hilborn & Mangel, 1997; Legendre & Legendre, 2012). This issue has recently taken on new significance following the advent of DNA-based biodiversity exploration methods, which are developing fast and hold the promise of rapid, repeatable and comprehensive biodiversity measurements (Bik *et al.*, 2012; Taberlet *et al.*, 2012). Yet they are also less direct than classic biodiversity surveys and entail poorly assessed noise sources. In this study, we ask how the parameter estimates of Hubbell's neutral theory, one of the most prominent quantitative biodiversity models of the last decade (Hubbell, 2001; Etienne & Alonso, 2007; Rosindell *et al.*, 2012), are affected by noise in taxa-abundance datasets. We focus on the type of noise generated in DNA-based surveys, and specifically in DNA metabarcoding surveys (see below; Taberlet *et al.*, 2012), currently the most popular method for environmental DNA analysis. Nevertheless, our results can apply more generally.

DNA metabarcoding is a multi-taxa extension of the DNA-based identification of single specimen from tissue samples using a universal DNA-barcode sequence (Hebert *et al.*, 2003). It consists in amplifying a short DNA barcode by PCR from the DNA extracted from an environmental sample (e.g. soil, water, bulk sample of organisms), and sequencing the product by high-throughput sequencing. This method is not restricted to the detection of known taxa and hence allows for comprehensive biodiversity measurement. DNA metabarcoding was initially developed to study bacterial communities (Giovannoni *et al.*, 1990; Huber *et al.*, 2007; Roesch *et al.*, 2007; Zinger *et al.*, 2012), but has since been extended to many other groups including archaea (Schleper *et al.*, 2005) and eukaryotic clades (e.g. plants, earthworms, insects,

fungi; Bienert *et al.*, 2012; Yoccoz *et al.*, 2012; Yu *et al.*, 2012; Tedersoo *et al.*, 2014). It is hence now possible to study patterns of diversity across all domains of life (Ramirez *et al.*, 2014; Tedersoo *et al.*, 2015). However, DNA metabarcoding observations have seldom been compared to the predictions of biodiversity models (Hubbell, 2001; Ricklefs, 2004).

Over the past decade, the neutral theory of biodiversity has represented a significant advance in interpreting empirical biodiversity patterns within an ecological guild (Hubbell, 2001; Etienne & Alonso, 2007; Rosindell *et al.*, 2012). Hubbell's neutral model is simple, easily generates biodiversity patterns, allows for exact maximum-likelihood parameter inference from taxa-abundance distributions, and neutral predictions on taxa-abundance distributions compare well with empirical surveys (Etienne, 2005; Etienne & Alonso, 2005; Jabot & Chave, 2009). In Hubbell's model, sites vacated by the death of an individual are replaced by the offspring of local individuals or by immigrants. Birth, death and immigration all occur irrespective of the taxon the organism belongs to (neutrality hypothesis). Immigrants are drawn from a much larger (regional) pool of individuals, and the addition of new taxa in the regional pool is made possible by (rare) speciation events. Hubbell's model has two parameters: θ describes the taxon diversity of the regional pool, and m is the immigration rate from the regional pool into the sampled community (see Supplementary Note 1).

The predictions of Hubbell's neutral model have so far been primarily compared to integrative patterns obtained for macroorganisms using classic census data, such as the abundance distribution of tropical forest trees (Hubbell, 2001). Some studies have also applied neutral models to environmental DNA data to interpret the composition of microbial communities. Sloan *et al.* (2006, 2007) and (Woodcock *et al.*, 2007) Woodcock *et al.* (2007) developed a continuous approximation to Hubbell's model adapted to large-sized bacterial populations. They focused on estimating the rate of immigration into the local community independently of assumptions on the regional pool of taxa, by comparing taxa occurrence in multiple samples (Sloan *et al.*, 2006; Drakare & Liess, 2010; Ostman *et al.*, 2010; Ayarza & Erijman, 2011; Roguet *et al.*, 2015) or by measuring the turnover of taxa over time (Ofiteru *et al.*, 2010). The composition of many microbial communities was found to be compatible with

stochastic immigration of taxa of equivalent fitness from a regional pool, at odds with the classic assumption that deterministic niche sorting explains the assemblage of microbial communities (Baas Becking, 1934; Fenchel & Finlay, 2004). Another approach is to simultaneously estimate the diversity and immigration parameters by fitting the taxa-abundance distribution, as it has been commonly done for classic censuses of macroorganisms. Dumbrell *et al.* (2010) and Lee *et al.* (2013) did so on fungal and bacterial DNA data using maximum-likelihood parameter inference based on the exact Etienne sampling formulas (Etienne, 2005, 2007, 2009), while Harris *et al*. (2015) followed a Bayesian approach inspired by the field of machine learning.

Most DNA-based studies comparing empirical abundance patterns to the predictions of neutral models have been limited by the poor detectability of rare taxa owing to the methods used (Sanger sequencing, DGGE, t-RFLP, ARISA). High-throughput sequencing now allows for improved sampling and provides better quality data. Nevertheless, metabarcoding data are not directly comparable with classic census data owing to both experimental and biological factors. First, both PCR amplification and sequencing produce artifacts. During the PCR amplification, DNA polymerase makes mistakes when replicating DNA strands, at a rate that depends on enzyme types. DNA strands suffer further damage during the high-temperature denaturation step (Pienaar *et al.*, 2006; Quince *et al.*, 2011; Degnan & Ochman, 2012). Furthermore, Illumina sequencing generates between 10^{-3} and 10^{-2} errors per base pair (Ross *et al.*, 2013). Clustering algorithms are used to cluster the reads displaying errors with respect to the original sequence into a single Molecular Operational Taxonomic Unit (MOTU; Sipos *et al.*, 2010; Coissac *et al.*, 2012; Mahe *et al.*, 2014). While these approaches strongly reduce the number of artifacts in the data, they do not exclude artifactual MOTUs that are more difficult to detect (e.g. chimerical fragments, highly degraded sequences). Second, unbalanced PCR amplification and sequencing among taxa distorts the relative abundances of MOTUs (Sipos *et al.*, 2007; Amend *et al.*, 2010; Aird *et al.*, 2011; Nguyen *et al.*, 2015). Third, relative abundances are further biased by noise sources inherent to the use of DNA barcodes, such as the strong variability of the barcode copy number among taxa (Kembel *et al.*, 2012; Weber & Pawlowski, 2013). This problem is even more serious for multicellular organisms because the read count should also depend on cell abundance. Abundances are further biased by the variable

rate of DNA release into the environment through excreted, sloughed or decaying material (Andersen *et al.*, 2012; Maruyama *et al.*, 2014; Klymus *et al.*, 2015).

In this paper, we conduct simulations to address how the sources of uncertainty mentioned above may distort parameter estimates in Hubbell's neutral theory, and we discuss the conceptual differences between individual-based and environmental DNA approaches to the measurement of biodiversity. We ask the following questions: 1) what is the effect of artifactual MOTUs and abundance noise on estimating the neutral diversity parameter? 2) Can we use the same approach for multicellular as for unicellular organisms? 3) What are the effects of the different noise sources on neutral parameter inference when accounting for dispersal limitation?

Methods

1. Sampling from Hubbell's neutral model

We generated samples of J individuals following the stationary taxa-abundance distribution of Hubbell's neutral model. The immigration from the regional pool of diversity parameter θ into the sampled community can be either characterized by the immigration rate m or by the normalized immigration parameter $I = \frac{m}{1-m}(J-1)$ that does not depend on the sample size J and is thus invariant by sampling. If $m \ll 1$, I is approximated by the product Jm, noted $N_T m$ in Sloan *et al.* (2006, 2007).

We first assumed no dispersal limitation (i.e. $m = 1$). We generated a sample by running J times the following algorithm parameterized by θ: at step j, draw individual $j+1$ from a new taxon with probability $\theta/(j+\theta)$, or draw one of the j individuals already present and add an individual $j+1$ of the same taxon. This algorithm, due to Hoppe (1984), partitions J individuals into a random number T of taxa according to the Ewens distribution of parameter θ (Ewens, 1972).

We then generated samples from a dispersal-limited neutral community using the two-step procedure provided in Etienne (2005) which partitions J individuals into a random number T of taxa. First, we run J times Hoppe's algorithm as described above but with parameter I, so as to partition the J individuals into A immigrating ancestors. Second, we run A times the algorithm with parameter θ, so as to partition the A immigrating ancestor into T taxa, thus taking into account the taxa-abundance distribution in the regional pool. Finally, we assign the J individuals to the taxonomic identity of their immigrating ancestor.

We generated samples of $J = 10^5$ individuals. We explored a realistic range of parameter values: θ in [1, 500] and m in [0.001, 1].

2. Simulating noise in DNA sequence reads: experimental noise

We simulated the DNA metabarcoding procedure by sampling N sequence reads from the relative taxa abundances of the neutral model, possibly after modifying the relative abundances according to simulated noise sources (see below). We present the results obtained for the value $N = 10^4$, a typical number of Illumina sequence reads for one environmental sample.

In order to test the effect of misidentification bias on neutral parameter inference, we added artifactual MOTUs to the data, while keeping the number of reads constant. We assumed that each true MOTU with a read abundance r generates a random number of artifactual MOTUs, drawn from a multinomial distribution with weight r. We added either singletons, or MOTUs with larger read abundances. We obtained an example of artifactual MOTUs with realistic abundance structure from a benchmark experiment (see below and Supplementary Methods). Drawing on these empirical data, we simulated read abundances in the following way: each artifactual MOTU was assumed to have an abundance of 1 read if $r < 50$, or an abundance x if $r \geq 50$, where x lies between 1 and $r/50$ with a probability density $p(x) = \frac{1}{\log(r/50)x}$.

Molecular experimental procedures introduce biases also in read abundances, because the efficiency of PCR amplification and sequencing is variable across MOTUs. For instance, PCR amplification is less efficient if PCR priming sites differ from the primer sequence (Sipos *et al.*, 2007), or if the barcode sequence is too long or GC-rich (Aird *et al.*, 2011). As a result, the read abundance distribution of MOTUs is noised with respect to the DNA barcode abundance distribution in the sample. We assumed that the noise takes the form of a lognormally distributed multiplicative noise on relative abundances, with mean 1 and log standard deviation σ_{log}. This choice is parsimonious because this noise is predominantly due to PCR (Aird *et al.*, 2011), and the multiplicative amplification of DNA strands by PCR generates a multiplicative noise on abundances. This multiplicative noise can be further assumed to result from the product of random independent variables and thus to be lognormally distributed by virtue of the central limit theorem. We tested the effect of noise intensity σ_{log} on

neutral parameter inference. For completeness, we also tested the effect of an additive Gaussian noise of standard deviation σ_{add} on MOTUs relative abundances, for different σ_{add} values. This type of noise can be regarded as simulating the noise generated in the sequencing step.

To illustrate our modelling choices with empirical data, we produced a benchmark dataset obtained by mixing the DNA of 16 plant species in known quantities. The experiment and its results are detailed in the Supplementary Methods. After following standard data curation protocols, we found that the dataset contained 33% of artifactual MOTUs and displayed a lognormally distributed multiplicative noise on relative abundances of log standard deviation $\sigma_{log} = 1.2$. We reported these values on the figures as examples of realistic noise intensities.

3. Simulating noise in DNA sequence reads: 'biological' noise

Irrespective of experimental noise, variability in the number of barcode copies per individual may cause bias in the interpretation of read abundances. For bacteria (16S rDNA) or protists (18S rDNA), barcode copy number variability in nuclear DNA is an important contribution to abundance noise (Kembel *et al.*, 2012; Weber & Pawlowski, 2013): Kembel *et al.* (2012) found that the barcode copy number of the 16S rDNA gene follows a zero-truncated Poisson distribution of parameter $\lambda = 4$ across a range of bacterial clades. For multicellular eukaryotes, organellic barcodes are typically used, and they similarly display variable copy numbers per cell across taxa and tissue types. To assess this issue, we tested how a zero-truncated Poisson-distributed multiplicative noise affects neutral parameter inference, for various values of the parameter λ. The intensity of this noise is measured by the coefficient of variation (i.e., standard deviation over mean) of the zero-truncated Poisson distribution. Since it reaches a maximum at $\lambda = 1.8$, noise intensity is maximal for this value.

For multicellular organisms, the variability in the number of barcode copies per individual is further amplified because the number of cells may vary vastly across individuals, owing to body-size differences. We simulated size differences between

individuals following a simple and generic approach. As in O'Dwyer *et al.* (2009), we assumed that all individuals, irrespective of the taxon they belong to, grow in size over time at a constant rate g from an initial number of cells n_0 at birth, and die at a constant rate d. The stationary probability density $p_{ind}(n)$ of having a number n of cells for a randomly chosen individual is given by the solution of the von Foerster equation (O'Dwyer *et al.*, 2009): $p_{ind}(n) = \frac{d}{g} e^{-\frac{d}{g}(n-n_0)}$ (see Supplementary Note 2). We used this distribution to draw a number n of cells between n_0 and infinity for each individual, and modified the MOTUs relative abundances accordingly. Note that we simulated size differences between individuals and not between taxa, which would have been akin to simulating a multiplicative noise on taxa abundances as above. We tested the effect on neutral parameter inference for a range of values of $\frac{g}{dn_0} + 1$, the ratio of the mean cell number $\frac{g}{d} + n_0$ divided by the initial cell number n_0. Noise intensity is measured by the coefficient of variation $1/(1 + \frac{d}{g} n_0)$ of the probability density $p_{ind}(n)$. It is bounded by 1 for $\frac{g}{dn_0} \gg 1$, which corresponds to the case of taxa spanning large ranges of body sizes, such as trees or vertebrates.

Organisms may be entirely contained in the environmental sample if they are sufficiently small, or when DNA is extracted from a mixture of directly sampled live organisms, such as insects from a light trap (bulk samples; Yu *et al.*, 2012). However, in most cases, only small fractions of these organisms are sampled (e.g. roots, pollen, seeds, spores, faeces, and different secretion types), or even only extracellular DNA resulting from cell death and subsequent destruction of cell structure (Levy-Booth *et al.*, 2007; Taberlet *et al.*, 2012). Thus, the abundance distribution of environmental DNA also depends on the kinetics of DNA release and degradation in the environment. We assumed that this dynamics is fast with respect to changes in community composition, so that the 'stock' of environmental DNA is in a steady state. Under this assumption, the rate of DNA release through the death of organisms is roughly proportional to the total number of cells of the currently living individuals. In addition, the rate of environmental DNA release by a living organism reflects its metabolic rate and we assumed it to scale as the power 3/4 of body mass (or cell number), as predicted by the metabolic theory of ecology[61]. DNA degradation rate was assumed

uniform across individuals. Even though we focus here on multicellular organisms, unicellular organisms do excrete DNA material and differ in metabolic rates as well.

Based on the assumptions of the previous paragraph, we simulated the abundance distribution of environmental DNA as follows. We (1) generated a neutral sample of individuals, (2) assigned a number of cells n between n_0 and infinity to each individual as above, (3) counted a first contribution dn of each individual to the stock of environmental DNA, with d the death rate, (4) and counted a second contribution $r_0 n^{\frac{3}{4}}$ of each individual to the stock of environmental DNA, with r_0 the rate of DNA release for a hypothetical one-cell individual. Thus, environmental DNA abundance per individual is proportional to $n + \frac{r_0}{d} n^{\frac{3}{4}}$ rather than n. We tested the effect on neutral parameter inference by varying r_0/d, the parameter controlling the relative contribution of living and dead organisms to environmental DNA.

4. Estimating the neutral model parameters from the taxa-abundance distribution

We estimated the parameters of Hubbell's neutral model by maximum-likelihood inference from the simulated taxa-abundance distribution for a number of simulated noise sources. To test the influence of noise, we compared the estimated parameter values $\hat{\theta}$ and $\hat{I}$ with the values of θ and I used to generate the initial samples of individuals. For each set of parameters and noise intensity, we generated 100 simulated samples. We reported the mean and standard deviation of the relative biases $(\hat{\theta} - \theta)/\theta$ and $\log_{10}(\hat{I}/I)$ over the 100 realizations.

In the absence of dispersal limitation, the Ewens distribution permits the inference of θ by likelihood maximization. The maximum-likelihood estimator of θ, hereafter referred to as the Ewens estimator, is implicitly given by $T = \sum_{j=0}^{J-1} \frac{\hat{\theta}}{\hat{\theta}+j}$ as a function of the number T of taxa and the number J of individuals (Ewens, 1972). In the dispersal-limited case, the Etienne distribution provides an exact likelihood expression for the simultaneous inference of θ and I (Etienne, 2005), as implemented in the

software Tetame (Jabot *et al.*, 2008). As noted previously in the literature, the likelihood landscape of the Etienne formula often displays two local maxima (Etienne *et al.*, 2006; Jabot & Chave, 2009). To find the true parameter values, we first estimated θ using the Ewens estimator, and selected the local maximum with the θ estimate closest to the value yielded by the Ewens estimator. Prior to these analyses, we tested the performances of both estimators on unbiased neutral data depending on parameter values and sample size (see Supplementary Note 3).

In typical environmental DNA data, the number *J* of individuals in the sample is unknown. As already done in previous studies (Lee *et al.*, 2013), we used the number of sequence reads as an effective number of individuals. This is possible owing to a mathematical property of the Ewens and Etienne distributions: both distributions are invariant by sampling without replacement (Etienne & Alonso, 2005), hence maximum-likelihood inference yields the same results on any random sample from the community, and on any random subsample from an initial sample (up to a possible bias in the estimator). As a consequence, read abundances can be used for neutral parameter inference, as long as the reads can be regarded as forming a subsample without replacement of the initial individuals. This assumption is however not always verified in empirical data (see Discussion). The invariance property of Etienne distribution only holds if the distribution is expressed as a function of *I*, therefore we used here the immigration parameter *I* instead of *m* for the purpose of inference. In the following, *m* always refers to the value in the initial sample of *J* individuals.

In the absence of dispersal limitation, θ can also be estimated from the slope of the ranked log-abundance curve, a method that has the advantage of being independent of *J*. Indeed, the logarithm of $\mathbb{E}[P_i]$, the expected relative abundance of the i^{th} most abundant taxon, is given by: $\log(\mathbb{E}[P_i]) = -\log\theta - i\log(1 + 1/\theta)$ (Ewens & Tavaré, 1997). For simulated abundance noise, we estimated θ using this method in addition to Ewens estimator. We restricted the linear regression to the linear domain of the ranked log-abundance curve. We also compared the performance of both inference methods in the absence of simulated noise for samples of 10^2, 10^3, 10^4 and 10^5 sequence reads and for initial samples of individuals of different sizes (see Supplementary Note 4).

Results

We first included artifactual MOTUs in a simulated sample and tested the effect on estimating the diversity parameter θ of the neutral model without dispersal limitation. The relative bias $(\hat{\theta} - \theta)/\theta$ increased with the proportion of artifactual MOTUs, first linearly and then faster than linearly (Fig. 1a-b). It did not depend on the initial θ value or on the read abundance of the introduced artifactual MOTUs. The standard deviation of $\hat{\theta}$ was not modified by the presence of artifactual MOTUs.

Next, we simulated PCR noise, modelled as a lognormally distributed multiplicative noise with log standard deviation σ_{log}. This noise had no effect on the inference of the θ parameter below a threshold $\sigma_{log,th}$. For $\sigma_{log} > \sigma_{log,th}$, θ was underestimated. The value of $\sigma_{log,th}$ decreased with increasing θ but remained of the order of 1 for θ between 1 and 500 ($\sigma_{log,th} \approx 5$ for $\theta = 1$ and $\sigma_{log,th} \approx 0.5$ for $\theta = 500$; see Fig. 1c-d). We also applied an additive Gaussian noise of standard deviation σ_{add} to the relative abundances. This type of noise introduced a bias in $\hat{\theta}$ for values of σ_{add} at least one order of magnitude larger than the relative abundance of the least abundant MOTUs (Supplementary Fig. S1). Neither type of noise affected the standard deviation of $\hat{\theta}$ (Fig. 1, Supplementary Fig. S1). These results held both in maximum-likelihood inference and when using linear regression on the ranked log-abundance.

We then simulated the variability in barcode copy number by applying a multiplicative noise distributed according to a zero-truncated Poisson distribution. This type of noise had no effect on θ inference, even for the maximum noise intensity at $\lambda = 1.8$ (Fig. 1e-f). We accounted for body size differences by assuming a steadily growing cell number n over the course of an individual's life, and by varying the ratio $g/dn_0 + 1$ of the mean number of cells $g/d + n_0$ divided by the initial number of cells n_0. We found that this ratio had no effect on the mean and standard deviation of $\hat{\theta}$, even at large values (Fig. 1g-h). We also tested the effect of assigning an environmental DNA mass proportional to $n + \frac{r_0}{d} n^{\frac{3}{4}}$ to individuals (where n is the cell number) to reflect the joint effect of mortality (n term) and cellular turnover ($n^{\frac{3}{4}}$ term, proportional

to metabolic rate). We did not find any effect on θ inference even for large values of r_0/d (Supplementary Fig. S1).

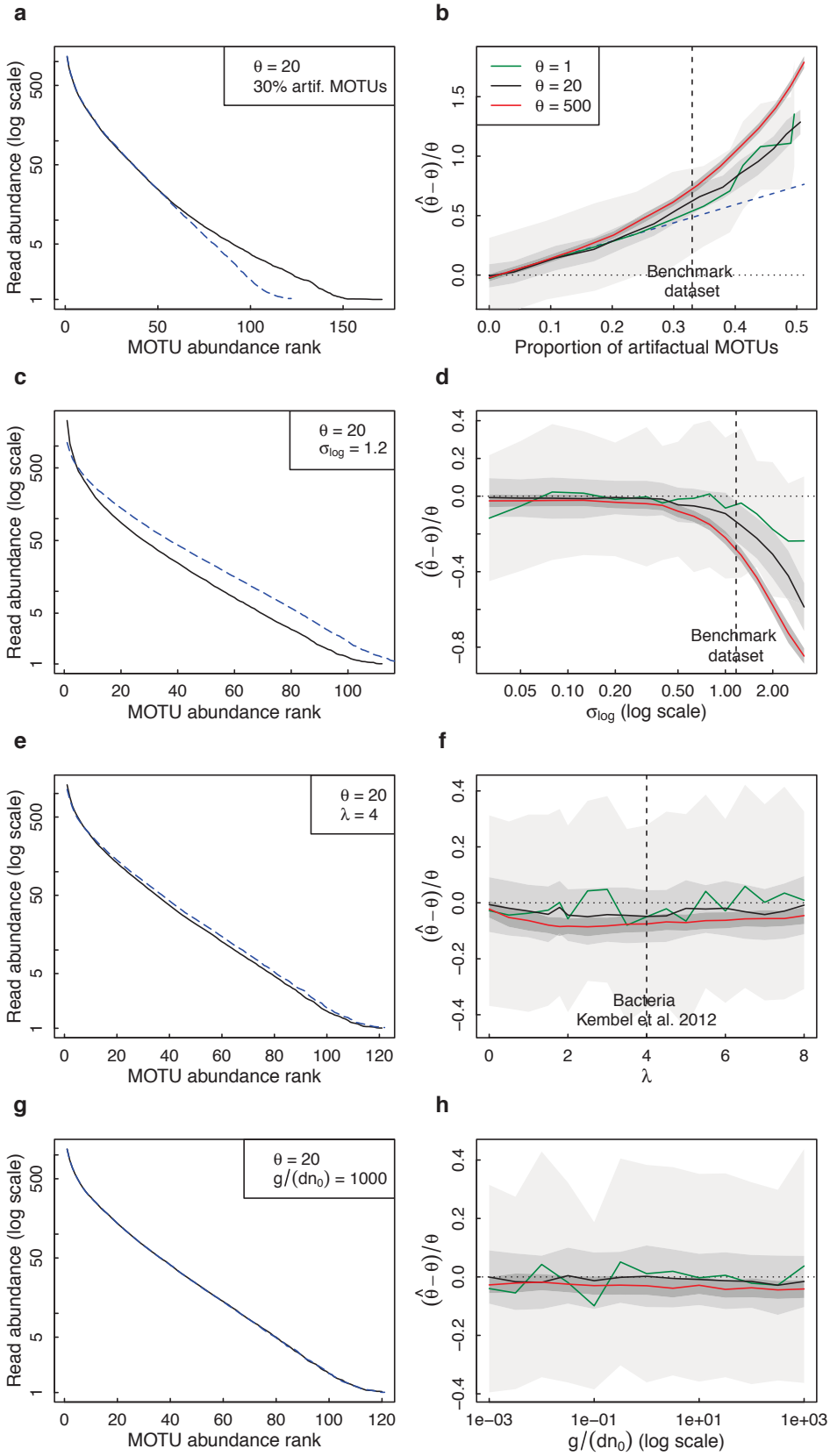

Figure 1: Neutral parameter inference without dispersal limitation. **Left panels**: mean MOTU rank abundance distributions over 100 realizations for $\theta = 20$ in a 10^4-read sample, without (dashed blue line) and with (black line) simulated noise: (a) 30% artifactual MOTUs added (as measured in benchmark dataset), (c) multiplicative lognormal noise of log standard deviation $\sigma_{\log} = 1.2$ (as measured in benchmark dataset), (e) multiplicative zero-truncated Poisson noise simulating barcode copy number variability (Poisson parameter $\lambda = 4$; cf. Kembel *et al.* 2012), and (g) size structure among individuals, for a ratio $\frac{g}{dn_0} = 1000$ (mean body mass over birth mass). **Right panels**: mean and standard deviation over 100 realizations of the relative bias on the θ estimate in a 10^4-read sample, for $\theta = 1$ (green), $\theta = 20$ (black) and $\theta = 500$ (red), as a function of (b) the proportion of artifactual MOTUs (dashed blue line underlines the linear dependence), (d) the lognormal noise intensity $\sigma_{\log}$, (f) the Poisson parameter λ, and (h) the ratio $g/(dn_0)\,\frac{g}{dn_0}$.

Finally, we replicated the analysis in the presence of dispersal limitation (i.e. assuming that $m < 1$). We found that the dispersal-limited maximum-likelihood estimator can be strongly biased even in the absence of simulated noise when dispersal limitation is too strong or too weak, especially for large θ values (see Supplementary Note 3). Therefore, we limited ourselves to parameter values that could be well estimated in the absence of simulated noise. Provided the immigration rate is large enough ($m > 0.1$), the relative bias $(\hat{\theta} - \theta)/\theta$ depended on the proportion of artifactual MOTUs similarly to the $m = 1$ $\mathrm{m} = 1$ case. For lower values of m, the dependence of $(\hat{\theta} - \theta)/\theta$ on the proportion of artifactual MOTUs was even stronger (Fig. 2a-b). The relative bias $\log_{10}(\hat{I}/I)$ on the normalized immigration parameter increased linearly with the proportion of artifactual MOTUs. Applying a lognormal multiplicative noise of log standard deviation σ_{log} on MOTUs relative abundances did not bias the estimation of (θ, I) below a noise threshold $\sigma_{log,th}$ identical to the one found without dispersal limitation. The threshold $\sigma_{log,th}$ decreased only slightly with decreasing m value. Above $\sigma_{log,th}$, θ was underestimated and I overestimated (Fig. 2c-d). Applying an additive Gaussian noise of standard deviation σ_{add} to the relative abundances introduced a bias for values of σ_{add} larger than the relative abundance of the least abundant MOTUs (Supplementary Fig. S2). A multiplicative noise distributed according to a zero-truncated Poisson had no influence on the parameter estimates (Fig. 2e-f), and likewise an exponentially distributed number of cells still had no effect on parameter inference in the dispersal-limited case (Fig. 2g-h, Supplementary Fig. S2).

Discussion

Although they provide an unparalleled amount of information, biodiversity studies based on environmental DNA also have limitations. One of them is that the abundance of sequence reads corresponding to a given molecular taxonomic unit does not necessarily reflect the true population abundance of the corresponding taxon. Our analysis offers a quantitative assessment of the importance of this issue in attempting to relate environmental DNA datasets with theoretical model predictions.

Our goal was to assess when amplicon-based DNA read abundance data can offer biological insights into the predictions of Hubbell's neutral theory. We selected Hubbell's model over other models predicting taxa-abundance distributions because it incorporates a number of key features for any biodiversity model such as demographic stochasticity and dispersal limitation (Vellend, 2010). Estimating the parameters θ and m of the neutral model is useful in interpreting biodiversity patterns even if the community is not governed by purely neutral mechanisms (Jabot *et al.*, 2008). Indeed, θ is closely related to Fisher's biodiversity index, and is an unbiased index of biodiversity, while m quantifies how the local sample is connected to its surroundings. We simulated taxa abundance datasets from a neutral model and added noise to them using a range of plausible noise types and intensities. We showed that the parameters θ and I could still be reliably estimated by maximum likelihood inference from the simulated sequence reads, provided that artifactual MOTUs are rare, and that lognormal noise on relative read abundances is below a log standard deviation threshold that depends on θ. We also showed that under our modelling assumptions, neutral inference is unbiased for assemblages of multicellular organisms and for variable barcode copy numbers. Finally, we found that the noise terms had a similar effect on parameter inference when fitting the one-parameter version of the model (without dispersal limitation) and when fitting Hubbell's dispersal-limited model.

One of the major differences between environmental DNA surveys and classic biodiversity surveys is that the number of sampled individuals is usually not measured.

Yet, most biodiversity measures assume the knowledge of the organisms' sample size. To solve this problem, we assumed in our simulations that the number of reads is several times smaller than the number of effectively sampled individuals: $N = 10^4$ sequence reads for $J = 10^5$ initial individuals. Under this assumption, sequence reads may be seen as a random subsample of the individuals, and because the maximum-likelihood approach of the neutral theory relies on sampling formulas that are invariant under subsampling, it follows that the inference on reads is unbiased (see Supplementary Note 4). Generating a larger number of individuals did not alter our results but was computationally prohibitive with our algorithm.

The assumption that the number of sampled individuals exceeds that of sequence reads is reasonable for prokaryotes (Whitman *et al.*, 1998) and microorganisms in general, but is unrealistic for larger organisms. One empirical method to test whether the sequencing data meet the requirement for neutral maximum-likelihood inference is to take a smaller subsample of reads and check that the parameter estimates are unchanged. If not, one should decrease sample size until stability is achieved (see Supplementary Note 4). If environmental DNA data do not consist of a discrete number of reads, as is the case in t-RFLP and ARISA, an arbitrarily set sample size may be used (Lee *et al.*, 2013). The number of individuals can also be estimated empirically, as in Woodcock *et al.* (2007) or Dumbrell *et al.* (2010). In the neutral model without dispersal limitation, a more straightforward approach is to infer θ from the slope of the ranked log-abundance distribution, but this requires an arbitrary delimitation of the linear domain of the curve, and it is reliable only if the read sample is large enough and contains a large enough taxonomic diversity. A general rule is that the sampling scheme should be suited to the size and spatial density of the target organisms: for large organisms, multiple spatially distributed environmental samples should be pooled so as to sample a sufficiently large number of individuals. For instance, capturing the abundance distribution of plant taxa from soil DNA samples requires pooling a sufficient number of soil samples over a sufficiently large area.

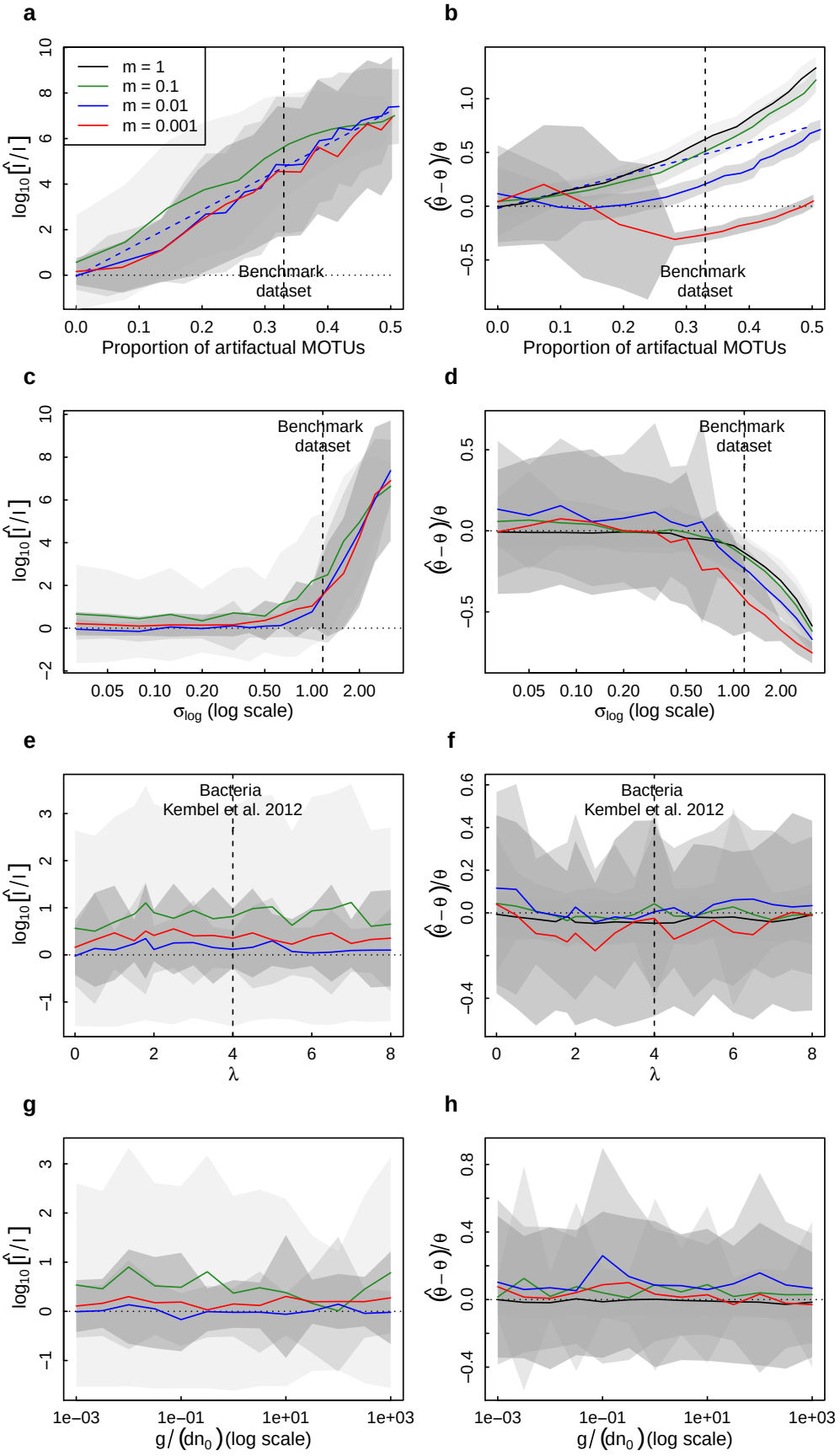
a
b
c
d
e
f
g
h
m = 1
m = 0.1
m = 0.01
m = 0.001
Benchmark dataset
Proportion of artifactual MOTUs
σlog (log scale)
Bacteria
Kembel et al. 2012
λ
g/(dn0) (log scale)

Figure 2: Neutral parameter inference in the presence of dispersal limitation. We simulated a 10^4-read sample and computed the mean and standard deviation over 100 realizations of $(\hat{\theta} - \theta)/\theta$ and $\log_{10}(\hat{I}/I)$. Results are plotted for $\theta = 20$ and for $m = 1$ (black), $m = 0.1$ (green), $m = 0.01$ (blue) and $m = 0.001$ (red). **Panels a-b:** variation with the proportion of artifactual MOTUs (dashed blue line underlines the linear dependence). **Panels c-d:** variation with the log standard deviation $\sigma_{\log}$ of a multiplicative lognormal noise on relative abundances. **Panels e-f:** variation with the parameter λ of a multiplicative zero-truncated Poisson noise. **Panels g-h:** variation with the body size ratio $g/(dn_0)$ $\frac{g}{dn_0}$.

When accounting for dispersal limitation, a single sample of sequence reads does not always provide enough information to reliably infer both θ and I from the taxa-abundance distribution, even in the absence of additional noise source. The maximum-likelihood estimator may be strongly biased when the immigration rate into the local community is either too low or too high, and increasingly so for larger θ values (see Supplementary Note 3). Since these biases decrease with larger read sample size, the number of sequence reads should be as large as possible as long as it does not preclude using the sequence reads for parameter inference. Moreover, in order to avoid bias in the case of weak dispersal limitation, the Ewens estimator should be favoured whenever it yields a higher likelihood value than the dispersal-limited estimator.

In practice, environmental DNA studies often sample the same regional species pool in different locations, which allows for more robust multi-sample maximum-likelihood inference (Etienne, 2007, 2009). It should be noted however that exact maximum-likelihood inference can be computationally prohibitive in the dispersal-limited case for larger numbers of reads than we used in this study or in the case of a multi-sample approach with large read samples (Lee *et al.*, 2013). Continuous approximations drawing on the work of Sloan *et al.* (2006, 2007) and Woodcock *et al.* (2007) might then be preferred, such as the Bayesian formulation of Harris *et al.* (2015).

Our analysis reveals that the presence of artifactual MOTUs is the most detrimental to neutral parameter inference. Bioinformatics methods aiming at limiting the number of artifactual MOTUs should be carefully applied to the sequencing data

before any attempt at estimating biodiversity indices (Sipos *et al.*, 2010; Coissac *et al.*, 2012; Mahe *et al.*, 2014). However, these methods do not guarantee a complete filtering of artifactual MOTUs from empirical datasets. In particular, chimeric sequences formed at the PCR stage may be misconstrued as MOTUs. Because these sequences are generated by rare error-generating PCR events, they should be predominantly represented by few reads. Thus one strategy for removing artifactual MOTUs consists in ignoring all MOTUs below an empirically set abundance threshold. However, in doing so, we lose the information on the relationship between the number of reads and the number of MOTUs. Hence we suggest that a more satisfactory method to mitigate this problem is to take a sufficiently small subsample of the sequence reads so as to trim out the artifactual MOTUs.

The presence of artifactual MOTUs in our simulated taxa assemblages manifests itself by a break in the slope of the ranked log-abundance curve (Fig. 1a, see also Fig. S3 in Supplementary Methods). Thus, the adequate subsample size for an empirical dataset may be chosen so as to trim out the MOTUs with abundances below an observed break in the ranked log-abundance curve. Another finding of our study is that for the same proportion of artifactual MOTUs, the θ estimate has a similar relative bias across θ values and the I estimate a similar relative bias across I values. Therefore, if artifactual MOTUs cannot be entirely excluded in an environmental DNA dataset, conclusions should be based on ratios of neutral parameter estimates among samples rather than on absolute values.

We modelled PCR noise using a lognormally distributed multiplicative noise term. We found a threshold noise value beyond which the inference of the neutral parameters becomes biased. This threshold was found to be lower for larger θ values. For instance, the empirical noise intensity $\sigma_{log} = 1.2$ measured on our benchmark dataset was near or below the threshold $\sigma_{log,th}$ for θ values up to ca. $\theta = 20$, while for larger θ values, it was responsible for a moderate underestimation of θ (20% for $\theta = 500$) and for a serious overestimation of I. Nevertheless, our benchmark dataset was here used for illustrative purposes, and noise intensity may differ in other datasets. In metabarcoding studies, noise intensity likely depends on the barcode, taxonomic group and wet laboratory protocol. Therefore we strongly advise to include

at least one benchmark dataset as part of any environmental DNA study to quantify noise intensity. Empirical noise assessments can then be compared to our simulation results.

We also simulated a Gaussian additive noise on abundance data and found that it had a disproportionate effect on the least abundant MOTUs, thus distorting the taxa-abundance distribution: parameter inference was biased if the standard deviation of the noise was larger than the abundance of the least abundant MOTUs. Here again, it is possible to correct for this type of noise in empirical datasets by subsampling the sequence reads. Additive noise can be considered to model the abundance noise generated by the sequencing step or by a single PCR cycle, while the succession of several PCR cycles produces a multiplicative abundance noise.

Another potential bias is due to the indirect relationship between the number of DNA barcode sequences in the sample and the number of sampled individuals. In particular, in the case of multicellular individuals, some of them may contribute disproportionately more than others. Given the variability and complexity of the associated noise structure, we chose to follow a modelling approach retaining as much generality as possible. We size-biased our samples by assuming that DNA availability in the environment is proportional to body mass, or to the turnover of body mass (i.e. the metabolic rate). We found that neutral parameter estimates are not modified by size structure in the community, irrespective of how strongly structured the community is, which is an interesting and general result.

Our approach to accounting for body size is directly inspired from the size-structured neutral model of O'Dwyer *et al.* (2009). This model integrates the growth of individuals into a neutral population dynamics without dispersal limitation, and may offer analytical predictions for the neutral "Species Biomass Distribution" (SBD) while accounting for the dependence of birth, death and growth rates on the size of individuals. When individuals grow in body size at a constant rate and neither birth nor death rates depend on size, this model predicts the same SBD as obtained analytically under our assumption of independent exponentially distributed sizes (see Supplementary Note 2). Our choice of a rate of environmental DNA release scaling with the $3/4^{th}$ power of body mass is motivated by a prediction of the metabolic theory of

ecology, which relates the metabolic rate to the body mass in one of the few general laws of ecology (West *et al.*, 1997).

Even though our modelling approach derives from theoretical considerations, it is also supported by some empirical evidence: it has been shown that the rate of DNA detection in the environment is biased by the size of organisms (Andersen *et al.*, 2012; Maruyama *et al.*, 2014; Klymus *et al.*, 2015), and the fact that DNA abundance should scale non-linearly with body mass has been experimentally verified in fishes (Maruyama *et al.*, 2014). Nevertheless, the noise introduced by size structure, fragments of organisms and extracellular DNA certainly has a far more complicated structure than we simulated. For instance, rates of DNA release into the environment and of DNA degradation both depend on taxa and on local conditions, and fluctuate temporally (Levy-Booth *et al.*, 2007; Barnes *et al.*, 2014; Strickler *et al.*, 2015). Moreover, the uneven spatial distribution of environmental DNA may prevent properly sampling the taxa-abundance distribution in the community, especially if whole pieces of living or decaying multicellular organisms are contained in the environmental sample. Pooling multiple spatially distributed samples should help average out local heterogeneity.

In this study, we considered that departure of the number of DNA barcode reads from the real taxon abundance is a source of bias. However, this source of bias may be generally seen as the accumulation of mutations during replication. In ecology, the only type of replication taken into consideration is demography, but DNA metabarcoding data are also the result of cellular and PCR replication processes. Since the assumptions of the neutral theory are generic and apply to any collection of replicating, mutating, and potentially dispersing entities, we could replace individual organisms by DNA barcodes as our basic replicating entities, and reinterpret the neutral parameters accordingly. As a consequence, we expect the taxa-abundance structure predicted by the neutral theory to be robust as long as the DNA barcodes do not differ too much in their replicating, mutating and dispersing abilities.

This study demonstrates that inferring the parameters of Hubbell's neutral model from the taxa-abundance distribution is possible even in noised biodiversity

datasets. We tested this hypothesis for a range of biologically plausible noise terms on simulated metabarcoding data, and we provide guidance for neutral parameter inference from such data. Our results indicate that whether an environmental DNA dataset really reflects the sampled community depends on noise intensity. They also suggest that this question can be answered by computing simple metrics on a benchmark dataset and comparing them to our simulations. The only way to quantify the noise level is to conduct careful benchmarking experiments, which will depend on the exact sampling and analysis protocol.

Acknowledgements

We thank Ryan Chisholm, Fabien Laroche and James O'Dwyer for fruitful discussion. This work has benefited from "*Investissement d'Avenir*" grants managed by the French *Agence Nationale de la Recherche* (CEBA, ref. ANR-10-LABX-25-01 and TULIP, ref. ANR-10-LABX-0041) and from an additional ANR grant (METABAR project; PI P. Taberlet).

References

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C. & Gnirke, A. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, **12**.

Amend, A.S., Seifert, K.A. & Bruns, T.D. (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K.H., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Ayarza, J.M. & Erijman, L. (2011) Balance of Neutral and Deterministic Components in the Dynamics of Activated Sludge Floc Assembly. *Microbial Ecology*, **61**, 486–495.

Baas Becking, L.G.M. (1934) *Geobiologie of inleiding tot de milieukunde.*, W.P. Van Stockum & Zoon, The Hague, the Netherlands.

Barnes, M.A., Turner, C.R., Jerde, C.L., Renshaw, M.A., Chadderton, W.L. & Lodge, D.M. (2014) Environmental conditions influence eDNA persistence in aquatic systems. *Environmental Science & Technology*, **48**, 1819–1827.

Bienert, F., De Danieli, S., Miquel, C., Coissac, E., Poillot, C., Brun, J. & Taberlet, P. (2012) Tracking earthworm communities from soil DNA. *Molecular Ecology*, **21**, 2017–2030.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233–243.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

Brown, J.H. (1995) *Macroecology*, University of Chicago Press.

Coissac, E., Riaz, T. & Puillandre, N. (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. *Molecular Ecology*, **21**, 1834–1847.

Degnan, P.H. & Ochman, H. (2012) Illumina-based analysis of microbial community diversity. *ISME Journal*, **6**, 183–194.

Drakare, S. & Liess, A. (2010) Local factors control the community composition of cyanobacteria in lakes while heterotrophic bacteria follow a neutral model. *Freshwater Biology*, **55**, 2447–2457.

Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C. & Fitter, A.H. (2010) Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME Journal*, **4**, 337–345.

Etienne, R.S. (2007) A neutral sampling formula for multiple samples and an "exact" test of neutrality. *Ecology Letters*, **10**, 608–618.

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Etienne, R.S. (2009) Maximum likelihood estimation of neutral model parameters for multiple samples with different degrees of dispersal limitation. *Journal of Theoretical Biology*, **257**, 510–514.

Etienne, R.S. & Alonso, D. (2005) A dispersal-limited sampling theory for species and alleles. *Ecology Letters*, **8**, 1147–1156.

Etienne, R.S. & Alonso, D. (2007) Neutral community theory: How stochasticity and dispersal-limitation can explain species coexistence. *Journal of Statistical Physics*, **128**, 485–510.

Etienne, R.S., Alonso, D. & McKane, A.J. (2007) The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, **248**, 522–536.

Etienne, R.S., Latimer, A.M., Silander, J.A. & Cowling, R.M. (2006) Comment on "Neutral ecological theory reveals isolation and rapid speciation in a biodiversity hot spot." *Science*, **311**, 610B–+.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical population biology*, **3**, 87–112.

Ewens, W.J. & Tavaré, S. (1997) *Multivariate Ewens Distribution*. *Discrete Multivariate Distributions* (ed. by Johnson,.

Fenchel, T. & Finlay, B.J. (2004) The ubiquity of small species: Patterns of local and global diversity. *Bioscience*, **54**, 777–784.

von Foerster, H. (1959) *Some remarks on changing populations*. *Kinetics of Cellular Proliferation*, pp. 382–399. Stohlman, F.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, **345**, 60–63.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2015) Linking statistical and ecological theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE*, **PP**, 1–14.

Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B-Biological Sciences*, **270**, 313–321.

Hilborn, R. & Mangel, M. (1997) *The ecological detective: confronting models with data*, Princeton University Press.

Hoppe, F.M. (1984) Polya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*, **20**, 91–94.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Huber, J.A., Mark Welch, D., Morrison, H.G., Huse, S.M., Neal, P.R., Butterfield, D.A. & Sogin, M.L. (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97–100.

Jabot, F. & Chave, J. (2009) Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecology Letters*, **12**, 239–248.

Jabot, F., Etienne, R.S. & Chave, J. (2008) Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos*, **117**, 1308–1320.

Kembel, S.W., Wu, M., Eisen, J.A. & Green, J.L. (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *Plos Computational Biology*, **8**, 11.

Klymus, K.E., Richter, C.A., Chapman, D.C. & Paukert, C. (2015) Quantification of eDNA shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp Hypophthalmichthys molitrix. *Biological Conservation*, **183**, 77–84.

Lee, J.E., Buckley, H.L., Etienne, R.S. & Lear, G. (2013) Both species sorting and neutral processes drive assembly of bacterial communities in aquatic microcosms. *Fems Microbiology Ecology*, **86**, 288–302.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, Elsevier.

Levy-Booth, D.J., Campbell, R.G., Gulden, R.H., Hart, M.M., Powell, J.R., Klironomos, J.N., Pauls, K.P., Swanton, C.J., Trevors, J.T. & Dunfield, K.E. (2007) Cycling of extracellular DNA in the soil environment. *Soil Biology & Biochemistry*, **39**, 2977–2991.

Mahe, F., Rognes, T., Quince, C., de Vargas, C. & Dunthorn, M. (2014) Swarm: robust and fast clustering method for amplicon-based studies. *Peerj*, **2**.

Maruyama, A., Nakamura, K., Yamanaka, H., Kondoh, M. & Minamoto, T. (2014) The release rate of environmental DNA from juvenile and adult fish. *Plos One*, **9**, 13.

Nguyen, N.H., Smith, D., Peay, K. & Kennedy, P. (2015) Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*, **205**, 1389–1393.

O'Dwyer, J.P., Lake, J.K., Ostling, A., Savage, V.M. & Green, J.L. (2009) An integrative framework for stochastic, size-structured community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6170–6175.

Ofiteru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C.A. & Sloan, W.T. (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15345–15350.

Ostman, O., Drakare, S., Kritzberg, E.S., Langenheder, S., Logue, J.B. & Lindstrom, E.S. (2010) Regional invariance among microbial communities. *Ecology Letters*, **13**, 118–127.

Pienaar, E., Theron, A., Nelson, A. & Viljoen, H.J. (2006) A quantitative model of error accumulation during PCR amplification. *Computational Biology and Chemistry*, **30**, 102–111.

Quince, C., Lanzen, A., Davenport, R.J. & Turnbaugh, P.J. (2011) Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, **12**, 18.

Ramirez, K.S., Leff, J.W., Barberan, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H. & Fierer, N. (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 9.

Ricklefs, R.E. (2004) A comprehensive framework for global patterns in biodiversity. *Ecology Letters*, **7**, 1–15.

Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K.M., Kent, A.D., Daroub, S.H., Camargo, F.A.O., Farmerie, W.G. & Triplett, E.W. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal*, **1**, 283–290.

Roguet, A., Laigle, G.S., Therial, C., Bressy, A., Soulignac, F., Catherine, A., Lacroix, G., Jardillier, L., Bonhomme, C., Lerch, T.Z. & Lucas, F.S. (2015) Neutral community model explains the bacterial community assembly in freshwater lakes. *Fems Microbiology Ecology*, **91**, 11.

Rosenzweig, M.L. (1995) *Species diversity in space and time*, Cambridge University Press.

Rosindell, J., Hubbell, S.P., He, F., Harmon, L.J. & Etienne, R.S. (2012) The case for ecological neutral theory. *Trends in Ecology & Evolution*, **27**, 203–208.

Ross, M.G., Russ, C., Costello, M., Hollinger, A., Lennon, N.J., Hegarty, R., Nusbaum, C. & Jaffe, D.B. (2013) Characterizing and measuring bias in sequence data. *Genome Biology*, **14**.

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2009) The map equation. *The European Physical Journal Special Topics*, **178**, 13–23.

Schleper, C., Jurgens, G. & Jonuscheit, M. (2005) Genomic studies of uncultivated archaea. *Nature Reviews Microbiology*, **3**, 479–488.

Sipos, M., Jeraldo, P., Chia, N., Qu, A.I., Dhillon, A.S., Konkel, M.E., Nelson, K.E., White, B.A. &

Goldenfeld, N. (2010) Robust computational analysis of rRNA hypervariable tag datasets. *Plos One*, **5**, 8.

Sipos, R., Szekely, A.J., Palatinszky, M., Revesz, S., Marialigeti, K. & Nikolausz, M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiology Ecology*, **60**, 341–350.

Sloan, W.T., Lunn, M., Woodcock, S., Head, I.M., Nee, S. & Curtis, T.P. (2006) Quantifying the roles of immigration and chance in shaping prokaryote community structure. *Environmental Microbiology*, **8**, 732–740.

Sloan, W.T., Woodcock, S., Lunn, M., Head, I.M. & Curtis, T.P. (2007) Modeling taxa-abundance distributions in microbial communities using environmental sequence data. *Microbial Ecology*.

Strickler, K.M., Fremier, A.K. & Goldberg, C.S. (2015) Quantifying effects of UV-B, temperature, and pH on eDNA degradation in aquatic microcosms. *Biological Conservation*, **183**, 85–92.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. & Willerslev, E. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic acids research*, **35**, e14–e14.

Tedersoo, L., Bahram, M., Cajthaml, T., Põlme, S., Hiiesalu, I., Anslan, S., Harend, H., Buegger, F., Pritsch, K., Koricheva, J. & Abarenkov, K. (2015) Tree diversity and species identity effects on soil fungi, protists and animals are context dependent. *Isme Journal*.

Tedersoo, L., Bahram, M., Polme, S., Koljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., Smith, M.E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Poldmaa, K., Piepenbring, M., Phosri, C., Peterson, M., Parts, K., Partel, K., Otsing, E., Nouhra, E., Njouonkou, A.L., Nilsson, R.H., Morgado, L.N., Mayor, J., May, T.W., Majuakim, L., Lodge, D.J., Lee, S.S., Larsson, K.H., Kohout, P., Hosaka, K., Hiiesalu, I., Henkel, T.W., Harend, H., Guo, L.D., Greslebin, A., Grelet, G., Geml, J., Gates, G., Dunstan, W., Dunk, C., Drenkhan, R., Dearnaley, J., De Kesel, A., Dang, T., Chen, X., Buegger, F., Brearley, F.Q., Bonito, G., Anslan, S., Abell, S. & Abarenkov, K. (2014) Global diversity and geography of soil fungi. *Science*, **346**, 1078–+.

Vellend, M. (2010) Conceptual synthesis in community ecology. *Quarterly Review of Biology*, **85**, 183–206.

Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, **424**, 1035–1037.

Weber, A.A.T. & Pawlowski, J. (2013) Can abundance of protists be inferred from sequence data: A case study of Foraminifera. *Plos One*, **8**, 8.

West, G.B., Brown, J.H. & Enquist, B.J. (1997) A general model for the origin of allometric scaling laws in biology. *Science*, **276**, 122–126.

Whitman, W.B., Coleman, D.C. & Wiebe, W.J. (1998) Prokaryotes: The unseen majority. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 6578–6583.

Woodcock, S., van der Gast, C.J., Bell, T., Lunn, M., Curtis, T.P., Head, I.M. & Sloan, W.T. (2007) Neutral assembly of bacterial communities. *Fems Microbiology Ecology*, **62**, 171–180.

Yoccoz, N.G., Brathen, K.A., Gielly, L., Haile, J., Edwards, M.E., Goslar, T., von Stedingk, H., Brysting, A.K., Coissac, E., Pompanon, F., Sonstebo, J.H., Miquel, C., Valentini, A., de Bello, F., Chave,

J., Thuiller, W., Wincker, P., Cruaud, C., Gavory, F., Rasmussen, M., Gilbert, M.T.P., Orlando, L., Brochmann, C., Willerslev, E. & Taberlet, P. (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.

Yu, D.W., Ji, Y.Q., Emerson, B.C., Wang, X.Y., Ye, C.X., Yang, C.Y. & Ding, Z.L. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zinger, L., Gobet, A. & Pommier, T. (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, **21**, 1878–1896.

Supplementary Information

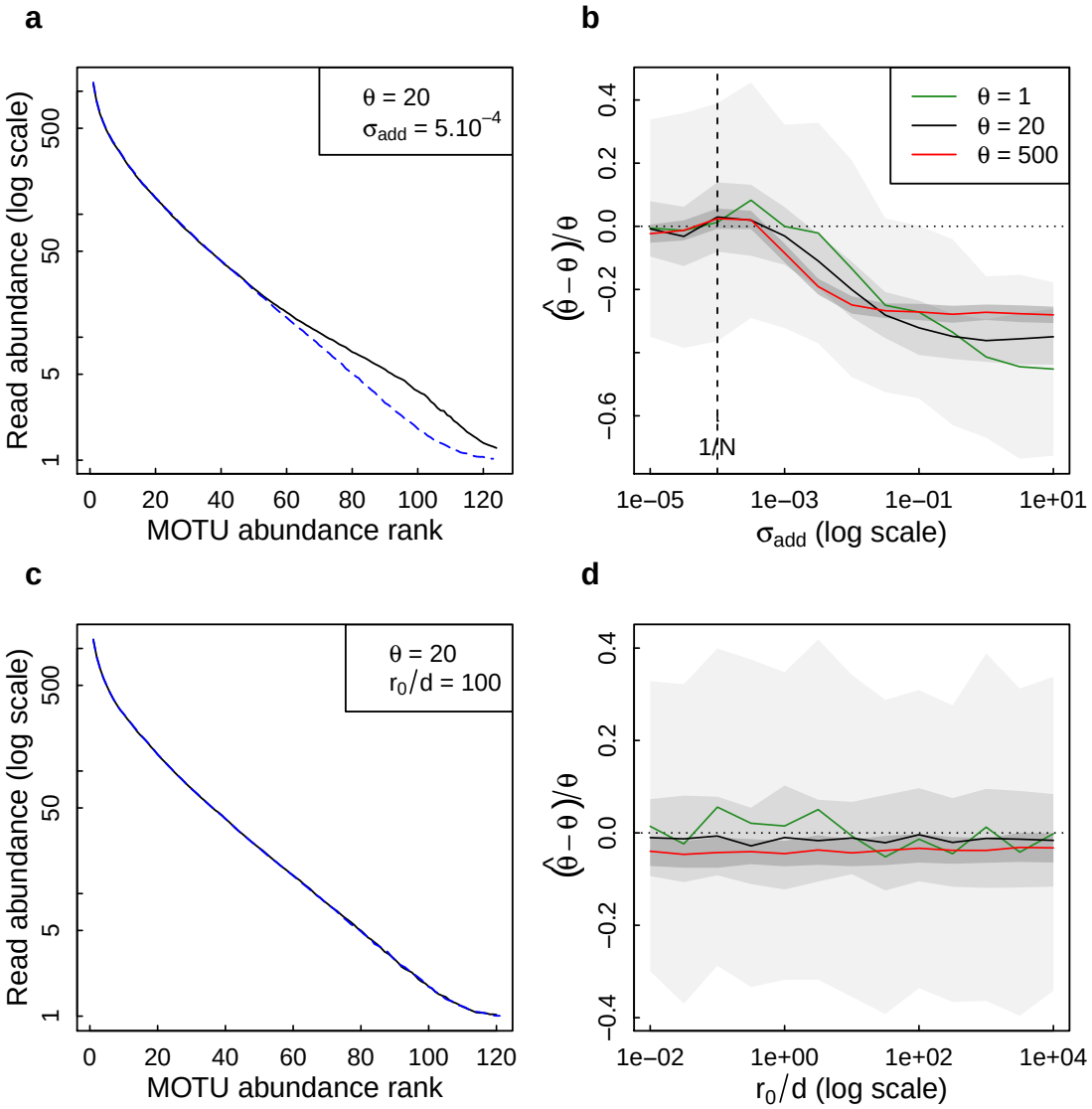

Figure S1: Effect of additive noise and metabolic rate on neutral parameter inference. **Left panels**: mean MOTU rank abundance distributions over 100 realizations for $\theta = 20$ in a 10^4-read sample, without (dashed blue line) and with (black line) simulated noise: (**a**) additive Gaussian noise of standard deviation $\sigma_{add} = 5.10^{-4}$ (5 times the relative abundance $1/N = 10^{-4}$ of the least abundant MOTUs), and (**c**) size structure among individuals and non-linear scaling of DNA release with body mass, for a body size ratio $\frac{g}{dn_0} = 1{,}000$ and a ratio $\frac{r_0}{d} = 100$ between metabolic rate and death rate. **Right panels**: mean and standard deviation over 100 realizations of the relative bias on the θ estimate in a 10^4-read sample, for $\theta = 1$ (green), $\theta = 20$ (black) and $\theta = 500$ (red), as a function of (**b**) the additive noise intensity σ_{add}, and (**d**) the ratio $\frac{r_0}{d}$.

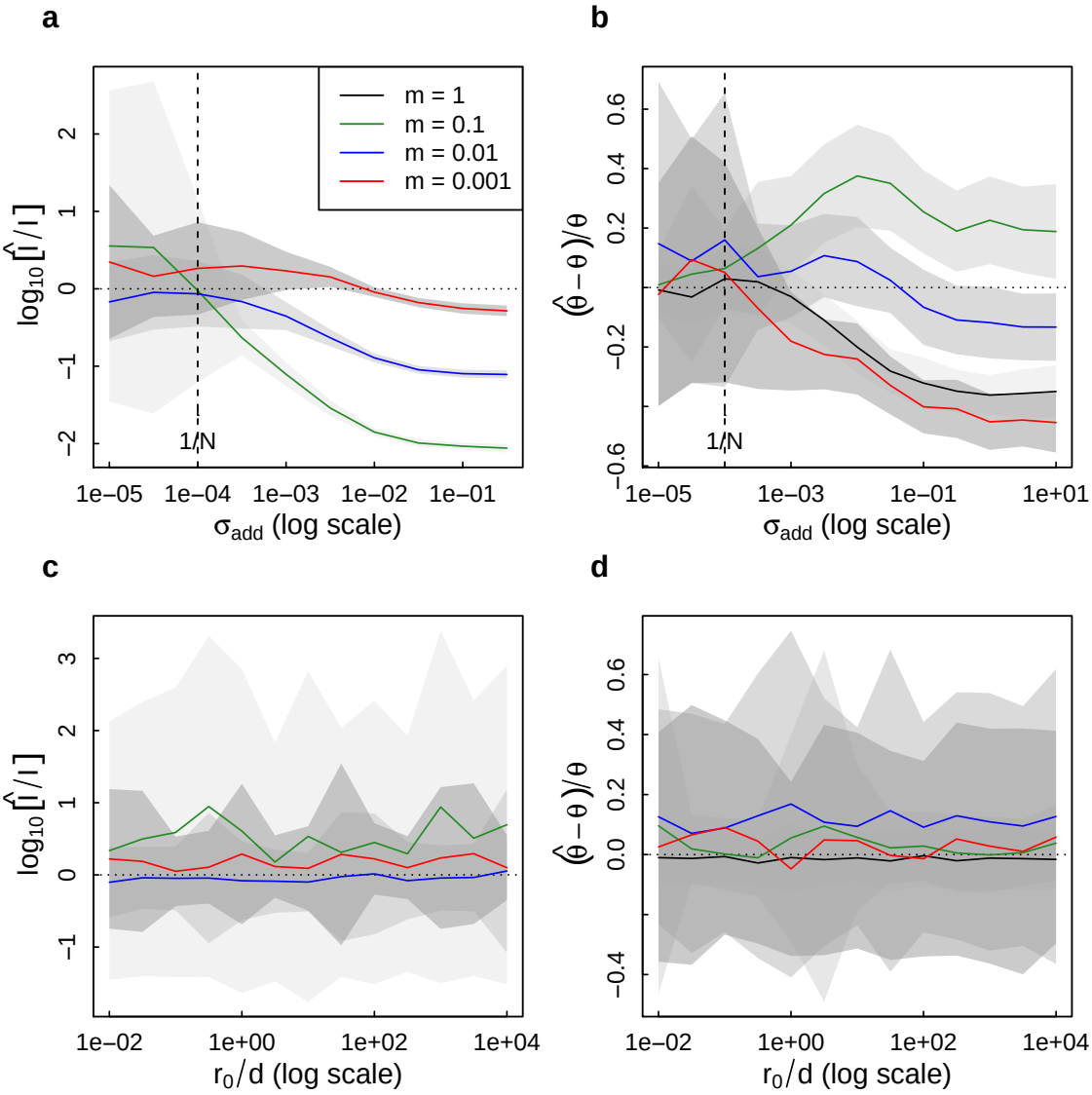

Figure S2: Effect of additive noise and metabolic rate on neutral parameter inference in the presence of dispersal limitation. We simulated a 10^4-read sample and computed the mean and standard deviation over 100 realizations of $(\hat{\theta} - \theta)/\theta$ and $\log_{10}(\hat{I}/I)$. Results are plotted for $\theta = 20$ and for $m = 1$ (black), $m = 0.1$ (green), $m = 0.01$ (blue) and $m = 0.001$ (red). **Panels a-b:** variation with the noise intensity σ_{add} of an additive Gaussian noise on relative abundances ($1/N = 10^{-4}$ is the relative abundance of the least abundant MOTUs). **Panels c-d:** variation with the ratio $\frac{r_0}{d}$ between metabolic rate and death rate.

Supplementary Methods: Quantifying noise using a benchmark dataset

To build our benchmark dataset, we mixed the genomic DNA extracted from 16 Alpine plant species in known quantities (Table S1), and we amplified and sequenced the chloroplast *trn*L P6-loop barcode (primer g-h; Taberlet *et al.*, 2007). Amplification and sequencing were replicated eight times. The DNA concentrations of the different species in the mixture scaled logarithmically, with a doubling in genomic DNA concentration from one species to the next more abundant. The 16 species thus spanned a large range of DNA concentration (1.10^{-5} ng/µL to 1 ng/µL), representative of the DNA abundances found in environmental samples.

The PCR mixtures comprised 2 ng DNA template, 10 µl of AmpliTaq Gold® Master Mix (Life Technologies, Carlsbad, CA, USA), 0.25 µM of each primer, 3.2 µg of BSA (Roche Diagnostic, Basel, Switzerland) for a final reaction volume of 20 µl. Thermocycling conditions consisted of an initial denaturation step (95°C, 10 min) followed by 35 cycles of denaturation at 95°C (30 s), primer annealing at 50°C (30 s) and elongation at 72°C (1 min), and by a final extension step (72°C, 7 min). Amplicons were then purified (MinElute™ PCR purification kit, Qiagen), pooled, loaded on a HiSeq Illumina lane and sequenced using the paired-end technology. The read coverage was about 10^5 Illumina sequence reads for each of the eight replicates.

The sequencing data were first curated following classical procedures using the OBITools package (Boyer *et al.*, 2016), consisting in paired-end read assembly, read assignation to their respective samples and dereplication. Sequences of length shorter than 10 nucleotides or containing ambiguous nucleotides were excluded. The sequences were then processed using the Infomap clustering algorithm (Rosvall *et al.*, 2009), to minimize the number of artifactual MOTUs by clustering sequences together based on their similarity. The dataset is considered as a network of sequences connected by links weighted according to sequence similarity. We used weights decreasing exponentially with the number of nucleotide differences between sequences and we discarded the links for more than 5 nucleotide differences. All replicates were lumped for this clustering analysis. In parallel, all sequences were

assigned to a taxon using the barcodes of the 16 species as a reference database (Table S1).

The clustering algorithm yielded 48 clusters (i.e. MOTUs), 24 of which were found only in some of the replicates (Fig. S3a). Each input species was represented as the most abundant sequence of a MOTU found in all 8 realizations. Taking only into account the MOTUs shared across replicates, the proportion of artifactual MOTUs in the curated dataset is 33% (Fig. S3b). Using the taxonomic assignation of all sequences to the most similar of the 16 species, we found that each artifactual MOTU originates from a single species and is at least 50 times less abundant than the species that generated it (Fig. S3a). Therefore, artifactual MOTUs have little impact on the abundance of the true MOTUs in the dataset. Moreover, the number of artifactual MOTUs generated by a species is proportional to the latter's read abundance r (Fig. S3c), and the log-abundance of these artifactual MOTUs is uniformly distributed between 0 and $\log(r/50)$. Our modeling choice for simulating artifactual MOTUs with realistic abundances built on these empirical observations.

The amplification factor, i.e. the ratio between the read abundance and the initial DNA concentration, was found to be approximately constant over the range of DNA concentrations spanned in the dataset (Fig. S3d). However, it varied across species and replicates. This results in a multiplicative noise on relative abundances that is approximately lognormally distributed, with logarithm standard deviation $\sigma_{log} = 1.2$ (Fig. S3e). Seventy-three percent of the variance of the logarithm is explained by differences among species (likely related to the variability in barcode copy number and in efficiency of PCR amplification) while the remaining variance corresponds to the variability among realizations (Fig. S3d).

References:

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2009) The map equation. *The European Physical Journal Special Topics*, **178**, 13–23.

Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G.,

Brochmann, C. & Willerslev, E. (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic acids research*, **35**, e14–e14.

Species	Dilution factor	Sequence	Sequence length (nt)	Sequence GC content (%)
Taxus baccata	1.000000	atccgtattataggaacaataattttattttctagaaaagg	41	24.39
Salvia pratensis	0.500000	atcctgttttctcaaaacaaaggttcaaaaaacgaaaaaaaaag	45	26.67
Populus tremula	0.250000	atcctatttttcgaaaacaaacaaaaaaacaaacaaaggttcataaagacagaataagaatacaaaag	68	25.00
Rumex acetosa	0.125000	ctcctcctttccaaaaggaagaataaaaaag	31	35.48
Carpinus betulus	0.062500	atcctgttttcccaaaacaaataaaacaaatttaagggttcataaagcgagaataaaaaag	61	27.87
Fraxinus excelsior	0.031250	atcctgttttcccaaaacaaaggttcagaaagaaaaaag	39	33.33
Picea abies	0.015625	atccggttcatggagacaatagtttcttcttttattctcctaagataggaaggg	54	38.89
Lonicera xylosteum	0.007813	atccagttttccgaaaacaagggtttagaaagcaaaaatcaaaaag	46	32.61
Abies alba	0.003906	atccggttcatagagaaaagggtttctctccttctcctaaggaaagg	47	44.68
Acer campestre	0.001953	atcctgttttacgagaataaaacaaagcaaacaagggttcagaaagcgagaaaggg	56	39.29
Briza media	0.000977	atccgtgttttgagaaaacaaggggttctcgaactagaatacaaaggaaaag	53	39.62
Rosa canina	0.000488	atcccgttttatgaaaacaaacaaggtttcagaaagcgagaataaataaag	51	31.37
Capsella bursa-pastoris	0.000244	atcctggtttacgcgaacacaccggagtttacaaagcgagaaaaaagg	48	45.83
Geranium robertianum	0.000122	atccttttttacgaaaataaagaggggctcacaaagcgagaatagaaaaaaag	53	33.96
Rhododendron ferrugineum	0.000061	atccttttttcgcaaacaaacaaagattccgaaagctaaaaaaaag	46	30.43
Lotus corniculatus	0.000031	atcctgctttacgaaaacaagggaaagttcagttaagaaagcgacgagaaaaatg	55	38.18

Table S1: List and characteristics of the 16 plant species included in the benchmark dataset.

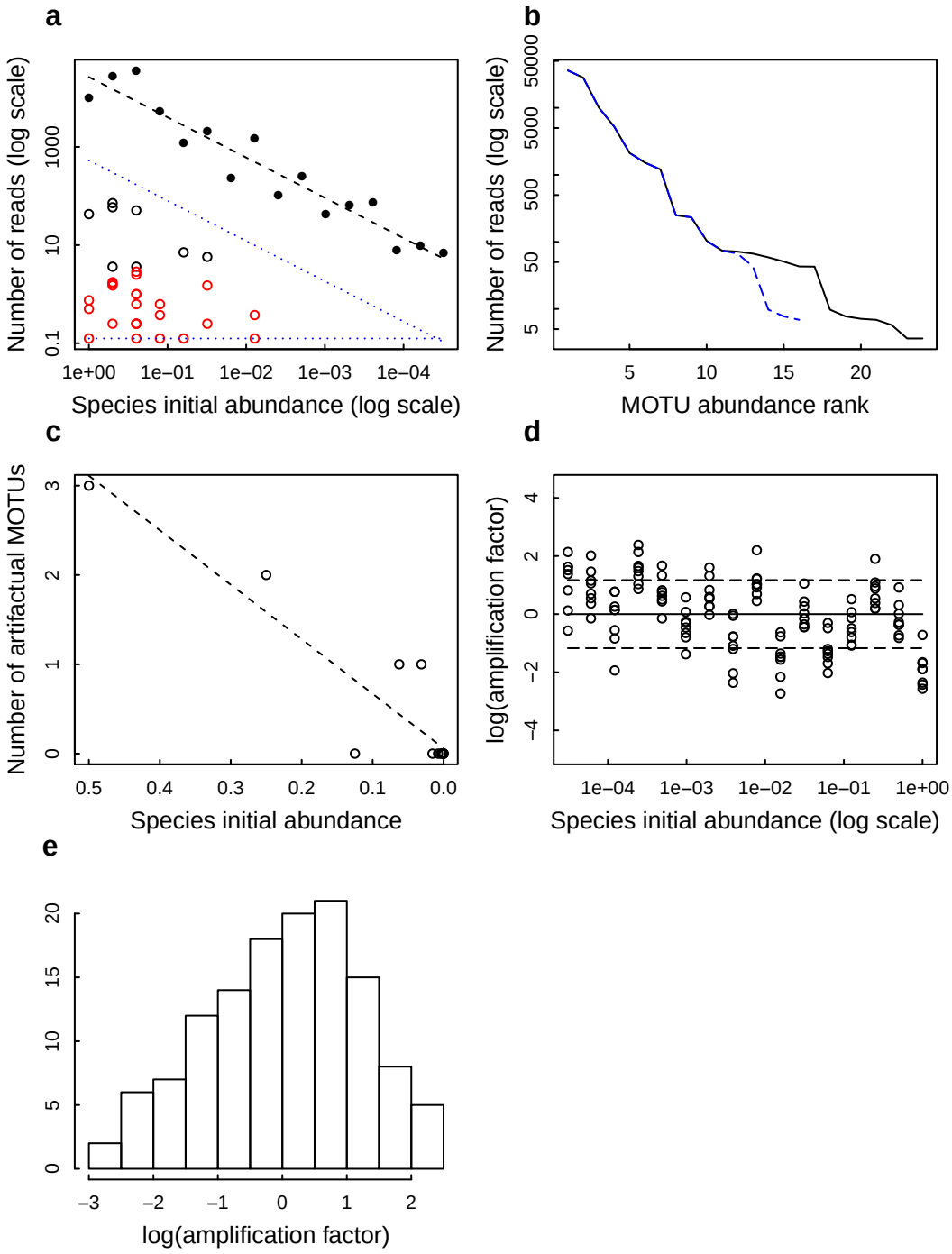

Figure S3: Empirical results for the benchmark dataset obtained by mixing the DNA of 16 plant species, then amplifying by PCR and sequencing on an Illumina platform the chloroplast trnL P6-loop barcode, with eight replicates. **Panel a**: Read abundance of the 16 species (•) and of the artifactual MOTUs (∘, ∘), averaged over the replicates, as a function of the species initial abundance. Some artifactual MOTUs were found in every realization (∘), but others were not (∘). The blue dotted lines delineate the abundance domain chosen to model the abundances of artifactual MOTUs. **Panel b**: Number of reads per MOTU as a function of the MOTU's abundance rank, including and excluding artifactual MOTUs (black and dashed blue, respectively). **Panel c**: Linear relationship between the number of artifactual MOTUs and the

relative abundance of the species that generated them (the most abundant species is excluded, as well as the MOTUs found in only some of the replicates). **Panel d**: Logarithm of the amplification factor, i.e. the ratio between the read abundance and the initial DNA concentration, as a function of the initial DNA concentration of the species (dotted lines: standard deviation $\sigma_{log} = 1.2$ over all species and all replicates). **Panel e**: Probability density of the logarithm of the amplification factor over the 16 species and the 8 realizations, approximately normally distributed.

Supplementary Note 1: Hubbell's neutral model

Hubbell's neutral model of biodiversity describes a large pool of J_M individuals undergoing random death, birth and speciation events in the following way: at each time step, one individual at random dies, and is replaced by a new individual. This new individual belongs to a taxon not previously found in the community with probability ν, or to one of the already existing taxa with probability $1-\nu$. In the latter case, each taxon has a probability to be picked proportionally to its abundance in the community[1]. In the absence of dispersal limitation, the multivariate steady-state distribution of taxa abundances is called the Ewens distribution and is characterized by the single parameter $\theta = \frac{\nu}{1-\nu}(J_M - 1)$ (Ewens, 1972; Etienne & Alonso, 2005). Any sample consisting of $J < J_M$ individuals drawn at random from the community follows also the Ewens distribution of parameter θ.

A dispersal-limited version of this model is defined as follows (Hubbell, 2001; Etienne & Alonso, 2005). New taxa disperse into a single local community by immigration from a regional pool, which follows the model without dispersal limitation described above. When an individual dies, it is replaced by an immigrating individual with probability m, and by the offspring of a local individual with probability $1-m$. Two immigrants may belong to the same taxon. The multivariate steady-state distribution of taxa abundances in the dispersal-limited local community depends on two parameters: the dispersal parameter $I = \frac{m}{1-m}(J-1)$, where J is the number of individuals in the local community, and the diversity parameter θ of the regional pool (Etienne, 2005). Any sample drawn at random from the local community also follows the Etienne distribution of parameters θ and I (Etienne & Alonso, 2005).

References:

Etienne, R.S. & Alonso, D. (2005) A dispersal-limited sampling theory for species and alleles. *Ecology Letters*, **8**, 1147–1156.

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical population*

biology, **3**, 87–112.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Supplementary Note 2: Modeling size differences

1. Modeling size differences using the von Foerster equation

The von Foerster equation (von Foerster, 1959; O'Dwyer *et al.*, 2009) describes a population where individuals grow in number of cells (or mass) n with a growth rate $g(n)$, and where they die with a death rate $d(n)$. The evolution of the number $j(n,t)\mathrm{d}n$ of individuals with a number of cells between n and $n+\mathrm{d}n$ at time t is given by:

$$\frac{\partial j(n,t)}{\partial t} = -\frac{\partial\big(g(n)j(n,t)\big)}{\partial n} - d(n)j(n,t)$$

When $g(n)$ and $d(n)$ are independent of n, the stationary (i.e., time-independent) solution of the von Forster equation is:

$$j(n) = J\frac{d}{g}e^{-\frac{d}{g}(n-n_0)}$$

where J is the constant population size, given by:

$$J = \int_{n_0}^{\infty} j(n)\,dn$$

Therefore, a randomly chosen individual in the population has a number n of cells with probability density:

$$p_{ind}(n) = \frac{j(n)}{J} = \frac{d}{g}e^{-\frac{d}{g}(n-n_0)}$$

We used this probability density to draw a number of cells between n_0 and infinity for each individual of the neutral sample. The mean $< n >$ and the coefficient of variation $\sigma_n/< n >$ of the number of cells of a randomly chosen individual are given by:

$$< n >= \frac{g}{d} + n_0$$

$$\frac{\sigma_n}{< n >} = \frac{1}{\frac{d}{g}n_0 + 1}$$

2. Comparison with O'Dwyer et al. (2009)'s size-structured neutral model

O'Dwyer *et al*. (2009) transformed the deterministic von Foerster equation into a probabilistic equation, and integrated it into the master equation of Volkov *et al.* (2003), which describes a neutral dynamics without dispersal limitation in a probabilistic way. The resulting size-structured neutral model predicts that, in steady state, if the growth rate $g(n)$, the birth rate $b(n)$ and the death rate $d(n)$ are independent of n, and if $\frac{g}{d} \gg n_0$ (i.e., individuals grow much larger than their size at birth), a randomly chosen species will have a total number of cells (or a total biomass) n with probability density:

$$p_{sp}(n) = \frac{\nu}{bn}\left(e^{-\frac{d-b}{g}n} - e^{-\frac{d}{g}n}\right)$$

where ν is the speciation rate of the neutral model ($\nu/b \ll 1$). Adding size structure does not modify the probability for a randomly chosen species to have J individuals:

$$P_{sp}(J) = \frac{\nu}{bJ}\left(\frac{b}{d}\right)^J$$

While the model of O'Dwyer *et al.* (2009) explicitly accounts for the coupling between the demographic dynamics and the growth of individuals, we generated a neutral sample of individuals and then assigned an independent number of cells to each individual. Therefore, under our assumptions, the numbers of cells of the different individuals are described by independent and identically distributed exponential random variables N_i, and for $\frac{g}{d} \gg n_0$, the total number of cells of a species with J individuals follows an Erlang distribution:

$$\sum_{i=1}^{J} N_i \sim Erlang\left(\frac{g}{d}, J\right)$$

with probability density:

$$p_{sp}(n|J) = p_{Erlang\left(\frac{g}{d},J\right)}(n) = \frac{1}{(J-1)!}\left(\frac{d}{g}\right)^J n^{J-1} e^{-\frac{d}{g}n}$$

The probability density for a species of having a total number of cells n is then given by:

$$p_{sp}(n) = \sum_{J=1}^{\infty} P_{sp}(J) p_{sp}(n|J)$$

Combining the expressions of $P_{sp}(J)$ and $p_{sp}(n|J)$ above, we obtain the same expression for $p_{sp}(n)$ as predicted by the size-structured model of O'Dwyer *et al*. (2009). Therefore, in the simple case where g(n), $b(n)$ and $d(n)$ are independent of the number of cells n, explicitly accounting for the coupling between demographic dynamics and individual growth is equivalent to assuming as we did that all individuals have independent and identically distributed numbers of cells.

The modelling approach of Volkov *et al*. (2003) and O'Dwyer *et al*. (2009) differs from that of Ewens (1972) and Etienne (2005). The former consists in describing the population dynamics of a single species with a fluctuating number of individuals, independently of the remaining of the community, and then considering that the results hold for every species in the community ("mean-field" approach). In contrast, the Ewens and Etienne distributions are obtained by explicitly considering a community with a constant number of individuals and a fluctuating number of species through time. However, the two approaches yield identical stationary distributions provided that the number of species is large enough (Etienne *et al.*, 2007).

References:

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Etienne, R.S., Alonso, D. & McKane, A.J. (2007) The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology*, **248**, 522–536.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theoretical population biology*, **3**, 87–112.

von Foerster, H. (1959) *Some remarks on changing populations*. *Kinetics of Cellular Proliferation*, pp. 382–399. Stohlman, F.

O'Dwyer, J.P., Lake, J.K., Ostling, A., Savage, V.M. & Green, J.L. (2009) An integrative framework for stochastic, size-structured community assembly. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 6170–6175.

Volkov, I., Banavar, J.R., Hubbell, S.P. & Maritan, A. (2003) Neutral theory and relative species abundance in ecology. *Nature*, **424**, 1035–1037.

Supplementary Note 3: Estimator performance without simulated noise

We explored how the maximum-likelihood neutral estimators behave in the absence of simulated noise over the range of tested parameter values (θ in [1, 500] and m in [0.001, 1]). We found that while the Ewens estimator is very little biased (Fig. S4a-b), the dispersal-limited estimator can be strongly biased depending on parameter values and sample size (Fig. S4c-f). The dispersal-limited estimator underestimates θ and overestimates I when the immigration rate into the local community is too small, and overestimates $\theta$and underestimates I when the immigration rate is too large. In the case of our 10^4-read sample, values of I around $I = 10^3$ (i.e. $m = 0.01$ in the 10^5-individual sample) allow for the least biased estimation of (θ, I). Biases are strongest for $\theta > 100$.

For both estimators standard deviation and bias decrease with sample size, but a much larger sample size is required to obtain accurate estimates in the dispersal-limited case than in the absence of dispersal limitation. While sample sizes of ca. $N = 100$ are sufficient for the Ewens estimator, sample sizes of $N = 10^4$ are still not sufficient for some parameter values in the dispersal-limited case. Larger θ values and smaller I values require larger sample sizes. Estimating the neutral parameters simultaneously from several read samples reduces these biases (Etienne, 2007).

Reference:

Etienne, R.S. (2007) A neutral sampling formula for multiple samples and an “exact” test of neutrality. *Ecology Letters*, **10**, 608–618.

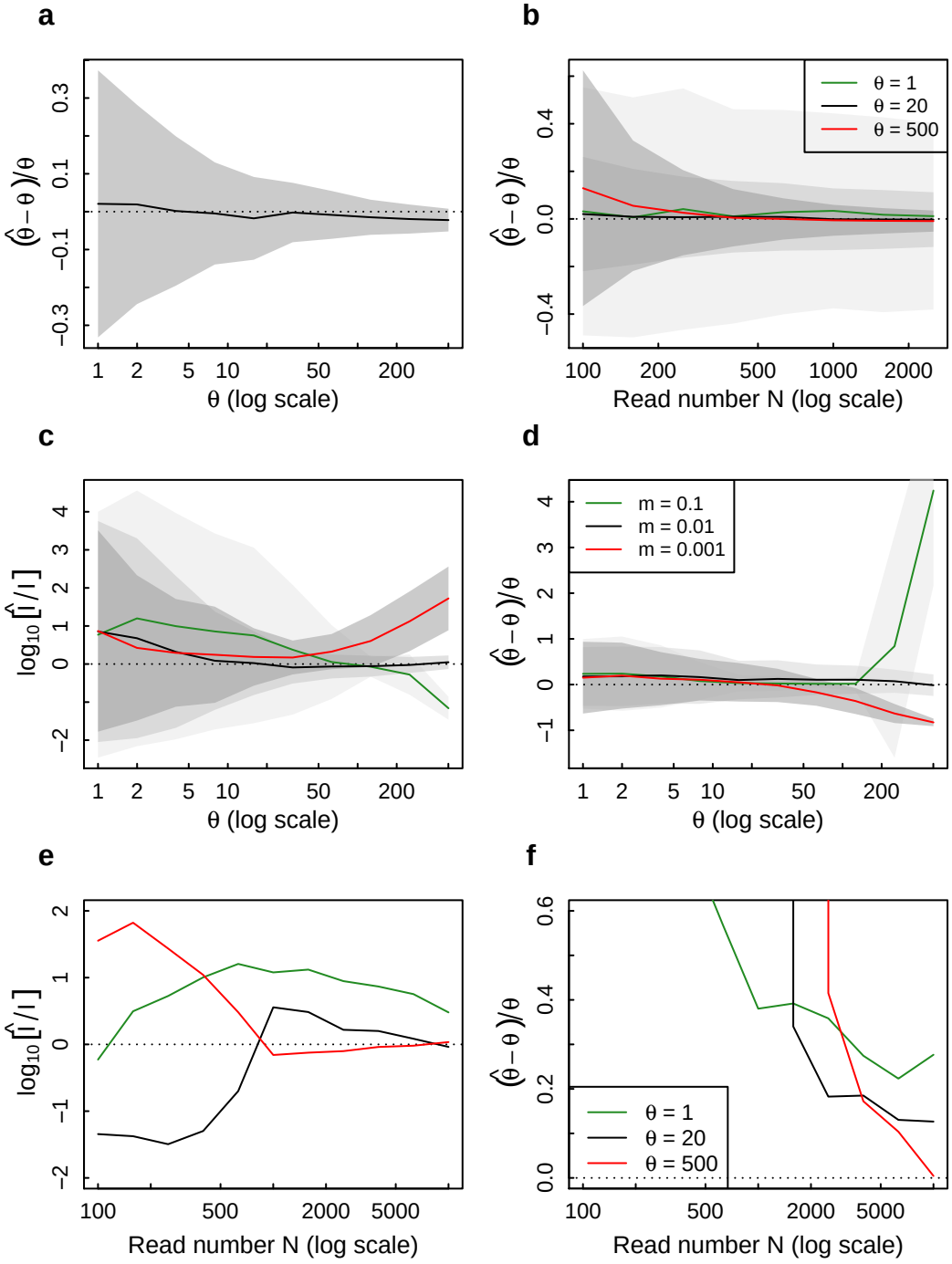

Figure S4: Neutral parameter inference without simulated noise, for different parameter values. The mean and standard deviation of the relative biases on parameter estimates are plotted over 500 realizations. **Panels a-b:** θ inference without dispersal limitation, as a function of (a) the input θ value and (b) the read number N, for θ equal to 1, 20, and 500. **Panels c-d:** θ and $\log_{10}(I)$ inference as a function of the input θ value, for m equal to 0.1, 0.01, and 0.001. **Panels e-f:** θ and $\log_{10}(I)$ inference as a function of the read number N, for $m = 0.01$ and for θ equal to 1, 20, and 500.

Supplementary Note 4: Neutral parameter inference with the number of individuals unknown

Because exact maximum-likelihood inference of the neutral parameters relies on sampling formulas that are invariant under subsampling, it is possible to use the sequence reads as effective individuals as long as we can consider the reads as a subsample from the initial individuals. Therefore, there should be less reads than individuals. A further complication is that the sequence reads are sampled with replacement from the initial individuals in our simulations (i.e. they are a multinomial sample from the relative abundances) instead of without replacement as required for the invariance property to hold. Hence there should be in fact several times less reads than individuals, because sampling with and without replacement are equivalent only in this case.

To illustrate this assumption, we explored how the Ewens maximum-likelihood estimator behaves in the absence of simulated noise depending on the initial number of individuals J, for $N = 10^2, N = 10^3, N = 10^4$ and $N = 10^5$, and for $\theta = 20$. As expected, the Ewens estimator yields an unbiased θ estimate as long as the initial number of individuals is ca. one order of magnitude larger than the number of reads (Fig. S5a-d). We then simulated a number of reads larger than the initial number of individuals ($N = 10^5$ reads for $J = 10^4$ individuals, and $N = 10^4$ reads for $J = 10^3$ individuals), and took smaller subsamples of reads from the original read sample until reaching a stable θ maximum-likelihood estimate. As expected, the θ estimate becomes stable under subsampling for samples at least one order of magnitude smaller than the initial number J of individuals. Using this method, we achieved an unbiased estimation of θ in spite of the small initial number of individuals (Fig. S5e-f). In the dispersal-limited case, we expect the maximum likelihood estimator based on the Etienne sampling formula to behave similarly.

We also compared estimating θ using the Ewens estimator and estimating θ by linear regression on the ranked log-abundance. We found that both methods perform similarly when the number of reads is one order of magnitude larger than the initial number of individuals, and that when this condition is not met, linear regression still

provides an unbiased θ estimate (Fig. S5a-d). However, unlike maximum likelihood inference, linear regression on the ranked log-abundance is not reliable when either the number of reads or the initial number of individuals is too low (lower than ca. 500 for $\theta = 20$; Fig. S5a-d), or when there is too little taxonomic diversity in the sample. Moreover, the θ estimate depends on the arbitrary delimitation of the linear domain of the curve.

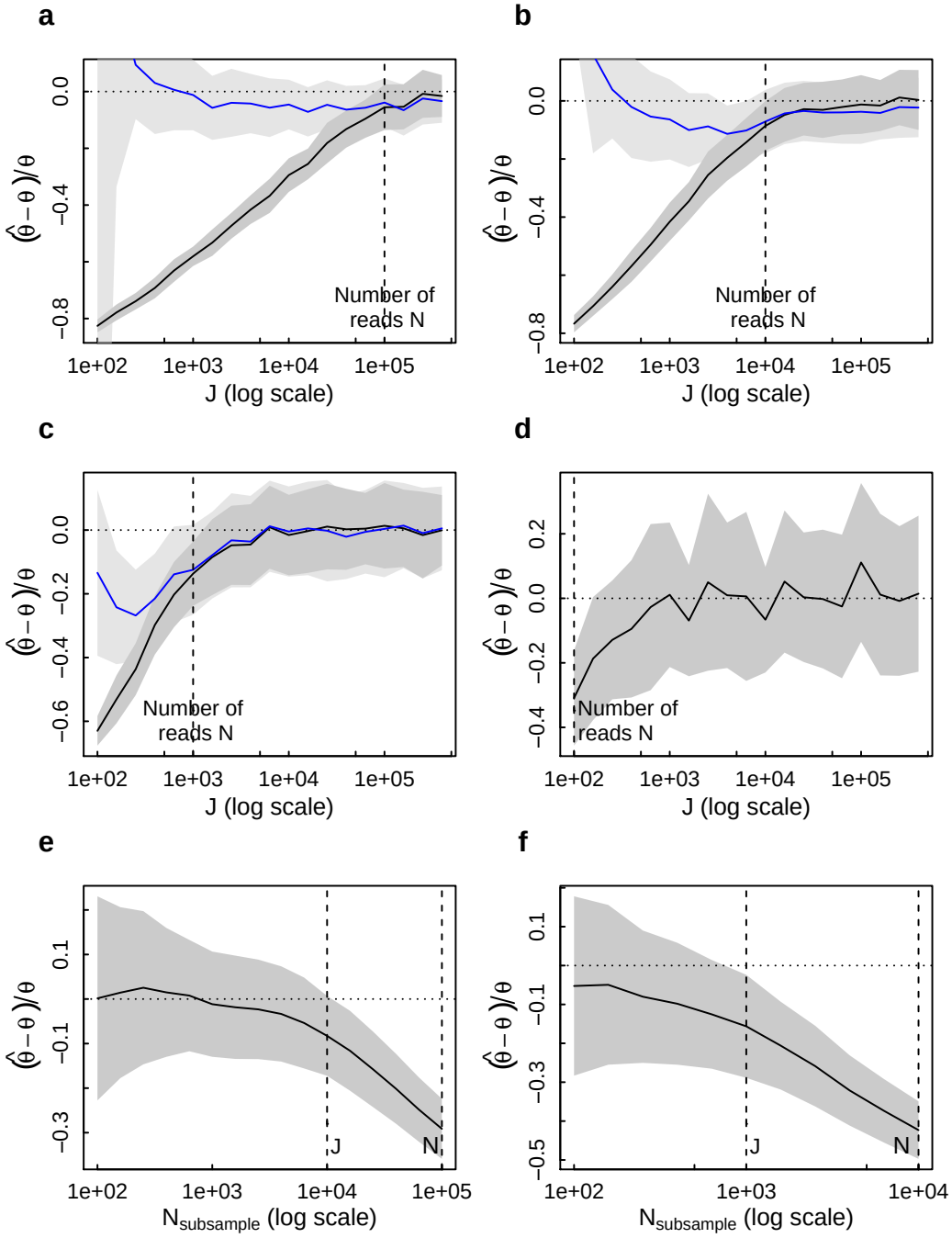

Figure S5: θ inference without dispersal limitation and without simulated noise for $\theta = 20$. The mean and standard deviation of the relative bias on the θ estimate are plotted over 100 realizations. **Panels a-d:** θ inference by maximum likelihood (black) and by linear regression on the ranked log-abundance (blue), as a function of the initial number of individuals J, (a) for $N = 10^5$ reads, (b) $N = 10^4$ reads, (c) $N = 10^3$ reads and (d) $N = 10^2$ reads (linear regression too inaccurate to be plotted for $N = 10^2$)**. Panels e-f:** Maximum-likelihood θ estimate as a function of the size $N_{subsample}$ of the read subsample used for estimation, starting from an original sample of (e) $N = 10^5$ reads or (f) $N = 10^4$ reads. An unbiased θ estimate is obtained when $N_{subsample}$ is at least one order of magnitude smaller than the initial number of individuals (e) $J = 10^4$ or (f) $J = 10^3$.

Chapter 3

Topic modelling reveals spatial structure in a DNA-based biodiversity survey

Guilhem Sommeria-Klein[1], Lucie Zinger[1,2], Eric Coissac[3], Amaia Iribar[1], Heidy Schimann[4], Pierre Taberlet[3], Jérôme Chave[1]

[1] Université Toulouse 3 Paul Sabatier, CNRS, IRD, UMR 5174 Laboratoire Evolution et Diversité Biologique (EDB), F-31062 Toulouse, France.

[2] Ecole Normale Supérieure, CNRS, UMR 8197 Institut de Biologie de l'ENS (IBENS), F-75005 Paris, France.

[3] Université Grenoble Alpes, CNRS, UMR Laboratoire d'Ecologie Alpine (LECA), F-38000 Grenoble, France.

[4] INRA, UMR 745 EcoFoG (AgroParisTech, CIRAD, CNRS, University of the French West Indies, University of French Guiana), F-97387 Kourou, France.

Chapter outline

The second chapter explored the effect of noise on the interpretability of environmental DNA data. In this third chapter, another challenge of environmental DNA data is addressed, namely the fact that microbial datasets typically yield a large number of rare OTUs, and that sampling effort cannot be controlled across samples. As in the first chapter, the focus is here on datasets containing many spatially distributed samples. However, while the first chapter aimed at comparing the taxonomic composition of samples with respect to their spatial layout and to environmental descriptors, this chapter describes a method to explore the structure of an environmental DNA dataset independently of any additional information. The results can then be interpreted in regard of contextual data. This method, which is closely related to methods already in use in microbiology, is suited to large and sparse datasets, and accounts for sampling effects. It consists in decomposing the data into assemblages of OTUs based on their propensity to co-occur across samples. In this chapter, it is tested using simulations and by applying it to a large soil DNA dataset collected over a forest plot following a regular sampling scheme. A measure of the stability of the decomposition is also proposed. Lastly, the application of this approach to ecological data is discussed more generally. Of particular interest is that this method is model-based, and could thus be extended by modifying the underlying model, including by the addition of more mechanistic elements.

Abstract

High-throughput sequencing of amplicons from environmental DNA samples has become a major method for rapid, standardized and comprehensive biodiversity assessments, allowing for the study of all life forms within a single sample. However, data interpretation is often difficult because a large number of rare taxa confound patterns. Hence, retrieving and describing the structure of such datasets requires efficient methods for dimensionality reduction. Here, we describe the first application of Latent Dirichlet Allocation (LDA) to an environmental DNA dataset. LDA uses a probabilistic model to decompose samples into overlapping assemblages based on the co-occurrence of taxa and the covariance of their abundances. It accounts for sampling effects and accommodates large and sparse datasets. We show that the grouping of taxa into assemblages can be tested statistically, and to this end develop a measure of assemblage stability. We then apply a LDA algorithm to a large soil survey of bacteria, protists and metazoans in a 12-ha plot of primary tropical forest. The LDA analysis reveals that bacterial and protist assemblages display a strong spatial structure while metazoans do not. Furthermore, bacteria and protists exhibit very similar spatial patterns, which match the topographical features of the plot. We conclude that LDA is a computationally efficient and robust method to detect and interpret the structure of large DNA-based biodiversity datasets. We discuss the possible future applications of this approach in biodiversity science.

Introduction

High-throughput sequencing is shedding a new light on the study of biodiversity patterns across domains of life. A simple and efficient method is 'DNA metabarcoding' (Taberlet *et al.*, 2012), which consists in amplifying and sequencing a genomic marker ('DNA barcode') in the DNA contained in environmental samples such as soil, water or feces (Thomsen & Willerslev, 2015). The resulting sequences can then be clustered into molecular Operational Taxonomic Units (OTUs), which serve as proxies of species in biodiversity assessments, and which can possibly be assigned to known taxa after comparison to reference databases. Metabarcoding data typically consist of a 'community matrix' that lists the OTUs found in each environmental sample, as well as their read counts.

A goal of community ecology is to understand patterns of species co-occurrence and turnover across space. Let us assume that many samples have been collected across space, in a regular fashion. So far, the search for community structure has been performed using multivariate ordination, as well as distance-based or partitioning-based clustering (Legendre & Legendre, 2012). These methods have proven their efficiency, but they have limitations when it comes to analysing datasets with a very large number of OTUs, and many rare OTUs, resulting in large and sparse community matrices (Holmes *et al.*, 2012). Their results are also biased by the uneven sampling effort across samples in metabarcoding data, since sampling effort depends on the amount of DNA retrieved and on PCR yield for each sample.

Probabilistic approaches to detecting data structure offer an alternative to ordination methods by explicitly modelling the sampling process that underlies the data (Holmes *et al.*, 2012). This can be achieved using a so-called mixture model, which assumes that the data are structured into a mixture of several (unobserved) component units, each with a distinctive taxonomic composition. Under this model, the observed discrete samples of sequence reads, which may be of different sizes, are sampled from

this mixture. The component units can then be inferred from the data using maximum-likelihood or Bayesian inference, which provide rigorous means of assessing goodness-of-fit and of selecting the number of component units. Mixture models have been successfully used in microbiology (Knights *et al.*, 2011; Holmes *et al.*, 2012; Ding & Schloss, 2014; Shafiei *et al.*, 2015) and in community ecology (Valle *et al.*, 2014), either in an unsupervised way (data clustering) or in a supervised way (data classification). In particular, Valle et al. (2014) used Latent Dirichlet Allocation (LDA) to cluster tree abundance data across forest plots into component assemblages – or 'component communities'. They showed that this method performed better than hierarchical and k-means clustering on simulated data. Here, we explore the potential of this method for the analysis of large metabarcoding datasets.

LDA decomposes samples into a mixture of component assemblages, which may themselves overlap in their taxonomic composition. The component assemblages can be interpreted as communities of co-occurring taxa. Because each sample is represented by a mixture of component assemblages, the model captures the smooth turnover in species composition along environmental gradients (Valle *et al.*, 2014). This model was originally introduced by Blei *et al.* (2003) to decompose large sets of text documents into topics (a problem known as 'topic modelling'), based solely on their word frequency, and has been subsequently extended to the analysis of large and complex datasets in various fields (see Blei (2012) for a review). The same model has been independently introduced in population genetics to model population structure using the distribution of alleles across individuals, and is now a cornerstone of population genetics analyses (model with admixture in the Structure software; Pritchard *et al.*, 2000).

One issue for the application of LDA to metabarcoding is that the interpretation that can be made of abundance information, i.e. the DNA read count per OTU, remains debated (Nguyen *et al.*, 2015; Sommeria-Klein *et al.*, 2016). For bacteria, it seems possible to relate the read count to the number of cells in the sample (Kembel *et al.*, 2012), while in the case of macro-organisms, the read count may be indicative of the taxon's biomass in the environment (Andersen *et al.*, 2012; Klymus *et al.*, 2015). Nevertheless, metabarcoding data are often best used as occurrence data, and it is thus important to evaluate the applicability of LDA to occurrence-based datasets. Second,

depending on how strongly structured the data are, the LDA algorithm may fail to converge to an optimal solution. It is indeed acknowledged in the literature that the result of LDA decomposition may vary from one run to the other (Steyvers & Griffiths, 2007; Balagopalan, 2012; Valle *et al.*, 2014). Hence, it would be important to quantify the robustness of the LDA decomposition, especially since environmental DNA data are noisy. We first address these problems on simulated data, and then turn to the analysis of an empirical metabarcoding dataset describing the soil biodiversity of bacteria, protists and metazoans over a large tropical forest plot in French Guiana (Zinger *et al.*, 2017). We thus address here the following questions: (1) can LDA accurately retrieve assemblages from occurrence data, (2) can we define a stability metric for the decomposition of metabarcoding data into component assemblages, and (3) can component assemblages retrieved from empirical data be related to variation in abiotic conditions? Finally, we discuss our results in light of those obtained by multivariate methods (Zinger *et al.*, 2017).

Methods

1. Latent Dirichlet Allocation

LDA decomposition takes as an input a community matrix representing samples by columns and OTUs by lines, where the entries are the read counts per OTU in each sample. Occurrence data can also be provided as an input, since they are a special case of abundance data where OTU abundances only take values 0 or 1. Inference consists in fitting a generative model to the observed community matrix. The generative model describes a way to generate the data based on two assumptions: the data are structured into K assemblages, where K is a fixed parameter, and each sample is a mixture of the K assemblages in Dirichlet-distributed proportions. The model involves unobserved ('latent') variables describing the underlying decomposition of the data into the K assemblages, and the fitting process consists in estimating the most likely value of the latent variables and of the model's parameters given the observed data (Fig. 1).

The generative model consists of the following steps. For sequence read n in sample m, assemblage membership z_n is generated by a categorical draw from a vector of K mixture weights $(\theta_k^m)_{k\in[\![1,K]\!]}$ (i.e., one out of K categories is chosen at random with probability weights $(\theta_k^m)_{k\in[\![1,K]\!]}$). Then, the OTU membership w_n is generated by a categorical draw from a vector of V mixture weights $\left(\phi_v^{z_n}\right)_{v\in[\![1,V]\!]}$, where V is the number of distinct OTUs in the whole dataset. The mixture weights θ_k^m represent the decomposition of each sample m into the K assemblages, while the mixture weights ϕ_v^k represent the taxonomic composition of each assemblage k. The model further assumes that the mixture weights θ_k^m follow for each sample m a symmetric Dirichlet distribution of mixing parameter α. Therefore, for each sample m:

$$\boldsymbol{\theta^m} = (\theta_k^m)_{k\in[\![1,K]\!]} \sim Dirichlet(\alpha)$$

And then, for each sequence read n in sample m:

$$z_n \sim Categorical(\boldsymbol{\theta^m})$$

$$w_n \sim Categorical(\boldsymbol{\phi}^{\boldsymbol{z_n}})$$

Thus, fitting the generative model to the observed data consists in finding the most likely assemblage mixtures $\boldsymbol{\theta}^{\boldsymbol{m}}$ for the M samples, the most likely OTU compositions $\boldsymbol{\phi}^{\boldsymbol{k}}$ for the K assemblages, and the most likely value for the mixing parameter α of the Dirichlet distribution. The value of α indicates whether the samples tend to be decomposed into an even mixture of component assemblages with similar abundances (case $\alpha > 1$) or into an uneven mixture dominated by one or a few component assemblages (case $\alpha < 1$). A sharp spatial segregation of the assemblages is associated with a α value markedly lower than unity. The Dirichlet distribution is used as a prior primarily because it is the conjugate prior of the categorical distribution, which eases analytical calculations.

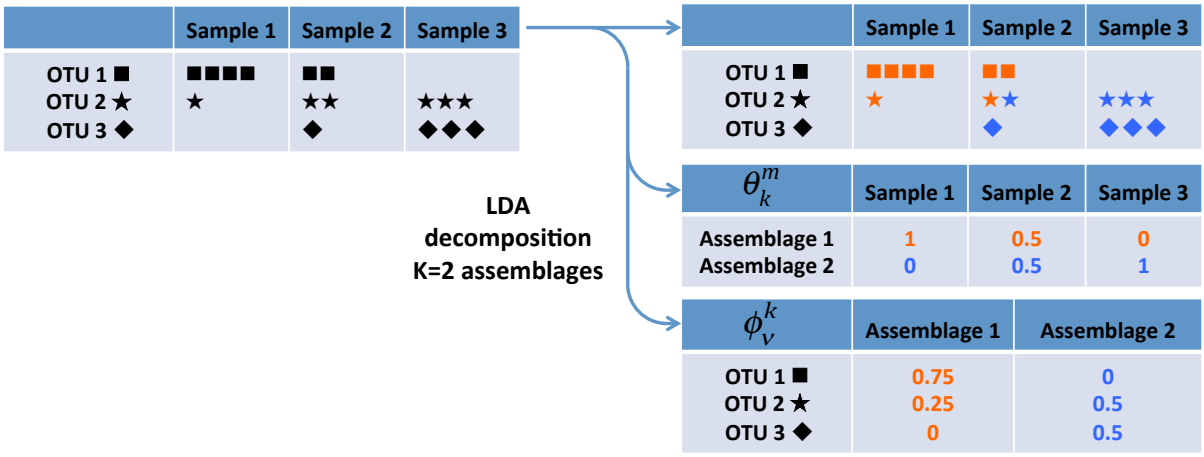

Figure 1. Illustration of Latent Dirichlet Allocation's (LDA) principle. LDA decomposes a community matrix with discrete abundance information (e.g., read count) into K assemblages based on the co-occurrence of OTUs and the covariance of their abundances across samples. K is fixed beforehand and can be selected using likelihood-based model selection methods. The assemblage mixture $(\theta_k^m)_{k\in[\![1,K]\!]}$ in each sample m, with $\sum_{k=1}^{K}\theta_k^m = 1$, and the taxonomic composition $\left(\phi_v^k\right)_{v\in[\![1,V]\!]}$ of each assemblage k, with $\sum_{v=1}^{V}\phi_v^k = 1$, are inferred from the data.

2. Inference using a Variational Expectation-Maximization algorithm

We fitted the generative model to the observed data using the Variational Expectation-Maximization (VEM) algorithm proposed and implemented by Blei *et al.* (2003), and

wrapped into the R package 'topicmodels' by Grün & Hornik (2011). Compared to the often-followed Bayesian approach of Griffiths & Steyvers (2004), this approach is computationally faster, estimates all parameters and allows for a better-justified use of AIC for selecting the number of assemblages. The algorithm uses approximate likelihood maximization to estimate the parameters α and $\boldsymbol{\phi} = \left((\phi_v^k)_{v\in[\![1,V]\!]}\right)_{k\in[\![1,K]\!]}$, as well as the posterior distribution of the latent variables $\boldsymbol{z} = \left((z_n)_{n\in[\![1,N_m]\!]}\right)_{m\in[\![1,M]\!]}$ and $\boldsymbol{\theta} = \left((\theta_k^m)_{k\in[\![1,K]\!]}\right)_{m\in[\![1,M]\!]}$ given the data $\boldsymbol{w} = \left((w_n)_{n\in[\![1,N_m]\!]}\right)_{m\in[\![1,M]\!]}$.

First, we set the model parameters to $\alpha = 0.1$ and to randomly chosen values for $\boldsymbol{\phi}$. Then, the following two steps are repeated until the likelihood (or more precisely, a lower bound for the likelihood) converges. The variational step approximates the posterior distribution $P(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{w}, \alpha, \boldsymbol{\phi})$ of $\boldsymbol{z}$ and $\boldsymbol{\theta}$, given the data $\boldsymbol{w}$ and given the current values of α and $\boldsymbol{\phi}$. This is achieved by minimizing the Kullback-Leibler divergence between a variational approximation and the true posterior. The Expectation-Maximization (EM) step estimates the parameters α and $\boldsymbol{\phi}$ by maximizing the marginal log-likelihood $L(\alpha, \boldsymbol{\phi}) = \ln[P(\boldsymbol{w}|\alpha, \boldsymbol{\phi})]$, making use of the approximation to the posterior distribution $P(\boldsymbol{z}, \boldsymbol{\theta}|\boldsymbol{w}, \alpha, \boldsymbol{\phi})$ found in the variational step (Blei *et al.*, 2003; Grün & Hornik, 2011). We used a convergence threshold of 10^{-7} for the EM step and a convergence threshold of 10^{-8} for the variational step in all our analyses.

This algorithm provides an estimate of the marginal log-likelihood $\ln[P(\boldsymbol{w}|\alpha, \boldsymbol{\phi})]$ of the final decomposition, that can be used to compare different realizations of the algorithm or to compute the model's AIC. It is a deterministic algorithm in the sense that it consists in a simple iterative optimization. However, the result may depend on the initialization for the taxonomic composition $\boldsymbol{\phi}$ of assemblages.

3. Computing the optimal number of assemblages

We selected the number K of assemblages based on AIC. There is no rigorous expression of AIC for a model such as LDA (Burnham & Anderson, 2002), but we chose to compute the AIC as $2(L(\alpha, \boldsymbol{\phi}) + K(V-1) + 1)$, where $L(\alpha, \boldsymbol{\phi}) = \ln[P(\boldsymbol{w}|\alpha, \boldsymbol{\phi})]$ is the marginal log-likelihood of the LDA decomposition. Indeed, there are $K(V-1)$ free parameters to

be estimated in $\boldsymbol{\phi} = \left((\phi_v^k)_{v\in[\![1,V]\!]}\right)_{k\in[\![1,K]\!]}$, plus the mixing parameter α. This is the same expression as the one used elsewhere (Than & Ho, 2012). We used the lower bound on the marginal log-likelihood computed as part of the VEM algorithm as an approximation for $L(\alpha, \boldsymbol{\phi})$. We also tried to correct the AIC for small sample size as $2[L(\alpha, \boldsymbol{\phi}) + (K(V-1)+1)\left(1+\frac{1}{M}\right)]$, where M is the number of samples (Burnham & Anderson, 2002), but this did not modify our results, and we do not report these analyses here.

4. Assessing the stability of the decomposition

The LDA decomposition reflects the co-occurrence structure of OTUs among samples, as well as the covariance structure of their abundances in the case of abundance data. If the data are not strongly structured, they may exhibit a complex likelihood landscape, which increases the chance that the algorithm reaches a local likelihood maximum. To address this issue, we ran the algorithm a hundred times starting from random initial assemblages $\boldsymbol{\phi^k}$, and we selected only the realization with the highest likelihood value for interpretation. We also measured the stability of the decomposition across the hundred realizations, with two goals in mind: measuring how strongly structured the data are, and assessing whether the realization with highest likelihood has indeed reached the optimal solution. We removed the occasional realizations with α values much larger than 1 from the analysis, because they correspond to non-informative solutions where all samples contain all assemblages in similar proportions.

To measure the stability of the decomposition across realizations, we first needed to define a measure of similarity between two possible decompositions of the data. We computed it as the mean similarity between the assemblages of the two decompositions. Therefore, it boils down to defining a measure of similarity between two assemblages. We used the symmetrised Kullback-Leibler (sKL) divergence, a measure of dissimilarity between two distributions that stems from information theory and that is commonly used in statistics and machine learning (Burnham & Anderson, 2002; Meila, 2006; Steyvers & Griffiths, 2007). The Kullback-Leibler divergence (or relative entropy) of a distribution $\boldsymbol{q} = (q_i)_{i\in[\![1,N]\!]}$ relative to a distribution $\boldsymbol{p} = (p_i)_{i\in[\![1,N]\!]}$ is defined as

$D(\boldsymbol{p}|\boldsymbol{q}) = \sum_{i=1}^{N} p_i \ln(p_i/q_i)$, with $\sum_{i=1}^{N} p_i = 1$ and $\sum_{i=1}^{N} q_i = 1$ (Kullback, 1959). It measures the amount of information lost when approximating the distribution $\boldsymbol{p}$ by the distribution $\boldsymbol{q}$. The symmetrised Kullback-Leibler divergence between $\boldsymbol{p}$ and $\boldsymbol{q}$ is then defined as $D_{\mathrm{s}}(\boldsymbol{p},\boldsymbol{q}) = \big(D(\boldsymbol{p}|\boldsymbol{q}) + D(\boldsymbol{q}|\boldsymbol{p})\big)/2$. Between two assemblages k_1 and k_2, the sKL divergence can be computed either based on their spatial distribution, i.e. $D_{\mathrm{s}}\big(\boldsymbol{\theta_{k_1}}/\sum_{m=1}^{M}\theta_{k_1}^{m}, \boldsymbol{\theta_{k_2}}/\sum_{m=1}^{M}\theta_{k_2}^{m}\big)$, or based on their OTU composition, i.e. $D_{\mathrm{s}}\big(\boldsymbol{\phi^{k_1}}, \boldsymbol{\phi^{k_2}}\big)$. Thus, we were able to measure both the spatial and the taxonomic similarity between two assemblages. Since $D_{\mathrm{s}}(\boldsymbol{p},\boldsymbol{q})$ is infinite as soon as there is at least one i in $[\![1, N]\!]$ that verifies $p_i = 0$ or $q_i = 0$, we avoided infinite sKL divergence values by setting a lower bound in every entry of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, equal to the inverse of the sum of all elements in the community matrix (i.e., the inverse of the total number of reads in the case of abundance data, or the inverse of the total number of occurrences in the case of occurrence data). Therefore, every point where both distributions take values below this threshold has a null contribution to $D_{\mathrm{s}}(\boldsymbol{p},\boldsymbol{q})$.

We used the sKL divergence to define the similarity measure $\sigma(k_1,k_2) = \big(\langle D_{\mathrm{s}}(k_1,k_2)\rangle_{\mathrm{rnd}} - D_{\mathrm{s}}(k_1,k_2)\big)/\langle D_{\mathrm{s}}(k_1,k_2)\rangle_{\mathrm{rnd}}$ between two assemblages k_1 and k_2, where $\langle D_{\mathrm{s}}(k_1,k_2)\rangle_{\mathrm{rnd}}$ is the average sKL divergence over 1000 randomizations of the assemblages. When computing spatial similarity, we performed randomizations by randomly shifting the spatial distribution of one assemblage with respect to the other, so as to account for spatial autocorrelation (Fortin & Payette, 2002). When computing taxonomic similarity, we performed random permutations of the OTUs in one distribution with respect to the other. The similarity $\sigma(k_1,k_2)$ is equal to 1 for a perfect match, and to 0 when the assemblages are as similar as expected by chance. We then defined the similarity between two decompositions d_1 and d_2 as the mean similarity between their best-matching assemblages, i.e. $S(d_1,d_2) = \sum_{k_1=1}^{K}\sigma\big(k_1, k_2^*(k_1)\big)/K$, where assemblage $k_2^*(k_1)$ is the best match in decomposition d_2 of assemblage k_1 in decomposition d_1, as deduced from the comparison of σ values. When more than one assemblage k_1 in decomposition d_1 had a best match with assemblage k_2^* in decomposition d_2, we forced a one-to-one correspondence between the assemblages of both decompositions by giving priority to higher σ values. This situation should be

rarely encountered however, since assessing stability mostly involves comparing decompositions that closely resemble each other.

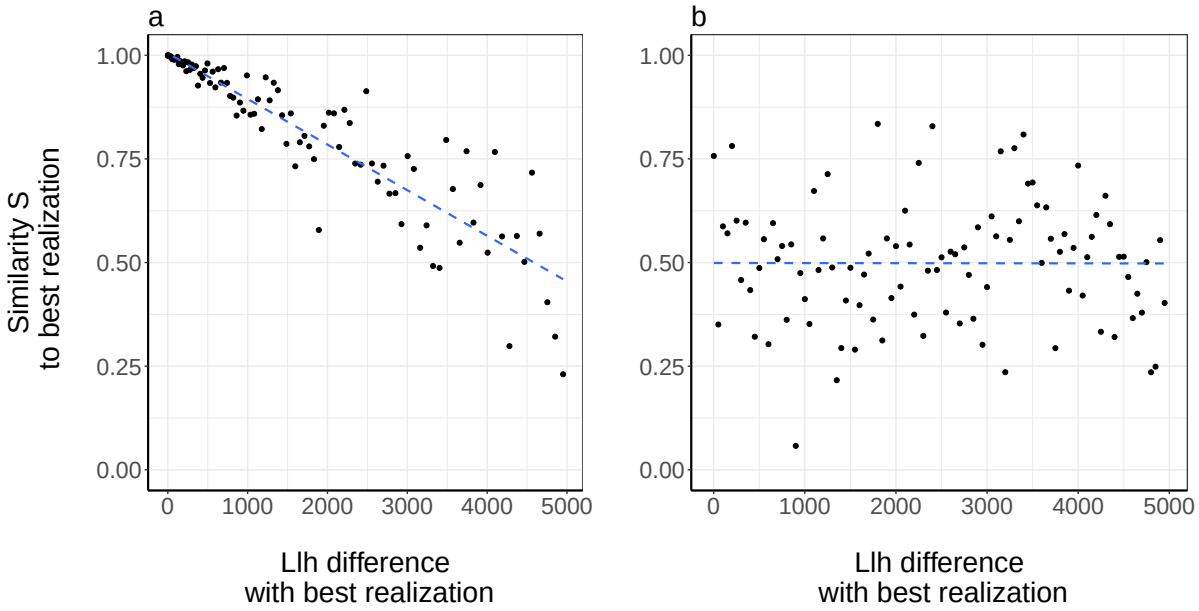

Figure 2. Assessing the stability of the LDA decomposition using the metric *I*. Each panel represents 100 realizations of the LDA algorithm with random assemblage initializations for a mock dataset. In both cases, the realization with highest likelihood (i.e., the best realization) is compared to each of the 99 others by plotting their similarity S as a function of their log-likelihood difference. The metric I is defined as the intercept of the linear regression (dashed blue line). Two cases are illustrated: (a) realizations grow increasingly similar to the best realization as their likelihood increases ($I = 1$), and (b) dissimilar realizations with similar likelihood coexist ($I = 0.5$). Values of I close to 1 indicate that the best realization is likely to have reached the optimal solution.

We measured the stability of the decomposition across $n = 100$ realizations by computing two metrics. First, we computed the mean similarity across all pairs of realizations $\langle S \rangle_n = \sum_{d_1,d_2} S(d_1, d_2)/(n(n-1)/2)$. The more similar the realizations are irrespective of the initial condition, the more strongly structured the data are likely to be. Second, we compared the realization with highest likelihood (i.e., the best realization) to each of the $n - 1$ others. To assess whether the best realization had indeed reached the optimal solution, we plotted for each pair their similarity S as a function of their log-likelihood difference (Fig. 2). We performed a linear regression of the similarity against the log-likelihood difference, and used the intercept I_n as a metric.

This metric I_n assesses whether the realizations tend to be increasingly similar to the best realization as their likelihood increases, i.e. to 'converge' toward the best realization. Values of I_n close to 1 mean that we can be confident that the best realization has reached the global likelihood maximum, provided that the space of possible initializations has been adequately sampled. We computed both metrics for spatial ($\langle S_{\text{spat.}} \rangle_n$, $I_{\text{spat.},n}$) and taxonomic ($\langle S_{\text{taxo.}} \rangle_n$, $I_{\text{taxo.},n}$) similarities between assemblages.

5. Simulated data

To test the performance of the LDA algorithm on occurrence-transformed data with respect to the original abundance data, we simulated a metabarcoding dataset. This simulated dataset comprised 1,131 samples containing a total of 1,000 OTUs and decomposed into 5 assemblages. We first defined the assemblages by drawing their OTU composition from a Dirichlet distribution of mixing parameter 0.02. We then assigned to each sample a mixture of assemblages in proportions determined by a sinusoidal function of the sample's index, so that the relative abundances of all 5 assemblages successively peak at 100% (Fig. 3). Combining the assemblage mixture and the taxonomic composition of assemblages, we obtained the relative abundances of OTUs in each sample. We generated the simulated dataset by sampling 1,000 sequence reads per sample from these relative abundances, which resulted in an average diversity of 105 OTUs per sample.

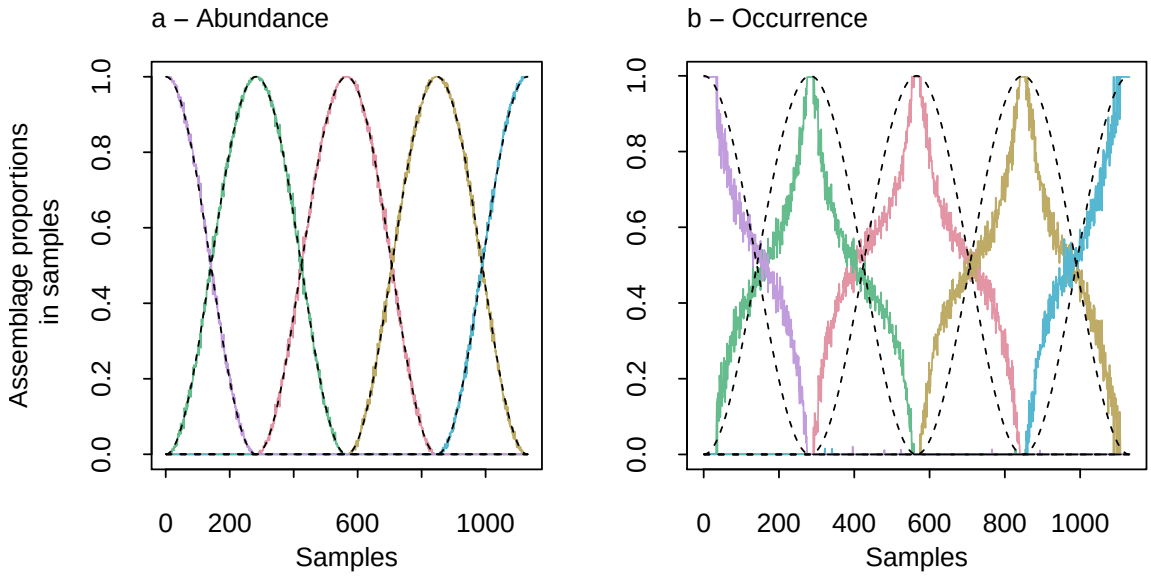

Figure 3: LDA decomposition of simulated occurrence and abundance data. LDA applied to a simulated dataset with 5 assemblages, 1,000 MOTUs, 1,131 samples, and 1,000 sequence reads per sample, (a) for the original abundance data, and (b) for the occurrence data derived from the same dataset. Each plot shows the assemblage proportions estimated by LDA for $K = 5$ (coloured lines; only the realization with highest likelihood out of 100 is shown) and the simulated assemblage proportions (dashed black lines).

6. Tropical forest soil metabarcoding dataset

We applied LDA to an empirical metabarcoding dataset describing the biodiversity of bacteria, protists and metazoans over a 300x400 m tropical forest plot (called Petit Plateau; Chave *et al.*, 2008) at the Nouragues Ecological Research Station, in a lowland tropical forest of central French Guiana (Bongers *et al.*, 2001). Site conditions, data collection, laboratory procedures, and sequencing filtering procedures are all described in detail in Zinger et al. (2017) and are only briefly summarized here.

The sampling campaign was conducted towards the end of the 2012 dry season. Soil samples were collected from the mineral horizon (~10 cm deep) using a soil auger every 10 m on a square grid covering the plot and excluding the edges, which resulted in 1,131 soil samples (Fig. S1). Extracellular DNA was extracted in the field from each sample (Zinger *et al.*, 2016). The present study uses data from two DNA barcodes

amplified by PCR and sequenced on high-throughput Illumina sequencers, targeting bacteria (16S rDNA), and all eukaryotes (18S rDNA). The sequencing data were curated using the OBITools package (Boyer *et al.*, 2016). Sequences were clustered into OTUs based on their similarity using the Infomap algorithm (Rosvall *et al.*, 2009) with a similarity cut-off of 3 mismatches, so as to cluster spurious sequences resulting from PCR and sequencing errors. Each OTU was given a taxonomic assignation by comparing its sequence to the following reference databases: GenBank r197 for the eukaryotic 18S marker, and SILVA for the bacterial 16S marker. Sequence matching to databases was conducted using the ecotag program included in the OBITools package. Based on these taxonomic assignations, we further split the eukaryotic 18S dataset into protists, arthropods, annelids, nematodes, and flat worms (Platyhelminthes).

Out of the 1,131 samples, a number of samples were excluded from the sequencing results for each barcode due to insufficient PCR yields (7.2% of samples for bacteria and 0.2% for eukaryotes). We interpolated the content of the missing samples by sampling with replacement the mean number of reads per sample from the (up to eight) non-empty nearest neighbouring samples on the grid. We then applied the LDA on either the read-abundance data, or on the occurrence-transformed data, defining the absence of an OTU in a sample strictly as zero read-abundance in the sample. We did not trim the data for rare OTUs, or for OTUs represented in a single sample.

A fine-grained description of the forest canopy structure and topography was obtained using a small-footprint LiDAR survey carried out over the sampling site in the same year as the soil sampling (2012; Rejou-Mechain *et al.*, 2015). This allowed the generation of maps of topography, slope, and canopy height from the LiDAR cloud of points. The topography of the plot is relatively smooth, with a maximal difference in elevation of 30 m. Maps of soil wetness (Beven & Kirkby, 1979) and light at ground level were also derived from the LiDAR measurements (Tymen *et al.*, 2017). We compared the LiDAR-derived data with the metabarcoding data by computing the mean values of the environmental variables over 10-m-by-10-m cells centred on the soil sampling points. We sought a biological interpretation for the retrieved assemblages by comparing their spatial distribution to the distribution of LiDAR-obtained environmental variables. To do so, we computed Pearson's correlation coefficient between the spatial distributions and assessed the significance of the correlation by

performing 100,000 spatial randomizations, i.e. shifting randomly one spatial distribution with respect to the other, so as to account for spatial autocorrelation.

			$K = 3$				
	Richness	$K_{\min(\mathrm{AIC})}$	$\langle S_{\mathrm{spat.}} \rangle_{100}$	$I_{\mathrm{spat.},100}$	$\langle S_{\mathrm{taxo.}} \rangle_{100}$	$I_{\mathrm{taxo.},100}$	$\alpha_{\mathrm{best\ real.}}$
Bacteria 16S	20,162	5	0.85	1.0	0.95	1.0	0.16
Protists 18S	1,648	2	0.68	1.0	0.95	1.0	0.082
Arthropods 18S	1,881	2	0.62	0.62	0.91	0.93	0.11
Nematodes 18S	378	2	0.33	0.49	0.88	0.94	0.05
Platyhelminthes 18S	126	2	0.52	0.50	0.86	0.88	7.0
Annelids 18S	51	2	0.41	0.57	0.83	0.90	0.035

Table 1: Stability of LDA decomposition for occurrence data. For each of the taxonomic groups under study: total number of MOTUs; optimal number of assemblages $K_{\min(\mathrm{AIC})}$ obtained from AIC minimization; spatial and taxonomic stability for three assemblages as measured by the $\langle S \rangle_{100}$ and I_{100} metrics; estimated value of the mixing parameter α in the best realization out of 100 for three assemblages.

Results

We first applied Latent Dirichlet Allocation decomposition to a simulated dataset, and compared the results for abundance and occurrence data. AIC minimization correctly recovered the simulated number of assemblages (five) in both cases (Fig. S2). In the case of occurrence data, LDA yielded more even assemblage mixtures than simulated (Fig. 3). The algorithm reached the optimal solution more reliably for occurrence data than for abundance data ($\langle \mathrm{S}_{\mathrm{spat.}} \rangle_{100} = 0.98$ for occurrence data, $\langle S_{\mathrm{spat.}} \rangle_{100} = 0.89$ for abundance data, $I_{\mathrm{spat.},100} = 1.0$ in both cases; cf. Fig. S2). Next, we applied the analysis to the tropical forest soil dataset. Using the read-abundance data, the optimal number of assemblages was always larger than 50, while it ranged between 2 and 5 for the occurrence data, depending on the taxonomic group (Table 1). As also observed on the simulated data, the LDA algorithm converged more reliably toward the optimal solution for occurrence data than for abundance data (Table S1). Thus we conclude that LDA can be effectively applied to occurrence-based biodiversity data. In the rest, we describe results obtained using occurrence data and assuming $K = 3$ assemblages, a value close to that minimizing the AIC across taxonomic groups.

We found clear differences when comparing bacteria and unicellular eukaryotes (henceforth denoted as protists) to metazoans (arthropods, annelids, flat worms and nematodes). Bacteria and protists displayed a stronger spatial structure at the scale of our study plot, as deduced from the spatial stability of the decomposition: the similarity intercept $I_{\mathrm{spat.},100}$ was equal to 1.0 (Table 1, Fig. S3), with a mean similarity across realizations $\langle \mathrm{S}_{\mathrm{spat.}} \rangle_{100}$ of 0.85 and 0.68, respectively (Table 1). In contrast, metazoans displayed a lower similarity intercept ($0.49 \leq I_{\mathrm{spat.},100} \leq 0.62$), and also a lower similarity across realizations ($0.33 \leq \langle \mathrm{S}_{\mathrm{spat.}} \rangle_{100} \leq 0.62$). We also found that spatial structure was positively correlated to taxonomic diversity, measured by OTU richness (correlation coefficient $\rho = 0.85$ between $\langle \mathrm{S}_{\mathrm{spat.}} \rangle_{100}$ and the number of OTUs; Fig. 4). The taxonomic stability of the assemblages was higher than their spatial stability, following the same trends as the spatial stability, but with less pronounced differences among taxonomic groups (Table 1).

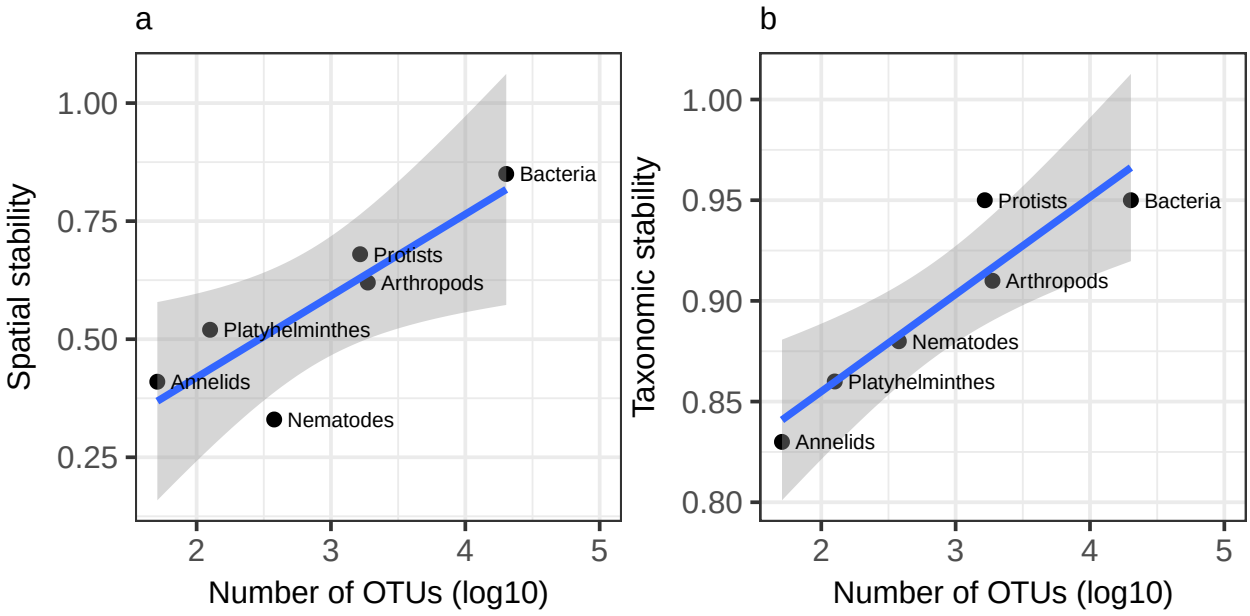

Figure 4. Stability for occurrence data measured as the mean similarity across realizations $\langle S \rangle_{100}$, as a function of the number of OTUs. The metric $\langle S \rangle_{100}$ is measured based on the (a) spatial and (b) taxonomic similarity between assemblages. The blue line figures a linear regression, and the shaded area its standard error. Pearson's correlation coefficient is $\rho = 0.85$ for spatial stability and $\rho = 0.92$ for taxonomic stability.

For all taxonomic groups except flat worms, the estimates of the mixing parameter α were much smaller than 1 (Table 1), indicating a strong spatial segregation among assemblages. In bacteria and protists, the decomposition into three assemblages was strongly linked to topographical features (Fig. 4, Table S2). The blue assemblage of Fig. 4 was associated with terra firme areas, defined as areas of higher topography, gentler slope, and lower soil wetness. The green assemblage was associated with hydromorphic areas, defined as displaying the opposite environmental correlations (Table S2). Finally, the spatial distribution of the red assemblage matched the location of exposed rock patches that are scattered across the forest plot, based on direct observations. In metazoans, we were unable to identify similar terra firme and hydromorphic assemblages (Fig. S4, Table S2), however one assemblage in arthropods and nematodes did match the exposed rock spatial pattern (Table S3). This exposed rock assemblage was indeed consistently found to be the most taxonomically distinctive in all taxonomic groups. Neither light at ground level nor canopy height explained the LDA decomposition in any of the taxonomic groups.

Discussion

Large environmental DNA datasets offer a unique opportunity to unlock some of the major challenges in community ecology, yet as a result data accumulation is accelerating, thus creating the need for novel methods adapted to these data. Here we have presented the potential of the Latent Dirichlet Allocation method for the analysis of metabarcoding data. This model-based method is adapted to large and sparse datasets. It assigns a probability weight for each sample to belong to an assemblage based on the OTUs in this sample, and also infers the composition in OTUs of each assemblage (see Table S4). It thus goes beyond a categorical classification of samples and generates biologically interpretable assemblages. Here, we further elaborate on the advantages and limitations of this approach, and on the implications to the analysis of the forest soil dataset.

Discussing the assumptions of LDA. Unlike in classical multivariate methods, no prior transformation of the data is required: input data consist of discrete OTU abundances, or occurrences, and sample sizes may vary across samples. Input data are not required to meet a normality assumption, the definition of a dissimilarity metric is not required, and LDA thus makes a more parsimonious use of the data. The assumptions made by the underlying model are minimal: the Dirichlet prior is the natural prior for the parameters of the categorical distribution, and it is sufficiently flexible to fit most datasets (O'Brien & Record, 2016). One could take a step toward more mechanistic modelling by adding more assumptions to the LDA approach. For instance, one could assume that a neutral dynamics takes place within assemblages, so that their taxonomic composition follows the taxa-abundance distribution predicted by Hubbell's neutral theory of biodiversity (Hubbell, 2001; Harris *et al.*, 2015). Assuming a Dirichlet prior also on the taxonomic composition of assemblages, as done in the Bayesian version of LDA (Griffiths & Steyvers, 2004; Valle *et al.*, 2014), is a first step in that direction, since the Dirichlet distribution approximates the neutral taxa-abundance distribution for a large number of taxa.

Assessing the robustness of the LDA decomposition and selecting the number of assemblages. In many applications of LDA, the question of the robustness of the decomposition is crucial. However the robustness of the algorithm, as measured by the similarity of the output across runs, has rarely been assessed, probably because it entails a serious computational burden. Here we have proposed a practical way to measure the similarity across runs based on the symmetrised Kullback-Leibler divergence, and have used it to quantify how stable the decomposition is with respect to initialization. We have computed two complementary stability metrics. First, $\langle S \rangle$ measures the mean similarity across pairs of realizations. This stability metric is general since it is not centred on the best realization, and measures how strongly structured the data are. Second, I is the similarity intercept obtained by comparing the highest-likelihood realization to all others through a linear regression of their similarity against their log-likelihood difference. This second stability metric takes account of the likelihood information, is less computationally intensive, and is used to assess whether the realization with highest likelihood has reached the optimal solution.

The symmetrised Kullback-Leibler (sKL) divergence is suited to assessing stability because it is sensitive to small differences between distributions. However, it is unbounded, which makes it difficult to interpret. By normalizing the sKL divergence by its mean value over randomizations, we defined a similarity index σ equal to 1 when the distributions are identical and to 0 when they are no more similar than expected by chance. This index also accounts for spatial autocorrelation in the data by performing spatial randomizations.

To compute the similarity between two decompositions, we consider only the similarity between the best-matching assemblages of both decompositions, thus discarding part of the information. This method works well when the decompositions are similar, however similarity is undesirably low when assemblages are merged or split between the two decompositions. This could be corrected by computing the sKL divergence between the full partitioning of the data in both decompositions, i.e. the assignment of every sequence read to an assemblage, instead of comparing pairs of assemblages. While this is the approach advocated for in the clustering literature (Meila,

2006; Vinh *et al.*, 2010), it would likely be very computationally intensive in the case of LDA.

There is no unique method to select the number of LDA component units (Airoldi *et al.*, 2010). Here we use AIC minimization as an indication of the optimal number of assemblages. Another commonly used method consists in splitting the data into a learning set and a test set, and optimizing the predictive power of the model on the test set, as measured by a perplexity function (Blei *et al.*, 2003). A more sophisticated method is to follow the non-parametric modelling approach of (Teh *et al.*, 2006), where the number of assemblages is modelled as a random latent variable that is estimated from the data. However, this method proved to have convergence issues on our empirical data. Stability of the algorithm's output could also be used as a criterion to select the number of assemblages. When a large number of LDA component units is selected, an additional step of analysis using simpler statistical methods may be needed to represent and interpret the result of the LDA decomposition (Mauch *et al.*, 2015).

Tropical forest soil biodiversity decomposition. By applying LDA to an environmental DNA dataset, we described the spatial structure of bacterial, protist and metazoan soil communities in a 12-ha tropical forest plot. The spatial patterns retrieved by LDA for these taxonomic groups allowed us to shed light on soil community structure (see also Zinger *et al.*, 2017).

We applied the LDA algorithm to metabarcoding data with no further transformation than clustering the sequences to avoid defining spurious OTUs. We verified that the interpolation of missing samples played no role in generating the observed patterns. The AIC minimization yielded between 2 and 5 assemblages for occurrence data depending on the taxonomic group, but we used the value $K = 3$ across groups to facilitate intercomparison and because the LDA decomposition is robust to the number of assemblages close to the optimum (Fig. S7). For the 20,162-OTU bacterial dataset, the largest dataset considered in this study, numerical inference of the LDA decomposition for three assemblages took about 25 minutes for occurrence data and 35 minutes for abundance data, which amounts to respectively 48 and 60 hours when running 100 realizations of the algorithm to test stability.

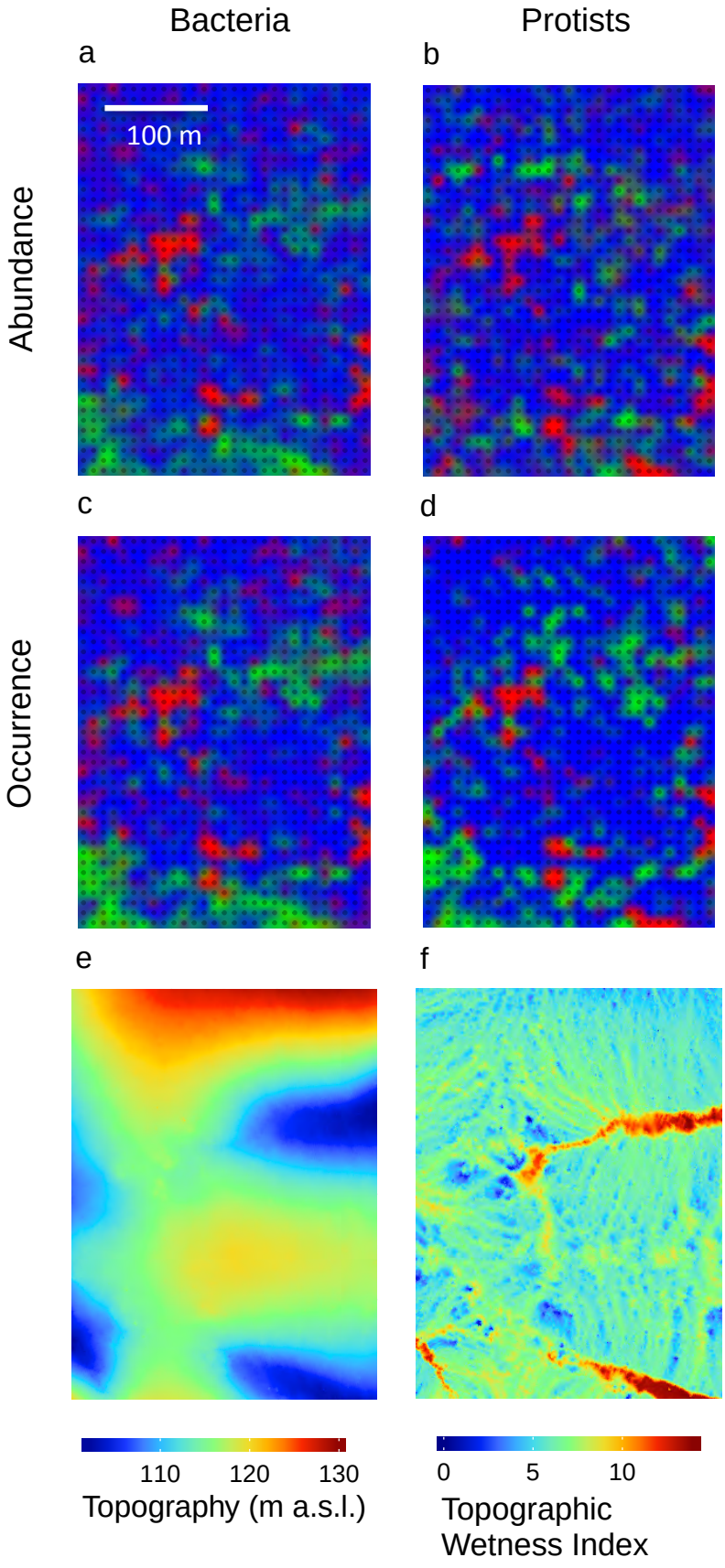

Figure 5: Spatial distribution of microorganism assemblages, for $K = 3$ assemblages.

Spatial distribution of the assemblages obtained from independent LDA decompositions of bacteria and protists, for (a-b) abundance and (c-d) occurrence data. Sampled locations are indicated by dark dots, and the assemblage mixture between samples has been interpolated using ordinary kriging. Terra firme (in blue), hydromorphic (in green) and exposed rock (in red) assemblages can be identified in each taxonomic group, based on correlations to (e-f) Lidar-derived topography, Topographic Wetness Index and slope, as well as on field observations. The spatial patterns retrieved for abundance data are similar to those obtained with occurrence data but less strongly correlated to topographic variables.

The stability analysis of the algorithm indicates that communities of unicellular organisms (i.e. bacteria and protists) are markedly structured at the scale of the plot, while metazoan communities are less so. The stability of the decomposition is also strongly correlated with the number of OTUs, which spans several orders of magnitude across taxonomic groups (Fig. 4, Table 1). Thus, the lower statistical power in groups containing fewer OTUs could explain this pattern. However, it is more likely due to ecological differences between groups. Indeed, this pattern is confirmed by Zinger et al. using ordination-based variation partitioning between environmental and spatial components.

Furthermore, the two unicellular organism groups can each be decomposed into three spatially segregated assemblages matching plot topography. While the covariation of microorganism composition with topography was already detected in Zinger et al., spatial patterns can here be directly represented under the form of assemblages that are characteristic of the different topographic conditions (Fig. 5, Table S4). These spatial patterns can also be shown to be similar between bacteria and protists, which is both a novel insight and a hint that the assemblages retrieved by LDA do reflect community structure. One assemblage associated with patches of exposed rock was retrieved in bacteria and protists but also in arthropods and nematodes. Its taxonomic composition is particularly distinctive (Fig. S7), which might be explained by the high amount of decaying organic matter retained between the boulders in these patches. A current limitation of LDA is that its ability to compare taxonomic composition to environmental data is limited to computing simple correlations between the spatial distribution of retrieved assemblages and environmental variables. This is in contrast to ordination-based methods such as Canonical Redundancy Analysis, and improving on this aspect would be a useful direction of research.

Using occurrence versus abundance data. The use of occurrence data was computationally faster, and led to more stable and more interpretable patterns. Because biodiversity data typically display a wide range of taxonomic abundances (Fig. S5), switching from abundance to occurrence data amounts to dramatically increasing the weight of rare taxa. In the empirical dataset, these OTUs constitute the bulk of the

diversity: OTUs tallying on average less than one sequence read per sample make up over 85% of the total number of OTUs in bacteria and protists (Fig. S5). They play a significant role in shaping the patterns, since removing them erases the retrieved occurrence-based spatial patterns (Fig. S6). This hints at the importance of rare taxa in defining communities of microorganisms. A possible caveat however is that some of those rare OTUs might be generated by remnant PCR errors in the data. If PCR errors are repeatable for a given DNA sequence, this would produce groups of consistently co-occurring OTUs and thus artificially increase the stability of occurrence-based patterns.

Conclusion. LDA is an efficient method to detect structure in the large and complex datasets generated by environmental DNA sequencing methods. The representation of spatial biodiversity patterns derived from LDA is easily interpretable, and the method comes with a measure of how strongly this representation is supported by the data. LDA could be used to explore the biogeographic patterns arising in larger-scale DNA-based biodiversity surveys such as the Earth Microbiome Project (Gilbert *et al.*, 2014) and the Tara Oceans Project (Sunagawa *et al.*, 2015). It could also be applied in non-spatial sampling designs, such as time series. Lastly, LDA is one example of a family of models, which could for instance find applications in the study of plant-microorganism interactions (Rosen-Zvi *et al.*, 2004). We hope this study will stimulate research on model-based methods of data analysis for the ecological interpretation of environmental DNA studies.

Acknowledgements

We thank Dylan Craven, Bart Haegeman, Hélène Morlon, Tim Paine, Mélanie Roy and Marc-André Selosse for fruitful discussions. We thank Blaise Tymen for his help with the Lidar data. This work has benefited from "*Investissement d'Avenir*" grants managed by the French *Agence Nationale de la Recherche* (CEBA, ref. ANR-10-LABX-25-01 and TULIP, ref. ANR-10-LABX-0041; ANAEE-France: ANR-11-INBS-0001), an additional ANR grant (METABAR project; PI P. Taberlet), funds from CNRS. We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrénées for providing computing resources.

References

Airoldi, E.M., Erosheva, E.A., Fienberg, S.E., Joutard, C., Love, T. & Shringarpure, S. (2010) Reconceptualizing the classification of PNAS articles. *PNAS*, **107**, 20899–20904.

Andersen, K., Bird, K.L., Rasmussen, M., Haile, J., Breuning-Madsen, H., Kjaer, K.H., Orlando, L., Gilbert, M.T.P. & Willerslev, E. (2012) Meta-barcoding of "dirt" DNA from soil reflects vertebrate biodiversity. *Molecular Ecology*, **21**, 1966–1979.

Balagopalan, A. (2012) Improving Topic Reproducibility in Topic Models.

Beven, K.J. & Kirkby, M.J. (1979) A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, **24**, 43–69.

Blei, D. (2012) Probabilstic Topic Models. *Communication of the Association for Computing Machinery*, **55**, 77–84.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, **3**, 993–1022.

Bongers, F., Charles-Dominique, P., Forget, P.-M. & Théry, M. (2001) *Nouragues: dynamics and plant-animal interactions in a Neotropical rainforest*, Springer Science & Business Media.

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P. & Coissac, E. (2016) OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, **16**, 176–182.

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multimodel inference*, Springer, New York.

Chave, J., Olivier, J., Bongers, F., Châtelet, P., Forget, P.-M., van der Meer, P., Norden, N., Riéra, B. & Charles-Dominique, P. (2008) Above-ground biomass and productivity in a rain forest of eastern South America. *Journal of Tropical Ecology*, **24**, 355–366.

Ding, T. & Schloss, P.D. (2014) Dynamics and associations of microbial community types across the human body. *Nature*, **509**, 357–+.

Fortin, M.J. & Payette, S. (2002) How to test the significance of the relation between spatially autocorrelated data at the landscape scale: A case study using fire and forest maps. *Ecoscience*, **9**, 213–218.

Gilbert, J.A., Jansson, J.K. & Knight, R. (2014) The Earth Microbiome project: successes and aspirations. *Bmc Biology*, **12**.

Griffiths, T. & Steyvers, M. (2004) Collapsed Gibbs Sampling for LDA. **101**, 5228–5235.

Grün, B. & Hornik, K. (2011) topicmodels: an R package for fitting topic models.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2015) Linking statistical and ecological theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE*, **PP**, 1–14.

Holmes, I., Harris, K. & Quince, C. (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *Plos One*, **7**.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Kembel, S.W., Wu, M., Eisen, J.A. & Green, J.L. (2012) Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *Plos Computational Biology*, **8**, 11.

Klymus, K.E., Richter, C.A., Chapman, D.C. & Paukert, C. (2015) Quantification of eDNA shedding rates from invasive bighead carp Hypophthalmichthys nobilis and silver carp

Hypophthalmichthys molitrix. *Biological Conservation*, **183**, 77–84.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, **8**, 761-U107.

Kullback, S. (1959) *Information Theory and Statistics*, John Wiley & Sons.

Legendre, P. & Legendre, L. (2012) *Numerical Ecology*, Elsevier.

Mauch, M., MacCallum, R.M., Levy, M. & Leroi, A.M. (2015) The evolution of popular music: USA 1960-2010. *Royal Society open science*, **2**, 150081–150081.

Meila, M. (2006) Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, **98**, 873–895.

Nguyen, N.H., Smith, D., Peay, K. & Kennedy, P. (2015) Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*, **205**, 1389–1393.

O'Brien, J. & Record, N. (2016) The power and pitfalls of Dirichlet-multinomial mixture models for ecological count data. *BioRxiv preprint*.

Pritchard, J.K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

Rejou-Mechain, M., Tymen, B., Blanc, L., Fauset, S., Feldpausch, T.R., Monteagudo, A., Phillips, O.L., Richard, H. & Chave, J. (2015) Using repeated small-footprint LiDAR acquisitions to infer spatial and temporal variations of a high-biomass Neotropical forest. *Remote Sensing of Environment*, **169**, 93–101.

Rosen-Zvi, M., Gri_ffiths, T., Steyvers, M. & Smyth, P. (2004) *The Author-Topic Model for Authors and Documents*.

Rosvall, M., Axelsson, D. & Bergstrom, C.T. (2009) The map equation. *The European Physical Journal Special Topics*, **178**, 13–23.

Shafiei, M., Dunn, K.A., Boon, E., MacDonald, S.M., Walsh, D.A., Gu, H. & Bielawski, J.P. (2015) BioMiCo: a supervised Bayesian model for inference of microbial community structure. *Microbiome*, **3**, 8.

Sommeria-Klein, G., Zinger, L., Taberlet, P., Coissac, E. & Chave, J. (2016) Inferring neutral biodiversity parameters using environmental DNA data sets. *Scientific reports*, **6**.

Steyvers, M. & Griffiths, T. (2007) *Probabilistic Topic Models*. *Latent Semantic Analysis: A Road to Meaning* (ed. by T. Landauer), D. McNamara), S. Dennis), and W. Kintsch), Laurence Erlbaum.

Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Bowler, C., de Vargas, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P. & Tara Oceans, C. (2015) Structure and function of the global ocean microbiome. *Science*, **348**.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. (2006) Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, **101**, 1566–1581.

Than, K. & Ho, T.B. (2012) *Fully Sparse Topic Models*.

Thomsen, P.F. & Willerslev, E. (2015) Environmental DNA - An emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.

Tymen, B., Vincent, G., Courtois, E.A., Heurtebize, J., Dauzat, J., Marechaux, I. & Chave, J. (2017) Quantifying micro-environmental variation in tropical rainforest understory at landscape scale by combining airborne LiDAR scanning and a sensor network. *Annals of Forest Science*, **2**, 1–13.

Valle, D., Baiser, B., Woodall, C.W. & Chazdon, R. (2014) Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, **17**, 1591–1601.

Vinh, N.X., Epps, J. & Bailey, J. (2010) Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, **11**, 2837–2854.

Zinger, L., Chave, J., Coissac, E., Iribar, A., Louisanna, E., Manzi, S., Schilling, V., Schimann, H., Sommeria-Klein, G. & Taberlet, P. (2016) Extracellular DNA extraction is a fast, cheap and reliable alternative for multi-taxa surveys based on soil DNA. *Soil Biology and Biochemistry*, **96**, 16–19.

Zinger, L., Taberlet, P., Schimann, H., Bonin, A., Boyer, F., De Barba, M., Gaucher, P., Gielly, L., Giguet-Covex, C., Iribar, A., Rejou-Mechain, M., Raye, G., Rioux, D., Schilling, V., Tymen, B., Viers, J., Zouiten, C., Thuiller, W., Coissac, E. & Chave, J. (2017) Soil community assembly varies across body sizes in a tropical forest. *bioRxiv*.

	Spatial stability (K=3)				Taxonomic stability (K=3)			
	Abundance data		Occurrence data		Abundance data		Occurrence data	
	$\langle S_{spat.} \rangle_{100}$	$I_{spat.,100}$	$\langle S_{spat.} \rangle_{100}$	$I_{spat.,100}$	$\langle S_{taxo.} \rangle_{100}$	$I_{taxo.,100}$	$\langle S_{taxo.} \rangle_{100}$	$I_{taxo.,100}$
Bacteria 16S	0.88	1.0	0.85	1.0	0.99	1.0	0.95	1.0
Protists 18S	0.72	0.87	0.68	1.0	0.92	0.96	0.95	1.0
Arthropods 18S	0.46	0.65	0.62	0.62	0.65	0.78	0.91	0.93
Nematodes 18S	0.43	0.67	0.33	0.49	0.69	0.87	0.88	0.94
Platyhelminthes 18S	0.45	0.69	0.52	0.50	0.66	0.83	0.86	0.88
Annelids 18S	0.63	0.81	0.41	0.57	0.75	0.85	0.83	0.90

Table S1: Stability of LDA decomposition for occurrence and abundance data. For each of the taxonomic groups under study, spatial and taxonomic stability for three assemblages as measured by the $\langle S \rangle_{100}$ and I_{100} metrics, for abundance and occurrence data.

		Topography	Wetness	Slope
Terra firme	**Bacteria 16S**	**0.36****	**-0.27****	**-0.26****
	Protists 18S	**0.23****	**-0.15****	**-0.17****
	Arthropods 18S	**0.18****	**-0.16*****	-0.027
	Nematodes 18S	**0.15****	**-0.091****	**-0.081****
	Platyhelminthes 18S	**0.12****	**-0.11****	-0.043
	Annelids 18S	0.023	0.042	**-0.091***
Hydromorphic	**Bacteria 16S**	**-0.43*****	**0.40*****	**0.31****
	Protists 18S	**-0.23****	**0.10***	**0.21*****
	Arthropods 18S	**-0.097***	**0.10****	-0.015
	Nematodes 18S	**-0.096*****	**0.10****	0.045
	Platyhelminthes 18S	0.044	-0.099**	0.0087
	Annelids 18S	-0.057	0.058	0.022
Exposed rock	**Bacteria 16S**	0.00024	-0.078**	0.0084
	Protists 18S	-0.052	0.084	-0.025
	Arthropods 18S	-0.12*	0.083	0.070
	Nematodes 18S	-0.071	-0.018	0.049
	Platyhelminthes 18S	-0.14**	0.19**	0.027
	Annelids 18S	0.0098	-0.072*	0.075**

Table S2: Correlation coefficients between the spatial distribution of assemblages and abiotic variables. p-values p were computed based on 100,000 spatial randomizations. Significant correlation coefficients are indicated by *,**,*** ($p < 0.05, p < 0.01, p < 0.001$), and additionally by bold font when they are consistent with a hydromorphic or terra firme interpretation. Taxonomic groups in bold are those that can be assigned a 'terra firme' or 'hydromorphic' label based on correlations to topography, wetness and slope, or an 'exposed rock' label based on correlation to the 'exposed rock' bacterial assemblage (see Table S3).

		Bacteria 16S		
		Terra firme	Hydromorphic	Exposed rock
1st assemblage - Terra firme	**Protists 18S**	**0.76*** **	-0.40***	-0.53***
	Arthropods 18S	0.23***	0.12**	-0.43***
	Nematodes 18S	0.24***	-0.19***	-0.098***
	Platyhelminthes 18S	0.32***	-0.10**	-0.29***
	Annelids 18S	0.23***	0.0046	-0.29***
2nd assemblage - Hydromorphic	**Protists 18S**	-0.45***	**0.51*** **	0.022
	Arthropods 18S	0.16***	-0.21***	0.022
	Nematodes 18S	0.10***	0.16***	-0.29***
	Platyhelminthes 18S	0.20***	-0.13***	-0.12**
	Annelids 18S	-0.064	0.13*	-0.059*
3rd assemblage - Exposed rock	**Protists 18S**	-0.56***	-0.055	**0.76*** **
	Arthropods 18S	-0.65***	0.12*	**0.69*** **
	Nematodes 18S	-0.48***	0.045	**0.56*** **
	Platyhelminthes 18S	-0.48***	0.20***	0.38***
	Annelids 18S	-0.18***	-0.076**	0.31***

Table S3: Correlation coefficients $\rho_{spat.}$ between the spatial distribution of bacterial assemblages and the assemblages in other taxonomic groups. p-values p were computed based on 100,000 spatial randomizations. Significant correlation coefficients are indicated by *,**,*** ($p < 0.05, p < 0.01, p < 0.001$), and correlation coefficients larger than 0.50 are indicated by bold font. Labels of assemblages are the same as in Table S2.

		OTU proportion	Taxonomic assignment
Occurrence-based assemblages	Terra firme	$8.1.10^{-4}$	Acidobacteria
		$8.1.10^{-4}$	Acidobacteriaceae (Subgroup 1) sp.
		$8.1.10^{-4}$	Acidobacteriaceae (Subgroup 1) sp.
		$8.0.10^{-4}$	Acetobacteraceae sp.
		$8.0.10^{-4}$	uncultured Holophaga sp.
			…
	Hydromorphic	$6.2.10^{-4}$	Acidothermaceae sp.
		$6.2.10^{-4}$	uncultured Holophaga sp.
		$6.1.10^{-4}$	Nitrosomonadaceae sp.
		$5.9.10^{-4}$	uncultured Holophaga sp.
		$5.9.10^{-4}$	Haliangiaceae sp.
			…
	Exposed rock	6.1.10-4	Rhizobiales Incertae Sedis sp.
		5.8.10-4	uncultured Acetobacteraceae bacterium
		5.8.10-4	uncultured Acidobacteriaceae bacterium
		5.7.10-4	Acidobacteriaceae (Subgroup 1) sp.
		5.7.10-4	Bacteria
			…
Abundance-based assemblages	Terra firme	4.0.10-2	Acidobacteria
		3.0.10-2	uncultured Nitrosococcus sp.
		2.6.10-2	uncultured Bacillaceae bacterium
		2.2.10-2	Acidothermaceae sp.
		1.9.10-2	Alcaligenaceae sp.
			…
	Hydromorphic	1.7.10-2	Alcaligenaceae sp.
		1.6.10-2	uncultured Thermosporotrichaceae bacterium
		1.6.10-2	uncultured Bacillaceae bacterium
		1.4.10-2	Acidobacteria
		1.3.10-2	Acidothermaceae sp.
			…
	Exposed Rock	3.8.10-2	Acidothermaceae sp.
		1.6.10-2	Acidothermaceae sp.
		1.5.10-2	uncultured Nitrosococcus sp.
		1.0.10-2	uncultured Steroidobacter sp.
		1.0.10-2	Xanthobacteraceae sp.
			…

Table S4: Five most abundant OTUs per bacterial assemblage (out of 20,162 bacterial OTUs)**, for occurrence and abundance data.**

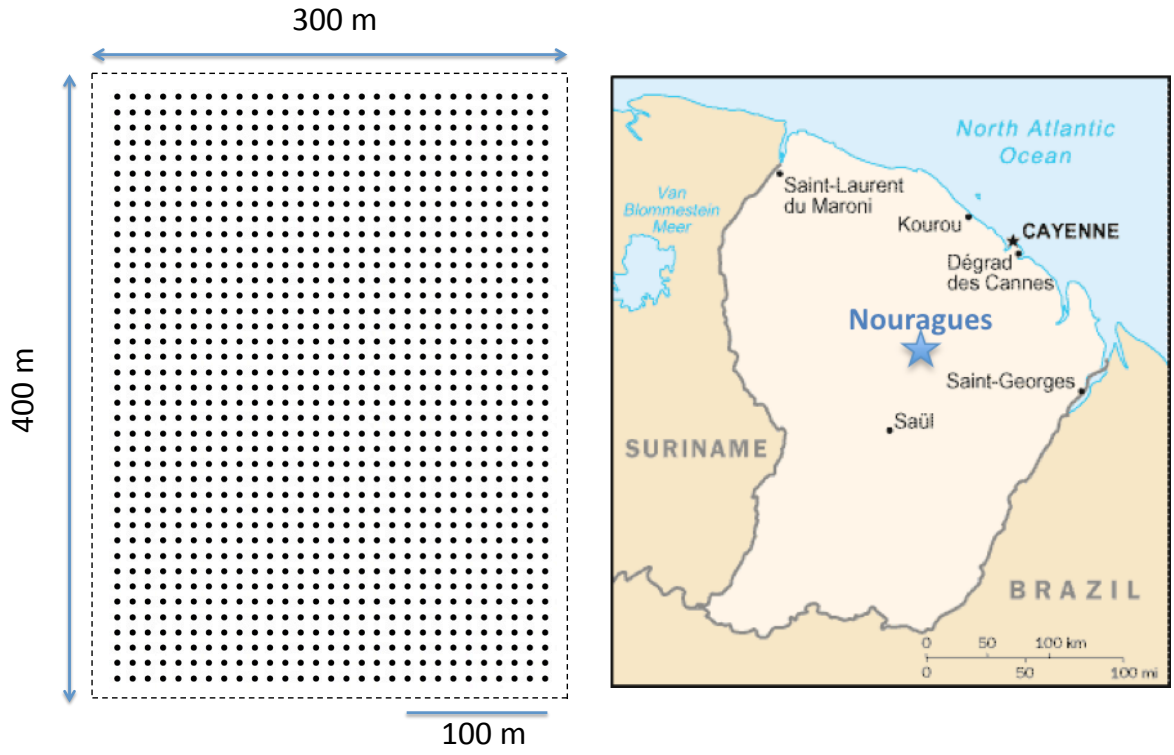

Figure S1. Soil sampling over 12 ha of tropical forest. 1,131 soil samples (one every 10 meters) were taken from the mineral soil horizon on a permanent plot of relatively homogeneous primary plateau forest at the Nouragues Ecological Research Station, French Guiana.

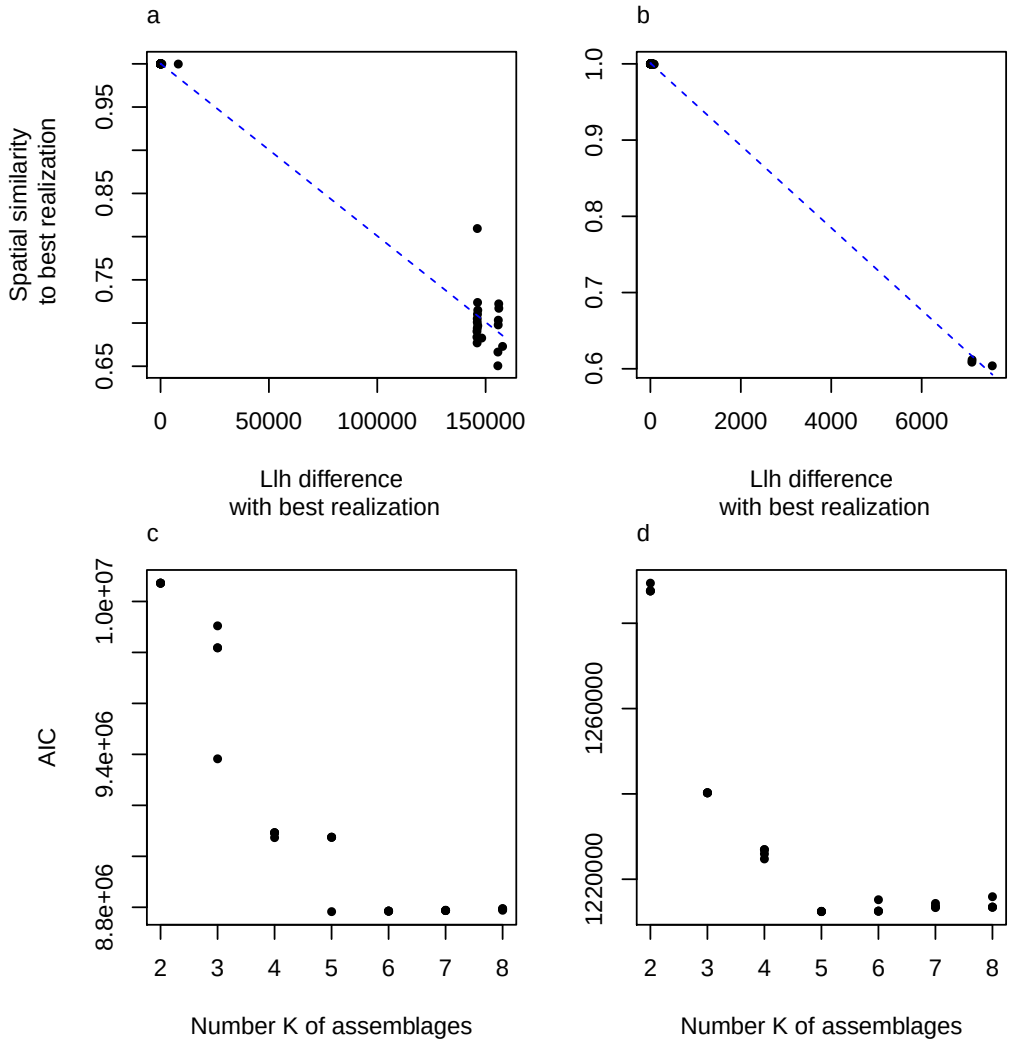

Figure S2. LDA applied to a simulated dataset with 5 assemblages, 1,000 MOTUs, 1,131 samples, and 1,000 sequence reads per sample, (a,c) for the original abundance data, and (b,d) for the occurrence data derived from the same dataset. Panels (a,b) show the comparison between the realization with highest likelihood and the 99 others using the spatial similarity $S_{\text{spat.}}$. $\langle S_{\text{spat.}} \rangle_{100} = 0.98$ for occurrence data, $\langle S_{\text{spat.}} \rangle_{100} = 0.89$ for abundance data, $I_{\text{spat.},100} = 1.0$ in both cases; cf. Fig. 2. Panels (c,d) show AIC comparison between different K values, with 3 realizations per K value.

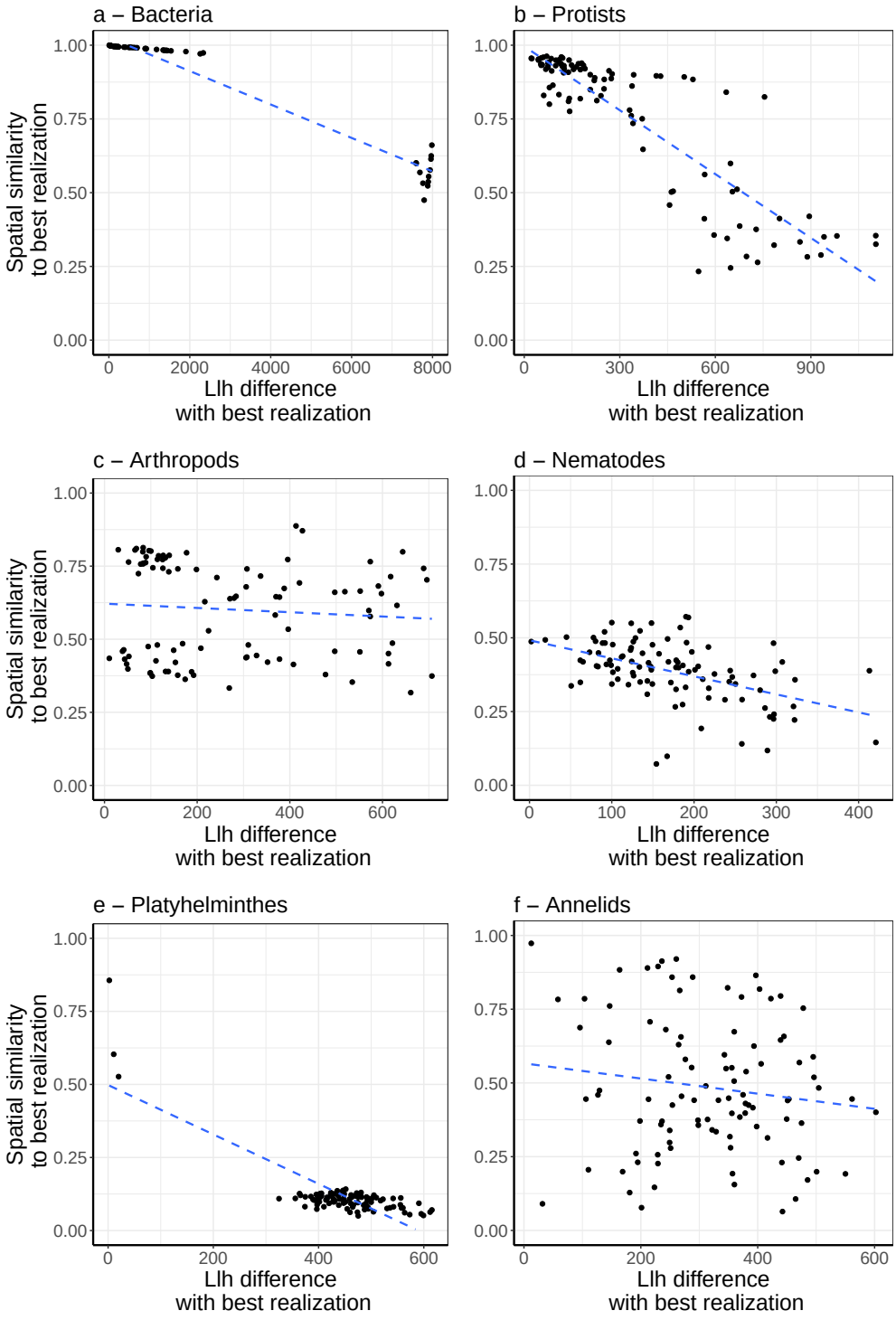

Figure S3. Stability of LDA decomposition ($K = 3$) for the different taxonomic groups. The realization with highest likelihood out of 100 is compared to the 99 others based on their spatial similarity (y-axis) and on their log-likelihood difference (x-axis), for occurrence data and for all the taxonomic groups under study. The intercept I of the linear regression (dashed blue line) shows a difference between unicellular organisms ($I = 1.0$) and metazoans ($I < 0.62$).

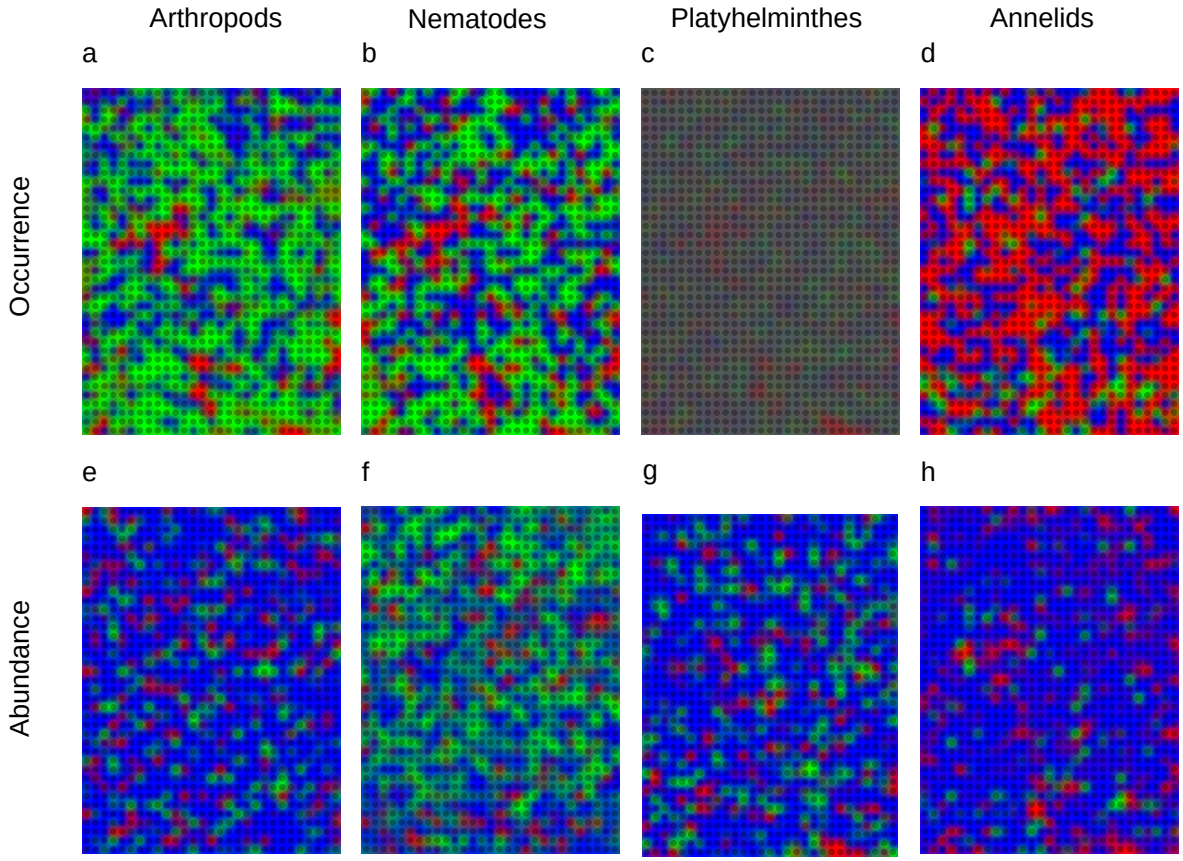

Figure S4: Spatial distribution of eukaryotic assemblages, for $K = 3$ assemblages. Spatial distribution of the assemblages obtained from independent LDA decompositions of arthropods, nematodes, flat worms (Platyhelminthes) and annelids, for occurrence (a-d) and abundance (e-h) data. As in figure 4, sampled locations are indicated by dark dots, and the assemblage mixture between samples has been interpolated using ordinary kriging. For occurrence data, an 'exposed rock' assemblage (in red) can be identified in arthropods and nematodes based on spatial correlation to the bacterial 'exposed rock' assemblage (Table S3). An 'exposed rock' assemblage may be distinguished in flat worms and annelids as well but is less conspicuous there.

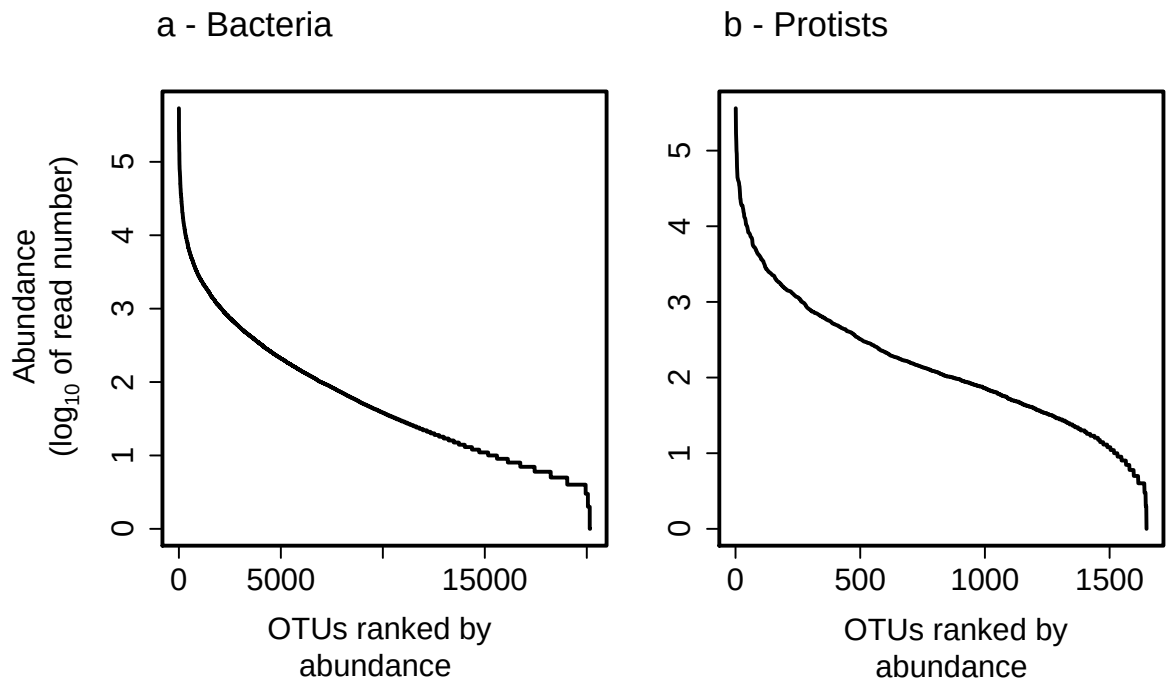

Figure S5: Ranked log-abundance distribution for bacteria and protists.

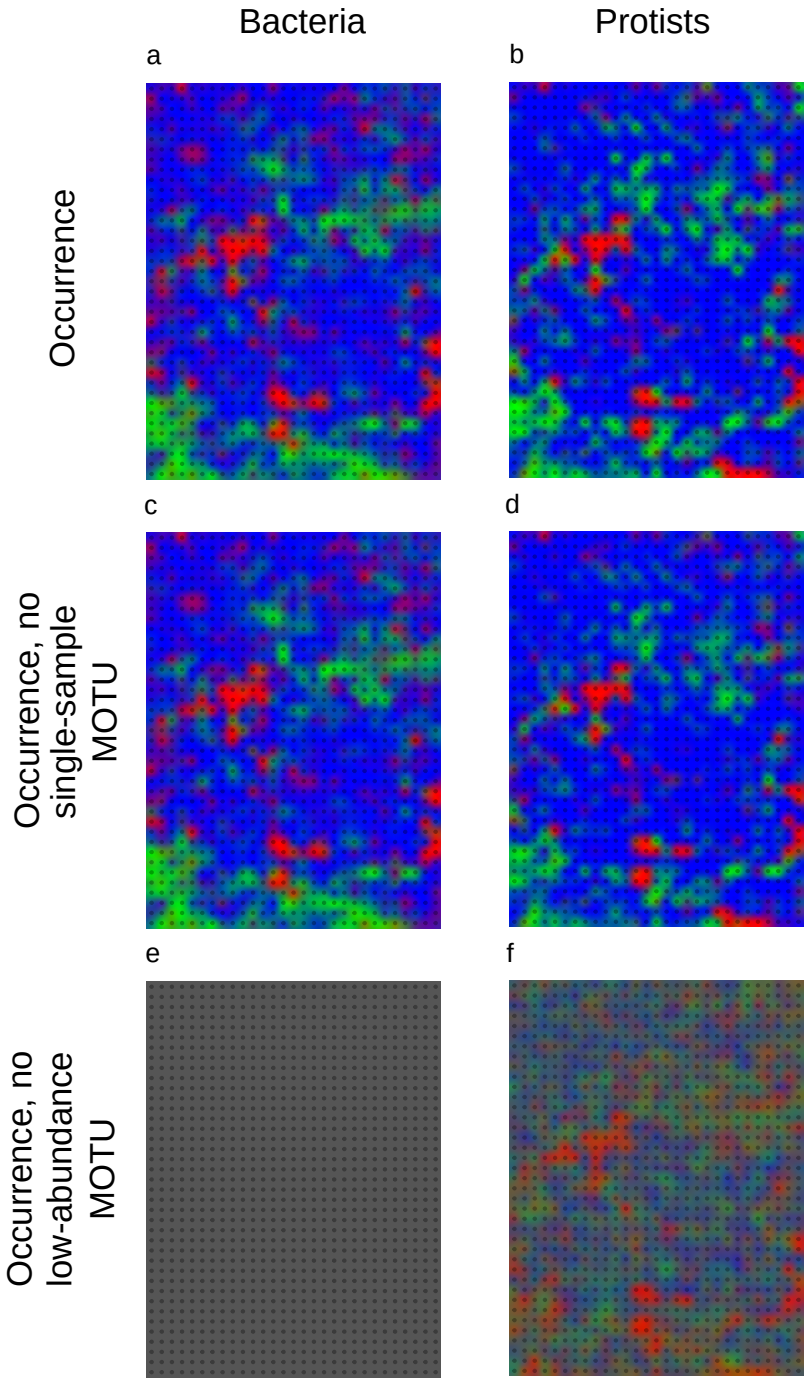

Figure S6: Effect of data pre-processing on LDA decomposition. Spatial distribution of the assemblages obtained from independent LDA decompositions of bacteria and protists for occurrence data, (a-b) without any filtering of rare OTUs, (c-d) after removing OTUs occurring only in a single sample, and (e-f) after removing OTUs with less than one read per sample on average (low-abundance OTUs). Removing single-sample OTUs brought little change to the decomposition. Removing low-abundance OTUs on the other hand yielded very degraded spatial patterns in bacteria and protists, hinting at the important role of rare MOTUs in defining the retrieved assemblages.

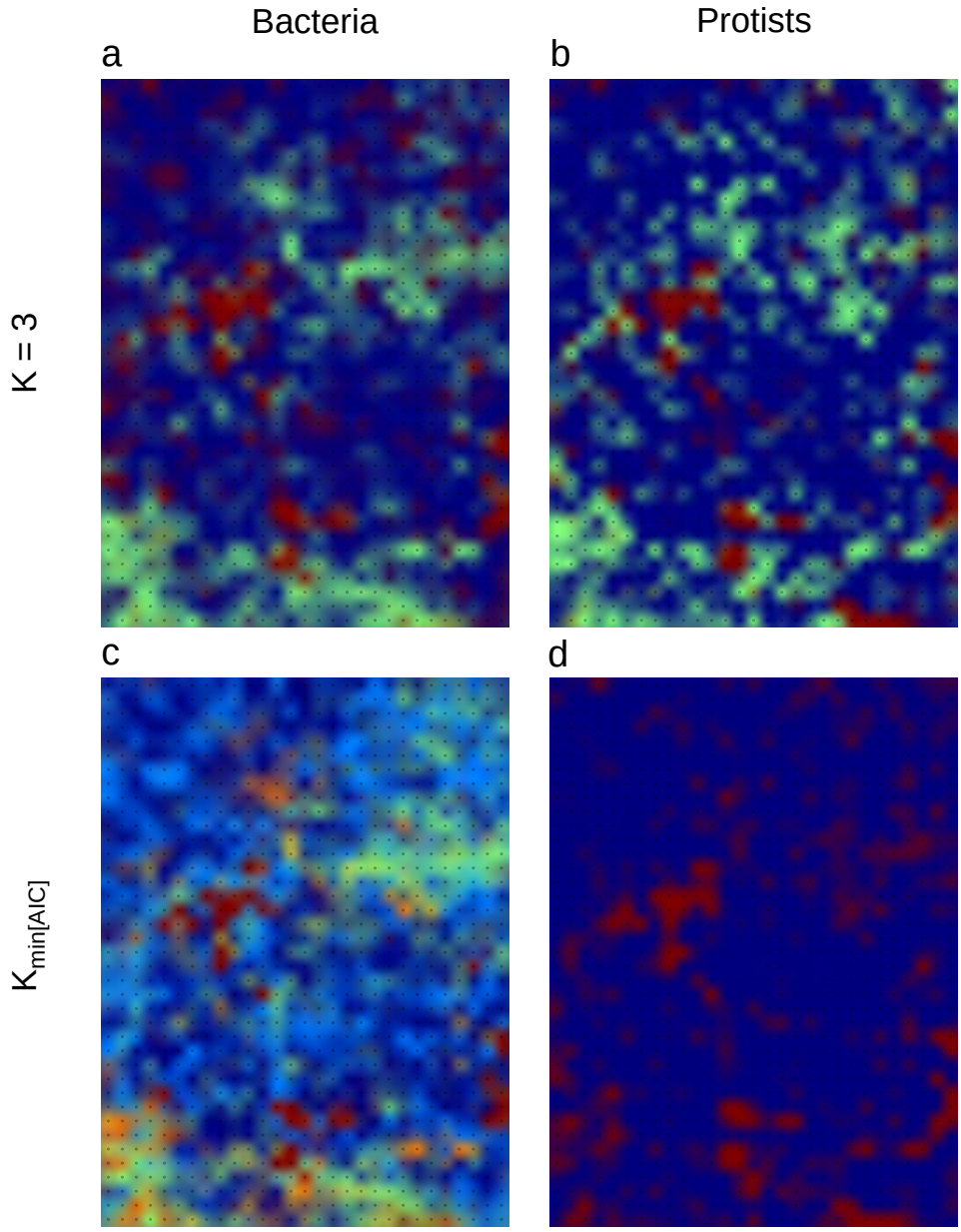

Figure S7: Spatial distribution of microorganism assemblages for occurrence data, for $K = 3$ and for $K = K_{\min[AIC]}$ assemblages (bacteria: $K_{\min[AIC]} = 5$; protists: $K_{\min[AIC]} = 2$). Decompositions for $K = K_{\min[AIC]}$ and for $K = 3$ differ primarily through splitting or merging of assemblages, without major disruption of the spatial patterns. This illustrates the robustness of LDA decomposition to the number of assemblages close to the optimum. The exposed rock assemblage (dark red) is left unchanged for K between 2 and 5 in bacteria and protists, which indicates a strong taxonomic distinctiveness.

Discussion

I. Synthesis

While data acquisition in ecology has long been dominated by low-technology approaches relying mostly on direct human observation, the field has recently witnessed a trend toward automated data acquisition. In particular, automated biodiversity measurements can now be obtained through the sequencing of environmental DNA. This method has originated in microbiology, where it is often the only means to obtain information on the organisms under study, but can now be applied with increasing ease to any type of organism. This provides ecologists with an unprecedented influx of data, which also creates new challenges.

While DNA-based data are a unique means to obtain exhaustive and standardized biodiversity measurements, it remains uncertain to what extent they will help solve the classical questions of community ecology. Indeed, these data are obtained through indirect observation of the targeted organisms, and lack in detail and accuracy compared to direct observations: in a sense, quality is traded for quantity. This entails a shift from studies rich in biological details toward the study of structure and patterns in large datasets. Moreover, the sheer amount of data produced is in itself an obstacle to the use of the classical statistical approaches of ecology. Conversely, current theoretical models in ecology are often not well suited for comparison with data.

The characteristics of environmental DNA data make them well suited to the study of integrative patterns of biodiversity, for which the quantity and exhaustiveness of available data matter more than detailed information on individual taxa. Integrative patterns have long been a key source of information for addressing one of the core questions of community ecology: what are the drivers of community assembly, and in particular, when do dispersal limitation and demographic drift supersede abiotic filtering and species interactions as the main drivers? The first and third chapters of this thesis explore the use of environmental DNA data for the study of spatially explicit biodiversity patterns, while the second chapter focuses on relative species abundances. These patterns are studied in the tropical forest of French Guiana, a ‘hyperdiverse’ and

poorly known ecosystem; two characteristics that make automated data collection most needed.

The first chapter shows how environmental DNA data can be used to investigate the drivers of beta diversity in a spatially explicit context, as it has been done previously for classical data such as tree censuses in monitored forest plots. On a spatial scale ranging from 40 m to 140 km between sampling points, a decay of taxonomic similarity with distance is observed in most groups, i.e. plants, fungi, arthropods, insects, annelids, bacteria, and protists, but not in nematodes and flat worms. Clear differences can be observed between domains of life regarding the relative influence of geographic distance and abiotic conditions on beta diversity: the data hint at a predominant effect of dispersal limitation in plants and annelids, a predominant effect of abiotic filtering in bacteria and protists, and a mixture of both in fungi, arthropods and insects. The beta diversity of fungi and soil insects appears to be especially high. These findings are in agreement with expectations and previous empirical results for plants and unicellular organisms (Condit *et al.*, 2002; Soininen *et al.*, 2007; Ramirez *et al.*, 2014), but bring some novel insight for annelids, fungi and insects. In addition, the inclusion of a few forest plots subject to past logging activities indicates that even after two decades at least, an effect can be detected on plant and annelid composition, as well as on fungi to a lesser extent, whereas it is not the case for other groups. Thus, large-scale patterns of biodiversity can now be readily measured and compared across a tropical forest's whole range of taxa using environmental DNA.

The second chapter focuses on relative species abundances, a pattern that has been extensively used to test the predictions of theoretical models of community assembly, especially since Hubbell's work on the neutral theory of biodiversity (Hubbell, 2001). A major obstacle in exploiting species abundance patterns generated using environmental DNA is that abundance information is unreliable, because it is noisy and difficult to interpret. However, simulations show that even if abundance measurements are unreliable for individual taxa, valuable information can still be retrieved from the species abundance distribution as a whole, as long as the noise is not too strong. In particular, the parameters that characterize diversity and connectivity in a neutral community may still be reliably estimated. Thank to the sampling-invariance property of neutral models, sequencing reads may be used as discrete abundance units in place of

individuals as long as the DNA originates from a number of individuals larger than the number of reads. While it is usually the case for microorganisms, this condition may not be verified for larger organisms. Lastly, great care should be taken in clustering spurious OTUs generated during PCR amplification and DNA sequencing, since they strongly bias neutral parameter estimates.

When the spatial distribution of species is shaped at least partly by niche processes or by limited dispersal, the structure of spatially distributed environmental DNA data should be marked by these processes. However, this signal may be faint and complex. Moreover, it is usually obscured by a large number of rare species and an uneven sampling effort across samples. The third chapter shows how a categorical mixture model similar to some of the models used in microbiology, population genetics or text document modelling, Latent Dirichlet Allocation, can be used to retrieve spatial patterns in a regularly-sampled 12-ha forest plot. Unlike the classical pattern-detection tools of community ecology, such as simple ordination and clustering algorithms, this model is designed to accommodate discrete abundance data in a large number of unevenly sized samples, and performs well on large and sparse community matrices. Even though the fitted model parameters may depend on the initialization of the inference algorithm, this uncertainty can be quantified by measuring the similarity between the outputs of different runs. The stability of the output across initial conditions may even be used as an empirical measure of how strong the spatial structure is.

In the 12-ha forest plot, the strongest structure is detected for bacteria and protists. Moreover, the spatial patterns of these two groups are very similar, and match the topography of the forest plot. This is in agreement with the findings of the first chapter, since abiotic filtering was found there to strongly influence the beta diversity of these groups. In contrast, spatial structure in arthropods and annelids is weak, which indicates that the spatial scale and the level of environmental heterogeneity in a 12-ha plot are insufficient to detect the processes that were found to act on these groups at larger spatial scales.

Overall, we conclude that environmental DNA data can offer a uniquely comprehensive, if somewhat crude, perspective on community structure in a complex

and species-rich ecosystem. In addition to the classical tools of community ecology, model-based statistical methods can be borrowed from fields more accustomed to large and complex datasets, and put to good use to take full advantage of these data. The development of ecology into a data-rich field should foster the development of theoretical models that can be compared to data using rigorous statistical approaches, following the example of Hubbell's neutral model and its subsequent theoretical developments (Etienne, 2005; Harris *et al.*, 2015). Building on generative models stemming from machine learning, such as Latent Dirichlet Allocation, is one possible avenue for the development of such models, as discussed in the following.

II. Perspectives

1. Aquatic communities

This thesis aimed at exploring general approaches for the analysis and interpretation of large biodiversity datasets, with the underlying goal of understanding community assembly processes from biodiversity patterns. Nevertheless, it mostly focuses on community assembly in land ecosystems, especially as studied through the amplification and sequencing of DNA extracted from soil samples. One may follow a similar approach for studying communities of aquatic organisms by extracting DNA from water samples. Experimentally, the method consists in filtering water through a mesh so as to collect small living organisms as well as fragments or sloughed material from larger organisms. In particular, this approach allows for the study of planktonic microorganisms (i.e., suspended in the water column and passively transported by water movements), the knowledge of which is so far very fragmented, despite them forming the basis of the ocean's food web and being responsible for the production of half of the atmospheric dioxygen (Field *et al.*, 1998).

The Tara project is an unprecedented and on-going effort to sample marine planktonic communities in various locations spread across the world's oceans (de Vargas *et al.*, 2015). Sampling was conducted chiefly in the open ocean from 2009 to 2012, with more recent campaigns focusing on more specific habitats. Samples were collected at different depths, and using different mesh sizes so as to assign the sampled organisms to different size ranges. The Latent Dirichlet Allocation approach of the third chapter is currently being applied to this dataset, so as to understand the biogeography and community structure of planktonic eukaryotes across the world's ocean.

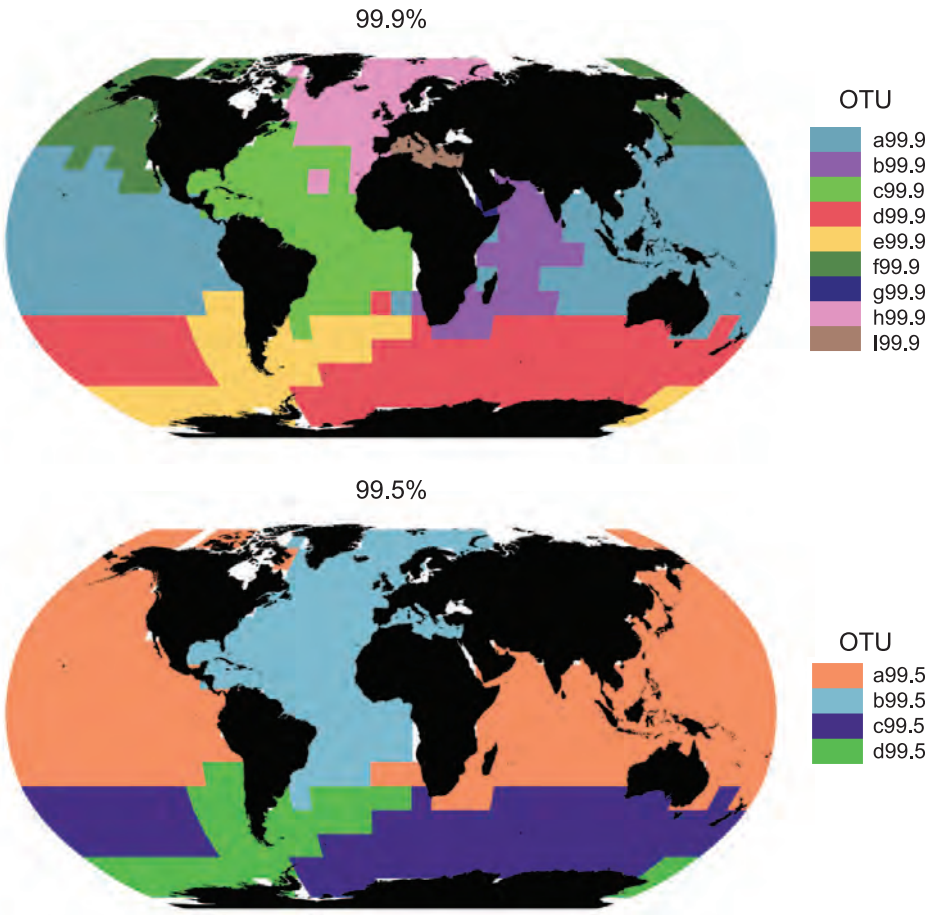

Figure 1: Biogeographic patterns in oceanic plankton predicted by a neutral agent-based model. Transport by oceanic current was simulated during 1,400 years with constant mutation rate starting from a single genome. Biogeographic regions are distinguished based on their dominant OTUs, defining OTUs at either (**top**) 99.9% similarity or (**bottom**) 99.5% similarity. Adapted from Hellweger *et al.* (2014).

However, it is unclear what a suitable neutral model would be for planktonic communities. Indeed, unlike land organisms, planktonic organisms do not actively disperse. Instead, the local community is transported over time along oceanic currents, and slowly mixes with surrounding communities along the way. Hubbell's model of a local community under constant immigration flow could be regarded as a suitable model for a planktonic community followed through time along an oceanic current ('Lagrangian' perspective). However, whether several simultaneously sampled communities can be considered as independent and undergoing immigration from the same metacommunity depends on their positions relative to oceanic currents.

Simulations of the transport of plankton by currents between stations could help measure their level of connectivity (see Fig. 1; Follows *et al.*, 2007; Ward *et al.*, 2012; Hellweger *et al.*, 2014), and serve as the basis for inference-oriented modelling efforts.

2. Topic modelling of biodiversity data

As discussed in the section III.5 of the Introduction, Latent Dirichlet Allocation is a very versatile method, that has been employed in a variety of contexts far beyond its original intended use as a 'natural language processing' method. It could become a routine tool for the analysis of environmental DNA data, as the very similar Structure software has become in population genetics. While the third chapter focuses on the analysis of spatially distributed samples, LDA could prove equally useful for the analysis of time series, or when both a spatial and a temporal dimensions are present, as in Valle *et al.* (2014). It could also be used to analyse samples that are neither spatially nor temporally distributed. This is for instance often the case of human microbiome data, which are currently collected in large quantities, and the interpretation of which is an active domain of research in medical sciences (Huttenhower *et al.*, 2012). Another potential application is the analysis of the bacterial communities found in sewage plants, the understanding of which is of critical importance for the optimization of wastewater treatment (Ofiteru *et al.*, 2010).

The use of generative mixture models is not new in microbiology: these methods have been first introduced to the field with the works of Knights *et al.* (2011) and Holmes *et al.* (2012). However, probably because ecology and microbiology are still relatively separate scientific fields, and because the use of environmental DNA data is more recent in ecology, generative mixture models have been little used so far in ecology, except for the effort of Valle *et al.* (2014) on classical tree census data. Furthermore, focus in microbiology appears to have been mostly on models without admixture (i.e., where samples belong to a single assemblage, cf. Introduction), unlike topic models. While LDA is one of the simplest topic models (along with the earlier Probabilistic Latent Semantic Analysis model, or PLSA; Hofmann, 2001), many extensions have been developed for the analysis of text documents since its original

introduction. The adaptation of these methods to bioinformatics, e.g. for the classification of DNA sequences or the identification of protein function, has been extensively explored (Liu *et al.*, 2016). Ecology, and microbiology, would benefit from a similar effort oriented toward biodiversity data. In the following, I review a few possible examples.

Figure 2: Terrestrial biogeographic units of the world inferred from the distribution of 21,037 species of amphibians, birds and mammals. Inference was performed using UPGMA hierarchical clustering on phylogenetic dissimilarity. Thick lines denote main biogeographic boundaries (separating 'realms') and dotted lines denote minor ones (separating 'regions'). Adapted from Holt *et al.* (2013).

First, the approach presented in the third chapter can be used irrespective of the spatial scale at which the data are collected, and may for instance be applied to the definition of biogeographic units. Aside from the sequencing of environmental DNA, the development of DNA sequencing methods now allows for efficiently and accurately assigning to a taxon any collected biological material, once a suitable reference database has been established. Thus, a better use can be made of the large number of specimens either collected in the field or stored in museum collections, and the resulting data may be used to study biogeographic patterns in a data-driven way. In recent years, alternatives to the classical hierarchical clustering approach (followed for instance in Holt *et al.*, 2013; see Fig. 2) have been sought to address this problem (Vilhena & Antonelli, 2015; Bloomfield *et al.*, 2017). The Appendix illustrates the potential of LDA in

this respect, by applying it to a large dataset of Amazonian frogs assembled with the help of DNA identification. LDA proved in this case an efficient way to infer the optimal number of biogeographic units, assess the strength of the underlying signal, and distinguish between sharp and diffuse boundaries between biogeographic units.

Nevertheless, unlike distance-based clustering, LDA does not allow for taking phylogenetic information into account, and improving on this aspect could be a useful research avenue. This problem could be for instance approached through the Generalized Polya urn LDA model (Mimno *et al.*, 2011). Furthermore, despite its shortcomings, distance-based hierarchical clustering is appreciated by ecologists because it provides additional hindsight on the relationship between the different samples, and because it offers the possibility of choosing the number of clusters based on the hierarchical tree. The hierarchical LDA model (or hLDA; Griffiths *et al.*, 2004), which describes a hierarchy of nested topics, could be appealing in this respect.

Second, the study of interactions between taxa constitutes a central interest of community ecology. Large environmental DNA datasets provide indirect information on potential interactions between taxa through the co-occurrence of OTUs and the covariance of their abundances (see Fig. 3; Faust & Raes, 2012). LDA assemblages are inferred based on this information, and thus reflect the presence of potential interactions within each assemblage. Nevertheless, the application of LDA decomposition separately to different taxonomic groups, as done in the third chapter, does not provide any information on the possible interactions between these groups. Conversely, when LDA is applied to the whole dataset, it is not possible to explicitly distinguish between subgroups of preferentially interacting taxa, such as plants and fungi for instance. This shortcoming could be addressed by using the 'author-topic model', an extension of LDA aiming at accounting for sample 'metadata', such as authors in a text document (Rosen-Zvi *et al.*, 2004, 2010). This model is identical to LDA except that each document (or sample) is not directly characterized by a mixture of topics (or assemblages), but by its authors, to each of which is assigned a mixture of topics. In practice, authors may be any discrete labels, and could for instance correspond to the one or few tree species surrounding each soil sample if one is interested in tree-fungi interactions. The method would thus yield a mixture of fungi assemblages for each tree species, and indirectly a mixture of assemblages for each sample based on the tree

species surrounding it. A simpler version of this model might also be considered where a single assemblage characterizes each tree species.

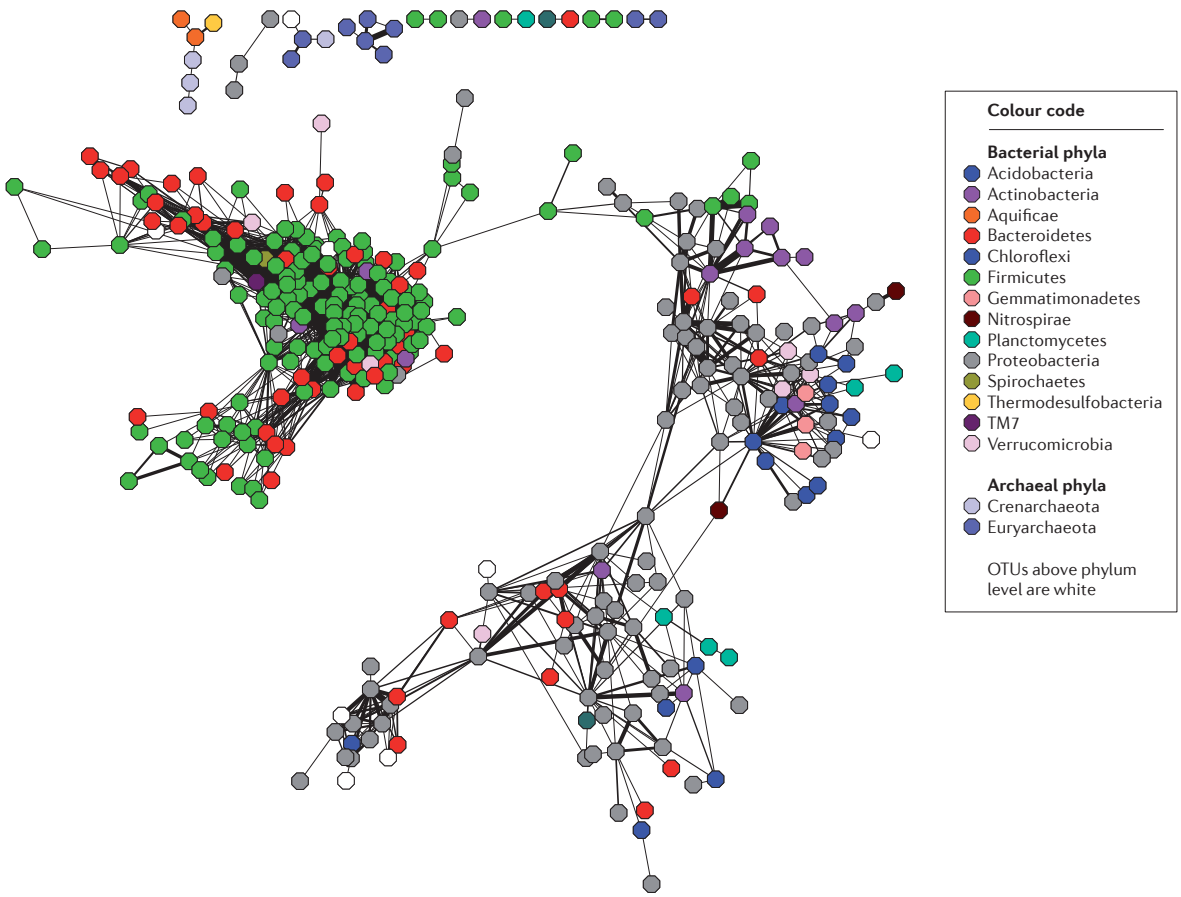

Figure 3: Occurrence-based inference of interactions between prokaryotic OTUs from a global data set (Chaffron *et al.*, 2010). Each node represents an OTU, and edges between nodes represent significant associations based on co-occurrence. Edge thickness increases with significance. Adapted from Faust & Raes (2012).

More generally, ecological studies often do not limit themselves to exploring the structure of a single type of data, but attempt at uncovering statistical relationships between different types of data, such as taxonomic and environmental data. While the author-topic model only allows for adding discrete labels to each sample, other models such as Dirichlet-multinomial Regression (Mimno & McCallum, 2012) can also accommodate continuous attributes, and could be used to account for environmental measurements in the model. The goodness-of-fit of the model without environmental

data could then be compared to that of the model accounting for these data, ideally using AIC. This would entail a slightly different use of topic modelling than that of the third chapter: namely, shifting the focus from the exploration of data structure toward hypothesis testing. However, both approaches have their own merits.

Finally, the application of topic modelling to ecology needs not be limited to taxa abundance and occurrence data. It could for instance be extended to exon sequencing data describing functional types, or to RNA sequencing data characterizing gene expression. Moreover, aside from the major technological revolution that is high-throughput DNA sequencing, other promising technologies are currently being adapted to automated data collection in ecology, notably Lidar and hyperspectral imaging. Topic modelling has been successfully used to retrieve patterns from images (Luo *et al.*, 2015), and could possibly also find application in the analysis of remote-sensing ecological data.

3. Statistical versus mechanistic modelling

Topic modelling is but one of many competing branches of machine learning that are currently actively developed to exploit the ever-increasing amount of data produced by current technologies (Bishop, 2006). Over the recent years, some branches of machine learning have become particularly prominent, especially multi-layered neural networks under the name of 'deep learning' (LeCun *et al.*, 2015). Such methods are indeed efficient at detecting structure in large datasets, and have been recently applied to bioinformatics problems such as DNA sequence classification (Rizzo *et al.*, 2016). However, these methods are not based on an easily interpretable model. As such, they can only be fruitfully applied to supervised learning tasks, i.e. to situations where correct and incorrect results can be told apart a priori, which are more typical of engineering than basic science.

In contrast, topic models have a mathematical structure that is similar to the multivariate formulation of neutral models, as discussed in Harris *et al.* (2015) and in the third part of the Introduction. This parallel could be exploited to build mixed models

combining the advantages of both types of models. For instance, a local community, such as an island, may receive immigrating individuals from different source communities that have distinct (known) taxonomic compositions. By assuming a neutral dynamics in the local community, and modelling the origin of immigrating individuals by a topic model, one could possibly infer from the local taxonomic composition the relative contribution of the different source communities. Conversely, starting from a topic model as in the third chapter, one could assume that a neutral dynamics takes place within each assemblage.

While topic and neutral models may seem to be of different nature, the distinction between statistical and mechanistic models is more tenuous than may appear at first glance. The first topic modelling papers mentioned mechanistic arguments to justify their models, arguing than they mirrored the way humans write text documents, and some subsequent developments try to better account for the structure of natural language (Wallach, 2006). When applied to ecological data, the assumption that local communities are a mixture of several assemblages of co-occurring taxa constitutes a genuine biological hypothesis. Conversely, the realism of the hypotheses in Hubbell's neutral model has been much debated (Rosindell *et al.*, 2012), and one might argue that its most valuable hindsight is on the nature of the species abundance distribution pattern itself: namely, that most empirical species abundance distributions can be approximately decomposed into orthogonal diversity and connectivity components, irrespective of their exact mechanistic interpretation (Jabot *et al.*, 2008).

While a very flexible model is undesirable when one aims at testing modelling hypotheses on data, it becomes an advantage when one aims at characterizing the system at hand through a limited number of relevant parameters. This is often a more realistic prospect when faced with large datasets resulting from automated data collection. However, as illustrated by the case of Hubbell's neutral model, relevant parameters cannot be determined without an understanding of the basic processes at play. Moreover, characterizing a system is of little use if this does not entail the possibility of predictions and generalization. A right balance is thus to find between flexibility and falsifiability in building models for the analysis of large datasets, and the

shift toward inference-oriented models should not preclude building them on first principles (Marquet *et al.*, 2014).

References

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer. Michael Jordan, Jon Kleinberg, Bernhard Schölkopf.

Bloomfield, N.J., Knerr, N. & Encinas-Viso, F. (2017) A comparison of network and clustering methods to detect biogeographical regions. *Ecography*.

Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. (2010) A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Research*, **20**, 947–959.

Condit, R., Pitman, N., Leigh, E.G., Chave, J., Terborgh, J., Foster, R.B., Nunez, P., Aguilar, S., Valencia, R., Villa, G., Muller-Landau, H.C., Losos, E. & Hubbell, S.P. (2002) Beta-diversity in tropical forest trees. *Science*, **295**, 666–669.

Etienne, R.S. (2005) A new sampling formula for neutral biodiversity. *Ecology Letters*, **8**, 253–260.

Faust, K. & Raes, J. (2012) Microbial interactions: from networks to models. *Nature Reviews Microbiology*, **10**, 538–550.

Field, C.B., Behrenfeld, M.J., Randerson, J.T. & Falkowski, P. (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.

Follows, M.J., Dutkiewicz, S., Grant, S. & Chisholm, S.W. (2007) Emergent Biogeography of Microbial Communities in a Model Ocean. *Science*, **315**, 1843.

Griffiths, T.L., Jordan, M.I., Tenenbaum, J.B. & Blei, D.M. (2004) *Hierarchical topic models and the nested chinese restaurant process*. *Advances in neural information processing systems*, pp. 17–24.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2015) Linking statistical and ecological theory: Hubbell's Unified Neutral Theory of Biodiversity as a Hierarchical Dirichlet Process. *Proc. IEEE*, **PP**, 1–14.

Hellweger, F.L., van Sebille, E. & Fredrick, N.D. (2014) Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. *Science*.

Hofmann, T. (2001) Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, **42**, 177–196.

Holmes, I., Harris, K. & Quince, C. (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics. *Plos One*, **7**.

Holt, B., Lessard, J.P., Borregaard, M.K., Fritz, S.A., Araujo, M.B., Dimitrov, D., Fabre, P.H., Graham, C.H., Graves, G.R., Jonsson, K.A., Nogues-Bravo, D., Wang, Z.H., Whittaker, R.J., Fjeldsa, J. & Rahbek, C. (2013) An Update of Wallace's Zoogeographic Regions of the World. *Science*, **339**, 74–78.

Hubbell, S.P. (2001) *The unified neutral theory of biodiversity and biogeography (MPB-32)*, Princeton University Press.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A.,

Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.-M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Dwayne Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren, W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu, Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K. & White, O. (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Jabot, F., Etienne, R.S. & Chave, J. (2008) Reconciling neutral community models and environmental filtering: theory and an empirical test. *Oikos*, **117**, 1308–1320.

Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. & Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, **8**, 761-U107.

LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, **521**, 436–444.

Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. (2016) An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, **5**.

Luo, W., Stenger, B., Zhao, X. & Kim, T.-K. (2015) *Automatic Topic Discovery for Multi-Object Tracking. AAAI*, pp. 3820–3826.

Marquet, P.A., Allen, A.P., Brown, J.H., Dunne, J.A., Enquist, B.J., Gillooly, J.F., Gowaty, P.A., Green, J.L., Harte, J., Hubbell, S.P., O'Dwyer, J., Okie, J.G., Ostling, A., Ritchie, M., Storch, D. & West, G.B. (2014) On Theory in Ecology. *Bioscience*, **64**, 701–710.

Mimno, D. & McCallum, A. (2012) Topic models conditioned on arbitrary features with dirichlet-multinomial regression. *arXiv preprint arXiv:1206.3278*.

Mimno, D., Wallach, H.M., Talley, E., Leenders, M. & McCallum, A. (2011) Optimizing Semantic Coherence in Topic Models. *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*.

Ofiteru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C.A. & Sloan, W.T. (2010) Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15345–15350.

Ramirez, K.S., Leff, J.W., Barberan, A., Bates, S.T., Betley, J., Crowther, T.W., Kelly, E.F., Oldfield, E.E., Shaw, E.A., Steenbock, C., Bradford, M.A., Wall, D.H. & Fierer, N. (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B-Biological Sciences*, **281**, 9.

Rizzo, R., Fiannaca, A., La Rosa, M. & Urso, A. (2016) *A Deep Learning Approach to DNA Sequence Classification*. *Computational Intelligence Methods for Bioinformatics and Biostatistics: 12th International Meeting, CIBB 2015, Naples, Italy, September 10-12, 2015, Revised Selected Papers* (ed. by C. Angelini), P.M. Rancoita), and S. Rovetta), pp. 129–140. Springer International Publishing, Cham.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. & Steyvers, M. (2010) Learning Author-Topic Models from Text Corpora. *Acm Transactions on Information Systems*, **28**.

Rosen-Zvi, M., Gri_ffiths, T., Steyvers, M. & Smyth, P. (2004) *The Author-Topic Model for Authors and Documents*.

Rosindell, J., Hubbell, S.P., He, F., Harmon, L.J. & Etienne, R.S. (2012) The case for ecological neutral theory. *Trends in Ecology & Evolution*, **27**, 203–208.

Soininen, J., McDonald, R. & Hillebrand, H. (2007) The distance decay of similarity in ecological communities. *Ecography*, **30**, 3–12.

Valle, D., Baiser, B., Woodall, C.W. & Chazdon, R. (2014) Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecology Letters*, **17**, 1591–1601.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Lukes, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E. & Tara Oceans, C. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**.

Vilhena, D.A. & Antonelli, A. (2015) A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, **6**.

Wallach, H.M. (2006) *Topic modeling: beyond bag-of-words*. *Proceedings of the 23rd international conference on Machine learning*, pp. 977–984. ACM.

Ward, B.A., Dutkiewicz, S., Jahn, O. & Follows, M.J. (2012) A size-structured food-web model for the global ocean. *Limnology and Oceanography*, **57**, 1877–1891.

Appendix

Large-scale DNA barcoding of Amazonian anurans leads to a new definition of biogeographical subregions in the Guiana Shield and reveals a vast underestimation of diversity and local endemism

Jean-Pierre Vacher[a], Guilhem Sommeria-Klein[a], Francesco Ficetola[b], Miguel Trefaut Rodrigues[c], Philippe J.R. Kok[d], Brice P. Noonan[e], Andrew Snyder[e], Rawien Jairam[f], Paul Ouboter[f], Jerriane Oliveira Gomes[g], Teresa C.S. Avila-Pires[g], Jucivaldo Dias Lima[h], Raffael Ernst[i], Michel Blanc [j], Maël Dewynter[k], Tim J. Colson[l], Sergio M. de Souza[m], Pedro Nunes[n], Augustin Camacho[o], Mauro Teixeira[p], Renato Recoder[q], José Cassimiro[r], Quentin Martinez[s], Christian Marty[t], Philippe Gaucher[u], Christophe Thébaud[a], Antoine Fouquet[a,u].

[a] Laboratoire EDB, UMR5174, CNRS-UPS-IRD, Toulouse, France;
[b] Laboratoire d'Écologie Alpine (LECA), UMR5553, Grenoble, France;
[c] Universidade de São Paulo, Instituto de Biociências, Departamento de Zoologia, Caixa Postal 11.461, CEP 05508-090, São Paulo, SP, Brazil;
[d] Amphibian Evolution Lab, Biology Department,Vrije Universiteit Brussel, 2 Pleinlaan, 1050 Brussels, Belgium;
[e] Department of Biology, 507 Shoemaker Hall, University, MS 38677, USA;
[f] National Zoological Collection Suriname (NZCS), Anton de Kom University of Suriname, Paramaribo, Suriname;
[g] Museu Paraense Emílio Goeldi, Laboratório de Herpetologia/CZO, Av. Perimetral, 1901; Terra Firme, Belém, Pará, Brazil;
[h] Centro de Pesquisas Zoobotânicas e Geologicas (CPZG), Instituto de Pesquisas Cientificas e Tecnológicas do Estado do Amapá (IEPA), Macapá, AP, Brazil;
[I] Museum of Zoology, Senckenberg Natural History Collections Dresden, Königsbrücker Landst. 159, 01109 Dresden, Germany;
[j] Pointe Maripa, RN2/PK35, 93711, Roura, French Guiana;
[k] Biotope, Agence Amazonie-Caraïbes, 30 Domaine de Montabo, Lotissement Ribal, 97300 Cayenne, French Guiana; [l]Adresse
[s] 8 rue Mareschal, 30900 Nîmes, France;
[t] impasse Jean Galot, Montjoly, French Guiana;
[u] Laboratoire Écologie, évolution, interactions des systèmes amazoniens (USR 3456 LEEISA), Université de Guyane, CNRS Guyane, Cayenne, French Guiana.

Introduction

Amazonia encompasses about 40% of the world's tropical forests (Sioli, 1984; Hubbell *et al.*, 2008; Hoorn & Wesselingh, 2010), and many taxonomic groups reach their highest species richness in this region (Antonelli & Sanmartín, 2011; Jenkins *et al.*, 2013). The processes that have given rise to this exceptionally high diversity have long intrigued biologists (Wallace, 1852; Bates, 1863). Amazonia gives an appearance of homogeneity, because it is a vast and seemingly uniform extent of forest that is faunistically very distinct from other Neotropical regions (Dinerstein *et al.*, 1995; Olson *et al.*, 2001; Antonelli & Sanmartín, 2011; Vilhena & Antonelli, 2015). However, this is misleading: temperatures and rainfall vary widely across Amazonia (Mayle & Power, 2008), and so do vegetation types (Anderson, 2012; Hughes *et al.*, 2013). Moreover, Amazonia had a tumultuous climatological and geological past, mainly caused by the Andean uplift and the setting-up of the Rio Amazonas watershed during the late Tertiary (Hoorn *et al.*, 2010).

The distribution of species within Amazonia is known to relate to this large-scale environmental heterogeneity. The observed congruence between the geographic distribution of birds and primates on the one hand and the major interfluves on the other hand (Wallace, 1852; Haffer, 1974) led to the definition of biogeographic subregions (BSRs), coined as "Amazonian areas of endemism" (Wallace, 1852; Haffer, 1974; Cracraft, 1985). However, there is still little consensus on how to best delimit and name BSRs, with many terms being used interchangeably (Vilhena & Antonelli, 2015). In fact, the very existence and boundaries of different BSRs across Amazonia and the relative degree of endemism within them have simply never been analysed using modern analytic tools (e.g., clustering) and large species assemblages having unambiguous distribution data (Nelson *et al.*, 1990; Morrone, 2005; Naka *et al.*, 2012). Moreover, current knowledge on the delimitation of Amazonian BSRs is mainly based on birds, the best-known taxonomic group, as well as primates and plants displaying limited distributions in Amazonia. The explanatory power of the Amazonian BSR as currently defined remains questionable until their boundaries are proven to match

across multiple taxonomic groups. However, this seems unlikely because these groups have overall high dispersal abilities, and their distribution patterns may be poor predictors for less vagile taxa (Claramunt *et al.*, 2011; Pigot & Tobias, 2015; Zizka *et al.*, 2016).

Because small terrestrial vertebrates such as anurans have more limited dispersal abilities and possibly a greater sensitivity to environmental variation, they are better suited to the delineation of relevant bioregions (Zeisset & Beebee, 2008). Anuran assemblages may display different or finer geographic patterns than those previously described at the continental (Vilhena & Antonelli, 2015) or the regional scale (Vasconcelos et al., 2014). For instance, one of the rare studies unambiguously delimiting BSRs in Amazonia found well-delimited bioregions in the Guiana Shield based on the distribution of bird species, including a large homogeneous region spanning the eastern part of the Guiana Shield (Naka *et al.*, 2012). Yet, studies on anuran amphibians suggest a finer biogeographic structure in the Eastern Guiana Shield, where divergent lineages of frogs exhibit concordant distribution limits (Fouquet *et al.*, 2012d, 2013, 2016). In this paper, we aim at delimiting Amazonian bioregions in a data-driven way based on a newly collected dataset of molecular anuran diversity, with a particular focus on the Eastern Guiana Shield.

Revealing the basic geographical structure of species diversity in Amazonia is not only of crucial importance for conservation (Da Silva et al., 2005), it is also an important prerequisite for the study of the processes that gave rise to present-day diversity patterns. Identifying BSRs in Amazonia may help identify the physical barriers relevant to speciation, define the contact zones between closely related parapatric taxa, and capture the effects of dispersal limitation in the structure of Amazonian communities (e.g., Moura et al., 2016). Many hypotheses have been proposed to explain heterogeneities in species distribution across Amazonia, including landscape change induced by late Tertiary climate fluctuations (Haffer 1969), the uplift of the Andes, and continuous dispersal across large rivers (Hayes & Sewlal, 2004; Antonelli *et al.*, 2010; Hoorn *et al.*, 2010), or past environmental gradients (Colinvaux *et al.*, 2000). These different hypotheses have been verified for some taxonomic groups at different spatial and temporal scales (Hall & Harvey, 2002), but there is still no consensus about the main drivers of diversification within Amazonia.

Two major challenges to our understanding of the basic structure of Amazonian biodiversity are the scarcity of occurrence data and the imprecision of species delineation (Wallacean and Linnean shortfalls). These shortfalls are particularly obvious in small terrestrial vertebrates such as anurans (Ficetola *et al.*, 2014). Almost all anuran taxa with large ranges in Amazonia exhibit deep divergences when analysed with genetic tools, suggesting that they comprise several species, each with a restricted distribution (Fouquet *et al.*, 2007a; Funk *et al.*, 2012; Gehara *et al.*, 2014; Fouquet *et al.*, 2015b; Ferrão *et al.*, 2016; Fouquet *et al.*, 2016). These studies typically imply that the actual species richness in these groups is more than twice that estimated from morphology only. Therefore, ranges of Amazonian amphibians used in broad biodiversity assessments such as the International Union for the Conservation of Nature (IUCN) Red list are likely to be largely inaccurate (Ficetola *et al.*, 2014). Out of 427 amphibian species inhabiting the 6 million km^2 of Amazonia according to IUCN, at least 150 species (35%) are distributed over more than 1 million km^2 (Fouquet *et al.*, 2007a). Such a high proportion of broadly-distributed species seems unlikely (Wynn & Heyer, 2001), because amphibians usually display low dispersal capacities and often have small niches (Duellman & Trueb, 1994; Wells, 2010). This gap in our understanding of the actual diversity and distribution of species could seriously invalidate conclusions drawn from IUCN data (Foden *et al.*, 2013; Jenkins *et al.*, 2013, 2015; Pimm *et al.*, 2014; Feeley & Silman, 2016).

The overall aims of this study were (1) to obtain a new georeferenced dataset of Amazonian anurans based on molecular diversity, with a focus on the eastern Guiana Shield (EGS) (east of the Tepuis, and north of Rio Negro and Rio Amazonas), (2) to provide estimates of the number of species and of their distributions in this part of Amazonia, (3) to infer data-driven spatial boundaries between BSRs, as well as to re-assess their rate of endemism. Given that anuran species boundaries and distributions are plagued with uncertainty in Amazonia and that IUCN data are often out-dated and imprecise, it is necessary to use occurrence records linked to taxonomic frameworks based on clear criteria. Therefore, we conducted extensive fieldwork to collect specimens representative of present-day diversity at the scale of the entire region, and obtained mitochondrial DNA sequences (16S rDNA) from these specimens. We also included in our analyses publicly available sequences from other specimens. Based on

these sequences, we generated two new taxonomic frameworks for Amazonian anurans. Our dataset represents the largest molecular diversity dataset gathered so far in Amazonia for any taxonomic group.

Material and methods

1. Fieldwork

We undertook fieldwork in several localities throughout the Guiana Shield, notably in southern Suriname, French Guiana, and the Brazilian states of Amapá and Roraima. We collected specimens of as many anuran species as possible per locality by nocturnal and diurnal active searches (visual and acoustic). Each specimen was identified and photographed. They were subsequently euthanized using an injection of Xylocaine® (lidocaine chlorhydrate). Tissue samples (liver or muscle tissue from thigh or toe-clip) were removed and stored in 95% ethanol, while specimens were tagged and fixed (using formalin 10%) before being transferred to 70% ethanol for permanent storage. These field surveys allowed us to cover the anuran communities of the EGS at an unprecedented fine scale (Fig. 1A). We completed these data for the rest of Amazonia with loans of material from several institutions, notably from Universidad de Sao Paulo for the upper Madeira, lower Xingu, Abacaxis and Purus Rivers. Ultimately, the total number of analysed samples reached 4,681.

2. Molecular data

We extracted DNA from the samples using the Wizard Genomic extraction protocol (Promega; Madison, WI, USA). We targeted a c.a. 400bp fragment of the mitochondrial 16S rDNA using MiSeq and Sanger techniques (Supplementary Methods). We eventually generated 4,492 sequences.

Additionally, we retrieved from GenBank (as of the 1st August 2015) all sequences of species congeneric with those occurring in the Guiana Shield, as well as sequences of *Adelphobates* and *Phyzelaphryne*, two genera restricted to southern Amazonia. We removed low-quality or too short sequences, as well as duplicates from

the same specimen. We obtained approximate geographical coordinates for most of these records searching the original papers, locality information, or collection databases.

The final dataset contained 11,166 sequences, 10,254 of which were geotagged. This barcode dataset is probably the most extensive gathered so far in Amazonia for any vertebrate group. 8,181 records are from Amazonia proper, including 4,634 from the EGS, while the remaining are from adjacent regions. The obtained sequences were aligned with MAFFT v.7 (Katoh & Standley, 2013). We used the resulting alignment to generate a neighbour-joining tree using pairwise deletion and *p*-distance model with MEGA v.7.0.16 (Kumar *et al.*, 2016).

3. Taxonomic frameworks

While there is valid criticism against reliance on simplistic single-sequence approaches to species delineation (Goldstein & DeSalle, 2011; Krishna Krishnamurthy & Francis, 2012), such approaches can take us further toward the comparative quantification of biodiversity over different spatial scales (Emerson *et al.*, 2011; Yu *et al.*, 2012; Ji *et al.*, 2013). In the case of Amazonian anurans, clear and exhaustive delimitation of species boundaries based on morphology, acoustics and molecular data remains out of reach. As a consequence, many species groups have a very confused taxonomy leading to frequent misidentification, lumping of undescribed species within a single taxon, and assigning species to polyphyletic groups. This results in largely inaccurate IUCN data. In order to compare our sequence dataset to IUCN data, we built two different taxonomic frameworks. The TAXO1 taxonomic framework is conservative, linking as much as possible each sequence to a nominal taxon so as to form a monophyletic group, while the TAXO2 taxonomic framework results from a purely DNA-based species delineation (see below).

For TAXO1, our goal was to group under nominal taxa the sequences forming a monophyletic group according to the neighbour-joining tree, so as to obtain the geographic range of the lineages already considered by the IUCN. Original fieldwork and GenBank assignments were often contradictory because of the above-mentioned

reasons and because of taxonomic changes subsequent to identification, and were thus often modified. We first identified the sequences that could be unambiguously linked to a nominal taxon by considering the literature (e.g. sequences from type series), the known range of the taxon, and the location of the type locality. Then, we checked wether this identification was in accordance with the ID of the most closely related sequences. If in accordance, this taxon ID was applied to the sequences until another taxon was applicable to more distant lineage. When a taxon was found to be paraphyletic, we checked for possible misidentification, and whether one of the lineages could be identified as another taxon. When paraphyly was ambiguous, we kept the original identification. When paraphyly was unambiguous, one of the lineages was identified as the nominal taxon while the other ones were identified as "sp." if they did not share affinities with another taxon. In a few cases, two or more taxa were largely intricate with shallow genetic distances among sequences and remained ambiguous despite the allopatric distribution of the lineages. We then considered them as single taxon (e.g. *Atelopus hoogmoedi*, *A. flavescens*) given they represent single lineage and single patch of distribution. Ultimately, we think that TAXO1 provides a representative update of the current taxonomic knowledge for Amazonian anurans. 941 species were considered in TAXO1, including 365 occurring in Amazonia.

For TAXO2 we applied the Automatic Barcode Gap Discovery (ABGD) species delineation method (Puillandre *et al.*, 2012) to our sequence dataset. We performed ABGD analyses from the source code with default settings (JC69, Pmin: 0.001, Pmax: 0.1, steps: 10, Nb bins: 20) on each genus, and attributed a number to each candidate species retrieved in the analysis. Computations were performed on the EDB-Calc Cluster hosted by the laboratory "Évolution et Diversité Biologique" (EDB), using a software developed by the Rocks(r) Cluster Group (San Diego Supercomputer Center, University of California, San Diego and its contributors. In 24 instances (17 concerning Amazonian taxa), different nominal taxa in TAXO1 were lumped into a unique candidate species in TAXO2 because of a shallow mtDNA divergence between them (notably in *Atelopus* spp. and *Osteocephalus* ssp.). As these correspond to clearly distinct species based on morphology and acoustic, and form monophyletic groups in previous studies (but herein with shallow divergence or recovered ambiguously paraphyletic due to the low resolution in our 400 bp-long 16S fragment), we considered them as false negative and

we applied to them the same taxonomic assignment as in TAXO1. Ultimately, 1,246 species were considered in TAXO2, including 746 occurring in Amazonia.

Third, we compiled amphibian species range data from the IUCN (http://www.iucnredlist.org/technical-documents/spatial-data#amphibians), which is the most widely used amphibian distribution database. In order to make this dataset comparable with TAXO1 and TAXO2, we excluded 22 genera (433 species) that are only partly overlapping with our focal area, i.e., western Amazonia, northern Andes, Caatinga and Cerrados. One genus from the Tepuis (*Metaphryniscus*) was also omitted given that no sequences were available, as well as two introduced species (*Eleutherodactylus johnstonei* and *Lithobates catesbeianus*). Overall, 51 genera were used in our analyses.

4. Study area and species distribution data

Our analyses focused on a rectangular area that includes the whole central, eastern and northern parts of Amazonia (excluding most of the western and southern parts). The limits of our study area were W 72° W 47° and S 11° N 9°. We applied a grid of 1° × 1° (500 cells) to this area. This includes the Guiana Shield (Lujan & Armbruster, 2011), the central and eastern parts of the Rio Amazonas drainage, and the northern parts of the Rio Purus, Rio Madeira, Rio Tapajós, Rio Xingú, and Rio Tocantins drainages (Fig. 1A) as well as peripheral non-Amazonian areas.

We then estimated the putative range of each species by creating convex polygons out of our occurrence datasets TAXO1 (358 species total within the focal area) and TAXO2 (596 species) with the *sp* package implemented in R (R Development Core Team, 2016). The numbers of Amazonian species included in TAXO1 and TAXO2 differ from those occurring within the focal area because this area encompasses non-Amazonian areas and excludes western and southern parts of Amazonia. We then interpolated the occurrence of species in each cell of our study area for the three datasets. We excluded species occurring in less than three localities and cells with less than five species in them, thus removing poorly sampled species, that did not provide enough information for range reconstruction, and poorly sampled peripheral cells. 118

species were discarded in TAXO1 and 318 in TAXO2. Finally, we considered 240 species in TAXO1, 278 in TAXO2, and 440 in the IUCN dataset within the focal area (Fig. 1D, E, F).

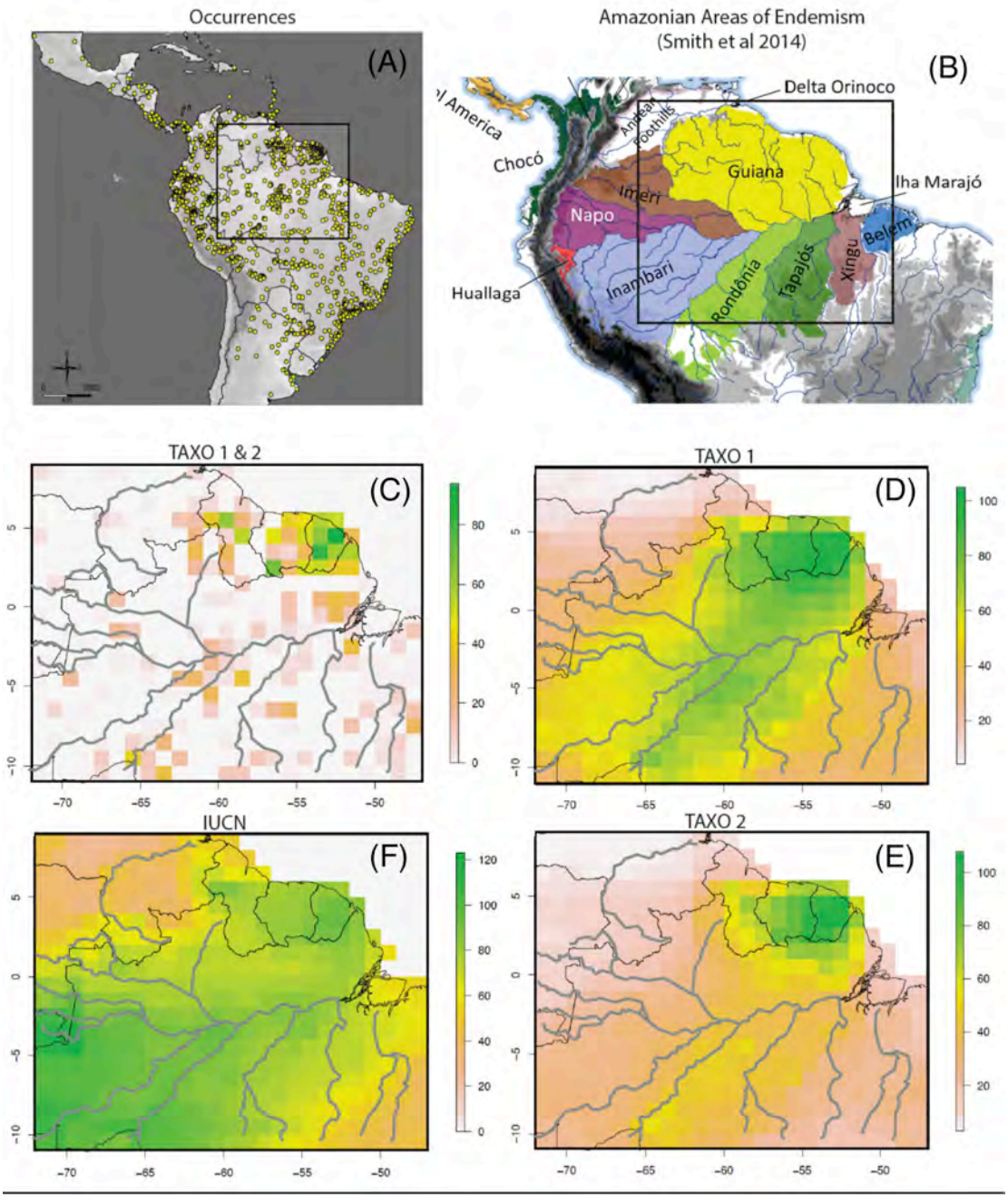

Figure 1. (A) All occurrences in the barcoding dataset and inset of the focal area; (B) Amazonian Areas of Endemism from Smith et al., 2014; (C) species richness mapped from occurrences data from TAXO1 and TAXO2, which provide identical results; (D) species richness mapped from TAXO1 after polygon transformation and exclusion of rare species; (E) species richness mapped

from TAXO2 after polygon transformation and exclusion of rare species; (F) species richness mapped from the distribution data of IUCN considered in our analyses.

5. Identification of Biogeographic Subregions

To delimit BSRs based on species occurrence, we decomposed the community matrix - i.e., the matrix listing the species occurring in each grid cell - using Latent Dirichlet Allocation (Blei *et al.*, 2003; Valle *et al.*, 2014). LDA is an unsupervised clustering method based on a probabilistic model, which assumes that several species assemblages coexist over the study area, the number *K* of which is fixed beforehand. This method has major advantages compared to classic clustering (e.g., hierarchical or k-means clustering). First, it is likelihood-based, thus providing rigorous tools for selecting the number of assemblages and comparing decompositions. Second, assemblages may partially overlap in taxonomic composition, and a given grid cell may either be dominated by one assemblage or contain a mixture of assemblages. Thus, it allows for modelling gradual changes in taxonomic composition over space. A mixing parameter α is estimated as part of the inference procedure, and indicates whether the samples tend to be decomposed into an even mixture of assemblages (case $\alpha > 1$) or into an uneven mixture dominated by one assemblage (case $\alpha < 1$).

We used the Variational Expectation Maximization (EM) algorithm implemented by Blei et al. (2003) and wrapped into the R package *topicmodels* (Grün & Hornik, 2011) for parameter inference, with a convergence threshold of 10^{-6} for the EM step and 10^{-8} for the variational step. We assessed the reliability of the solution by comparing the taxonomic composition of assemblages between 100 realizations of the algorithm starting from random initial conditions. We only interpreted the decomposition corresponding to the realization with the highest likelihood out of 100. We selected the number *K* of assemblages by AIC minimization. We represented the spatial distribution of assemblages on a map after ordinary Kriging between cells (R package *gstat* ; Pebesma, 2004). We also computed the Jaccard taxonomic dissimilarity between assemblages and displayed it as a dendrogram. Additionally, we decomposed the

datasets into *K*=3 assemblages to assess the coarser biogeographic structure of the study area. See Sommeria-Klein et al. (*in prep.*) for further methodological details.

Results

Underestimation of species richness. Based on our analyses, among the 363 Amazonian species found in TAXO1, 53 genetic lineages could not be associated with any nominal taxa. In the EGS, most of these undescribed lineages were already documented (e.g., *Adelophryne* sp., *Scinax* sp. 2, or *Pristimantis* sp. 1) (Fouquet *et al.*, 2007b, 2012b). In southern and western Amazonia however, several lineages are reported here for the first time (e.g., *Allobates* sp. "Divisor", *Amazophrynella* sp. "Acre", *Dendropsophus* sp. "Xingú"), This suggests that species diversity has been well sampled in the lowlands of the Guiana Shield, but not in the rest of Amazonia. Our datasets also provide evidence of range extension for many taxa compared to previous knowledge. This is for example the case of *Scinax nasicus*, which extends to the Sipaliwini savannah (Suriname), *Pristimantis koheleri,* to the southern part of the Guiana Shield, or *Synapturanus mirandariberoi*, to the southern part of the Amazonas drainage. However, most of these newly documented populations are highly genetically divergent from the populations lying within the known range of the species and are considered as independent species in TAXO2.

In fact, 246 TAXO1 species display splits, yielding 568 species (× 2.3) in TAXO2. TAXO2 provides 1,548 pairwise comparisons among species that are lumped as conspecific in TAXO1. 39% of these average pairwise distances (p-distance pairwise deletion) were above 6%, a threshold believed to conservatively delimit species (Vences *et al.*, 2005; Fouquet *et al.*, 2007a) and 85% were above 3% (Fig. 2A). In terms of taxonomy, 436 TAXO2 species cannot be assigned to any of the 310 nominal taxa of TAXO1. These observations suggest that the TAXO1 framework remains overconservative in many instances.

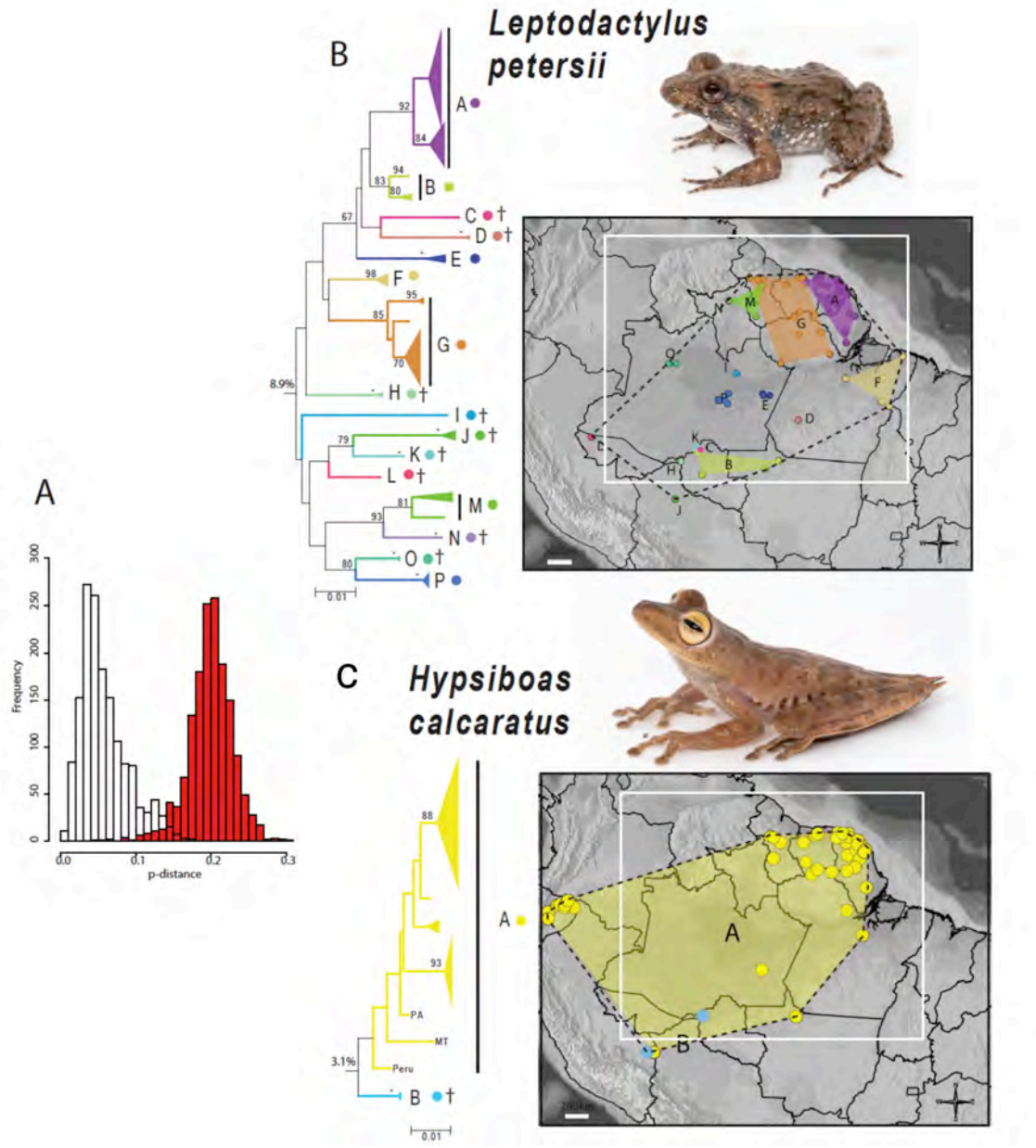

Figure 2. A: Histogram of the average pairwise distances among TAXO2 species considered as a single TAXO1 species (white bars) and among TAXO2 species considered as different TAXO1 species (red bars; this last distribution was randomly sampled to harbour the same number of comparisons than in the previous one); (B-C) Examples of genetic and geographic patterns for two Panamazonian single TAXO1 species that provide drastically different patterns in TAXO2; *Leptodactylus petersii* being split into 16 species whereas *Hypsiboas calcaratus* is only split in two candidate species in TAXO2. The colours of the lineages on the tree correspond to the colours of the occurrence points and areas on the map. † indicates candidate species that were discarded from the analyses in TAXO2 (less than three locality records).

A number of distinct patterns of distribution emerge from the occurrence data of TAXO1 and TAXO2. We highlight three of them that segregate groups of species occurring in the EGS: Guiana Shield endemic groups; Panamazonian allopatric groups and widespread species. The first pattern concerns five groups that are endemic to the Guiana Shield and occur in both the highlands and the lowlands: *Adelophryne* (4 species in TAXO1 *vs.* 4 in TAXO2), *Otophryne* (3 *vs.* 3 species), *Synapturanus* (3 *vs.* 4 species), *Anomaloglossus* (15 *vs.* 29 species), *Vitreorana ritae* clade (3 *vs.* 3 species), *Hypsiboas benitezi* clade (3 *vs.* 3 species). Among them, only *Anomaloglossus* seems to have substantially diversified in the lowlands. Secondly, the vast majority of species occurring in the EGS are nested in widespread Amazonian or lowlands Neotropical clades (Fig. 2B). Most of these clades display deep divergence among populations (above 6%; e.g. *Leptodactylus petersii* – 16 candidate species in TAXO2) and contain several candidate species with more restricted ranges. Finally, 78 species out of 358 (22%) in TAXO1, 45 out of 596 (8%) in TAXO2 and 142 out of 440 (32%) in IUCN actually have broad distributions (>1 millions km^2) within our focal study area (e.g., *H. calcaratus*) (Fig. 2C).

Biogeographical subregions. We decomposed the TAXO1, TAXO2 and IUCN datasets using Latent Dirichlet Allocation. AIC minimization yielded an optimal number of species assemblages close to K = 8 for all three datasets (Fig. S2). The retrieved assemblages were found to be spatially segregated (mixing parameter α much smaller than 1: $\alpha_{IUCN} = 0.021, \alpha_{TAXO1} = 0.019, \alpha_{TAXO2} = 0.016$) and contiguous. We could thus interpret them as BSRs. The LDA decomposition was found to be reliable for the three datasets based on its stability over 100 realizations (Fig. S2).

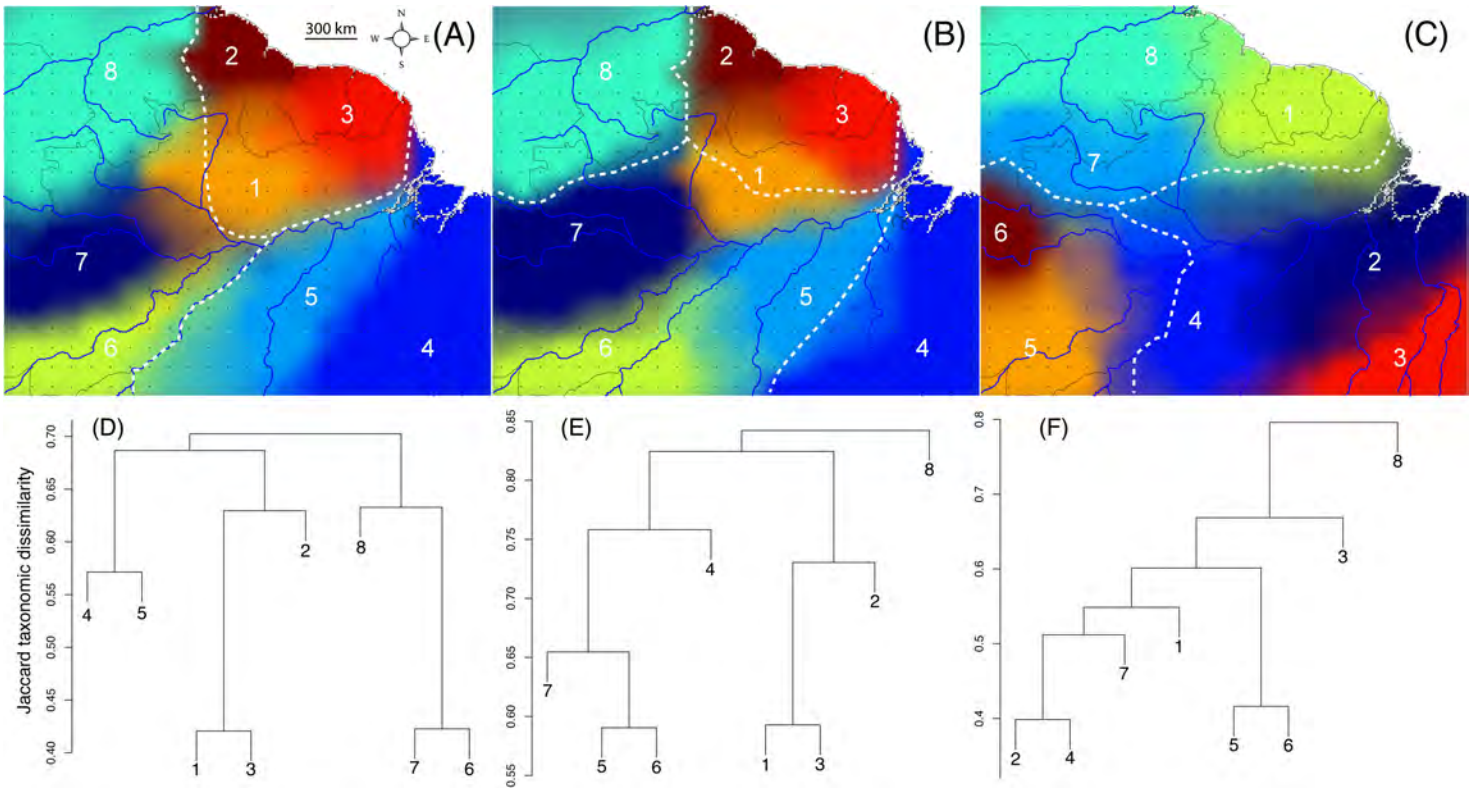

Figure 3. Maps generated by interpolating the eight-assemblage Latent Dirichlet Allocation (LDA) decomposition of the species occurrence data (A, B, C), and corresponding dendrograms showing the relationships between the eight assemblages recovered in the LDA decomposition using average Jaccard taxonomic dissimilarity (based on the presence/absence of species in assemblages). (A) TAXO1; (B) TAXO2; (C) IUCN data. The white dashed lines represent the approximate boundaries of the BSR for a three-assemblage LDA decomposition (in panel [B], the north-western and south-eastern regions belong to the same assemblage). The numbers on the maps correspond to the numbers attributed to assemblages for each dataset. (D) TAXO1; (E) TAXO2; (F) IUCN data.

Even though not identical, the spatial boundaries of the eight BSRs retrieved for TAXO1 and TAXO2 were very similar (Fig. 3A-B). The lowlands of the EGS were clearly separated from the rest of the study area by the Rio Amazonas and the Pantepui region. Moreover, the EGS was also found to exhibit some internal structure, since this area was composed of three independent BSRs, all found in both TAXO1 and TAXO2 despite large differences in the distribution of the species considered (e.g., *Leptodactylus petersii*). One of these three BSRs (BSR1 on Fig. 3A-B) comprised the southern part of Guyana, Roraima and the Northern parts of Pará and Amazonas states (Brazil). A second one (BSR2 on Fig. 3A-B) comprised the northern part of Guyana and adjacent Venezuela. Finally, a third one (BSR3 on Fig. 3A-B) comprised the state of Amapá (Brazil), French

Guiana, and Suriname. These three BSR were retrieved as a single cluster in the coarser 3-assemblage LDA decomposition. Taxonomic comparison between assemblages indicated that among these three BSR, BSR1 and BSR3 were more similar to each other, in both TAXO1 and TAXO2 (Fig. 1D, E). The only notable difference between TAXO1 and TAXO2 in the EGS area was that the boundaries of BSR1 matched well the Rio Negro and Rio Amazonas in TAXO2, while BSR1 extended somewhat further west across the Rupununi savannah in TAXO1. The boundaries between BSRs in this specific area were also sharper in TAXO2 than in TAXO1. Outside of the EGS area, there was a striking match between BSR boundaries and Rio Madeira in TAXO1 that was already recovered in the 3-assemblage decomposition. In contrast, the Purus and Tapajòs Rivers were found to be each at the center of a BSR in both TAXO1 and TAXO2.

The distribution of BSRs using the IUCN database provided a markedly different pattern, notably not matching the EGS boundaries. The three Guianas (Guyana, Suriname, and French Guiana) were grouped together in one BSR, excluding the north-western part of Guyana and including adjacent areas of Amapá and Pará (Brazil). The southern part of the EGS was grouped with the southern part of the Amazon drainage, thus encompassing Rio Amazonas (Fig. 3C).

Species richness and endemism. In terms of species richness and endemism, the three datasets are radically different. The BSR1 of IUCN is composed of 119 species, 27.7 % of which are endemic (Table 1), and is geographically comparable to the lumping together of BSR2 and 3 in TAXO1 and TAXO2. Yet, despite encompassing a smaller geographical area, the BSR3 of TAXO1 alone displays similar values of richness and endemism as the BSR1 of IUCN. When considering the three Guiana Shield BSRs together in TAXO1, richness (184 species) and endemism (57 %) are much higher than in the BSR1 of IUCN. These metrics increase to 250 species and 82.4 % endemism in TAXO2 for the EGS (Table 1). BSR2 (Northern Guyana) contains the highest number of endemic species in both taxonomic frameworks, reaching 75 % endemism in TAXO2 (Table 1), while the highest species richness (130 in TAXO2) is found in BSR3 (Suriname, French Guiana and Amapá).

		UICN			TAXO1			TAXO2		
Partition	BSR	Species richness	Endemic species	Endemism rate (%)	Species richness	Endemic species	Endemism rate (%)	Species richness	Endemic species	Endemism rate (%)
K = 8	1	119	33	27.7	89	4	0.4	71	25	35.2
	2	–	–	–	85	46	54.1	90	68	75.5
	3	–	–	–	118	30	25.4	130	77	59.2
K = 3	1	–	–	–	184	105	57	250	206	82.4

Table 1: Species richness and endemism in each of the BSRs covering the EGS. The figures presented in this table include singletons (species with only one occurrence point) and species that occur in less than three cells. BSR numbers correspond to those displayed in Fig. 3. For K=3, assemble 1 actually corresponds to the EGS.

Discussion

Underestimation of species richness and regional endemism in Amazonia. We analysed our molecular diversity data using two alternative taxonomic frameworks: a conservative framework TAXO1 in which sequences were as much as possible clustered into monophyletic groups around previously described nominal taxa, and a framework TAXO2 in which species were delineated solely based on the molecular distance between sequences. Our species delineation analysis corroborates previous suggestions that the actual number of anuran species occurring in Amazonia remains vastly underestimated (Fouquet *et al.*, 2007a; Funk *et al.*, 2012; Ferrão *et al.*, 2016). The number of species retrieved in TAXO2 (746) and the level of divergence among them are particularly striking in many groups.

Our TAXO1 dataset comprises 363 Amazonian species, which is close to the 427 species recorded by the IUCN. However, our sampling effort is low outside the EGS, as illustrated by the fact that we do not retrieve several nominal taxa included in the IUCN database. Therefore, the actual number of species is likely to be largely underestimated in TAXO1 outside the EGS. Moreover, TAXO1 remains over-conservative in many instances, as the level of genetic divergence within species is often very high. TAXO2 suggests the existence of more than twice the number of species found in TAXO1. Considering that uneven sampling is even more of an issue in TAXO2 than in TAXO1, as many of our candidate species are only retrieved in one or a few localities, the actual species count for Amazonia is likely to be substantially more than twice the current count. Hence, comparisons between taxonomic frameworks should be limited to the EGS, where our sampling effort is highest. When considering solely the EGS, the number of candidate species retrieved in TAXO2 is 1.34 times higher than for TAXO1 (Table 2).

A species delineation solely based on mtDNA divergence remains overly simplistic and cannot reliably delineate the species occurring in the region since it necessarily overestimates the actual number of species in some cases (false positives) and underestimates in others (false negatives) (Hickerson *et al.*, 2006). The pitfalls inherent to the sole use of short mtDNA sequences for species delineation have been

already extensively discussed (Hubert & Hanner, 2015). Nevertheless, in most groups for which the boundaries among species have been investigated using integrative taxonomy, mtDNA divergence of similar magnitude as used in this study to differentiate between intra- and interspecific genetic divergence was generally associated with phenotypic or acoustic differentiation as well (Funk *et al.*, 2012; Fouquet *et al.*, 2015b; Ortega-Andrade *et al.*, 2015; Fouquet *et al.*, 2016). Moreover, TAXO2 subdivisions have already been proven to be associated with morphological or acoustic differences in several groups (Jansen *et al.*, 2011; Fouquet *et al.*, 2013; Ferrão *et al.*, 2016). Thus, the TAXO2 taxonomic framework takes into account finer subdivisions that certainly correspond to phenotypically distinct species in many cases, and it is highly probable that the prevalence of false positives remains limited. In contrast, some false negatives were detected since several nominal taxa were retrieved as a single candidate species using ABGD (e.g., *Atelopus flavescens* and *A. hoogmoedi*, *O. oophagus* and *O. taurinus*). These were corrected in TAXO2 but the prevalence of false negatives remains difficult to evaluate in most groups where species boundaries have not been investigated using phenotypic traits. Overall, the present work provides an important update to the documentation of Amazonian anuran diversity, which will undoubtedly contribute to stimulate the process of species delineation and description.

If our work provides a glimpse of how far we still are from reaching a realistic estimate of the number of species occurring throughout Amazonia, it also provides an even more striking view of the degree of regional endemism. Our estimates of the rate of endemism for the frogs of the EGS reach 57.0 % based on TAXO1 and 82.4 % based on TAXO2. These figures are two to four times higher than the estimate of the IUCN for the same area. They are also 1.0 to 1.4 times higher than the rate of endemism of frogs in the whole geologically defined Guiana Shield, which also encompasses Venezuela and part of Colombia (Señaris & MacCulloch, 2005). In comparison, only 7.7% of bird species are endemic to the whole Guiana Shield, 29 % of reptile species, and 11 % of mammal species (Hollowell & Reynolds, 2005). These figures are still certainly underestimated (Lim, 2012), especially for reptiles (Geurgas & Rodrigues, 2010; de Oliveira *et al.*, 2016), but taxonomy has probably reached a much more stable level for birds and mammals in the Guiana Shield than for anurans. In comparison with other tropical American regions,

51.3% of the vertebrate species from the Atlantic Forest of Brazil are endemic, and 46.2 % of the vertebrates from the tropical Andes are endemic (Myers *et al.*, 2000).

A simple and rough extrapolation based on the species richness and endemism we obtained for the EGS (184–250 species with 57–82 % endemism) applied to the eight Amazonian BSRs retrieved in our analysis leads to ca. 1,472–2,000 species in our focal area, which represent about three to five times the 427 species that are supposed to occur in Amazonia according to the IUCN. Enhancing data coverage in order to refine these estimations would require extensive fieldwork in remote areas. Nevertheless, new predictive approaches based on the detection of cryptic diversity (Espíndola *et al.*, 2016) may permit to get a more precise estimate of species richness and endemism in each BSR, and therefore would help targeting areas where to focus sampling.

Biogeographic division of the eastern Guiana Shield. The extent of the BSRs retrieved for TAXO1 and TAXO2 are very similar in spite of the use of two drastically different taxonomic frameworks. In contrast, the BSRs retrieved from the IUCN database are very different and do not correspond to any landscape feature. No barrier effect of the lower Rio Amazonas is even distinguishable. This is most likely resulting from the artificially large distribution of many species contained in this database on both sides of this river.

The location of the Rio Madeira matches well the boundary between BSR5 and BSR6 in TAXO1, which is in accordance with what has already been shown for other groups of terrestrial vertebrates, such as birds (Fernandes *et al.*, 2012; Ribas *et al.*, 2012) and primates (Cortés-Ortiz *et al.*, 2003). The sharpness of this pattern is not obvious in TAXO2, but this is probably due to the removal of many singletons from the dataset after species delineation. Another interesting aspect is the lack of apparent suture effect between the Purus and the Solimões drainages, also in accordance with what has previously been found for other group of terrestrial vertebrates (Cortés-Ortiz *et al.*, 2003; Fernandes *et al.*, 2012; Ribas *et al.*, 2012). These rivers display a meandering behaviour associated with an unstable course over time, thus enabling gene flow through connection between populations located on both sides and dispersal of species from one interfluve to the other (Aleixo, 2004, 2006; Bates *et al.*, 2004; Jackson *et al.*,

2013). On the contrary, wide rivers in the Brazilian shield such as Rio Madeira display a putatively more stable course over time and are more likely to act as long lasting suture zones that might have promoted diversification or at least been more efficient in preventing dispersal (Antonelli *et al.*, 2010; Moraes *et al.*, 2016). Such characteristics are also found in rivers of the EGS (Fernandes *et al.*, 2012; Fouquet *et al.*, 2012a, 2015a), but except for the Rio Branco and Rio Negro, the impact of the Guiana Shield rivers on gene flow through dispersal limitation might not be as important as for the Amazonian rivers of the Brazilian Shield, owing to the smaller extent of the catchments and the smaller width of the rivers themselves. This is reflected in our results, as the suture zones between the three BSRs of the EGS do not correspond to any major drainage. In fact, it is more likely that the delimitation of these assemblages resulted from the combined influence of past climatic and landscape changes (Fouquet *et al.*, 2012c). The current climatic characteristics of the EGS are heterogeneous, with a large dryer corridor observed in the southern part (Mayle & Power, 2008), where patches of savannahs are found today. This corridor also matches the suture zone between BSR1 vs. BSR2 and BSR3. The strong climatic fluctuations in the Neotropics during the Miocene and Pliocene played a crucial role in the diversification of several organisms (Antonelli *et al.*, 2010). More recent climate fluctuations and associated landscape modifications during the Pleistocene certainly helped maintain the diversity that resulted from diversification events during the Miocene and Pliocene periods (Carnaval & Bates, 2007).

The outer limits of the three BSRs match well the delimitation of the Guianan area retrieved for birds (Naka, 2011), confirming the relevance of qualifying the EGS as a biogeographic area. Nonetheless, using anuran assemblages as a model revealed biogeographic heterogeneity within this region that could not be detected with bird assemblages, likely because birds have much higher dispersal abilities than anurans (Pigot & Tobias, 2015). The distinctiveness of the BSRs compared to the remaining of the dataset is also reflected in the structure of the dendrogram illustrating the level of taxonomic similarity between assemblages (Fig. 3D, E). The southern limit of BSR1 corresponds to Rio Amazonas for both TAXO1 and TAXO2. This is congruent with previous studies on terrestrial vertebrates indicating that this river is a strong barrier to gene flow and that it structures species assemblages (Cortés-Ortiz *et al.*, 2003; Haffer, 2008; Ribas *et al.*, 2012). The delineation of the western part of BSR1 differs across

datasets. It coincides perfectly with the lower Rio Negro, and the Rio Branco and associated savannahs (Rupununi) in TAXO2 but extends further west in TAXO1. These differences are inherent to the scarcer sampling west and south-west of the Rio Negro and Rio Branco, weakening the sharpness of the analysis in that zone, a phenomenon that becomes even more prevalent in TAXO2 because of the further taxonomic subdivisions. Another reason could be the inclusion of both forest and open habitat species in our analysis, which could blur the pattern in areas where both savannah and forest are found.

It is interesting to note that the limits of the BSRs of the EGS are rather similar when considering either a K=3 or a K=8 decomposition, for both TAXO1 and TAXO2. This indicates that a strong co-occurrence signal underlies the delineation of these BSRs, especially in the case of the two northernmost ones (BSR2 and BSR3) whose western and eastern boundaries coincide perfectly with the ones retrieved in the three-assemblage decomposition (Fig. 3).

Conclusion. Despite being far from exhaustive, our barcoding dataset is the largest ever gathered for Amazonia, and we argue that it is close from being exhaustive within the EGS. Of course, the patterns we obtained need to be confirmed in other taxonomical groups, and need even for the anurans to be much improved outside the EGS. Nevertheless, our results help us understand the spatial scale of the sampling efforts needed to capture the actual diversity of Amazonia. It implies notably that the magnitude of the Linnean and Wallacean shortfalls in Amazonia is so large that we could question the conclusions of large-scale studies based on currently admitted biodiversity data in Amazonia (Feeley & Silman, 2011; Foden *et al.*, 2013). In fact, even with very coarse data (IUCN), they estimated that Amazonian amphibians are highly threatened by climate change. Considering that many species were not included and that they actually harbour much narrower distributions, we can hypothesise that the situation is even more worrying. If a degree of endemism similar to the one we estimated within the EGS actually occurs across Amazonia, the impact of habitat loss could have been underestimated. It is especially the case along the Arc of deforestation (Vedovato *et al.*, 2016), where entire faunal assemblages that may harbour a high degree of endemism

are at risk of extinction (Da Silva *et al.*, 2005). Moreover, only BSR3 encompasses a large proportion of protected areas in the EGS. In contrast, BSR2 (northern Guyana) only harbours two protected areas and the BSR1 only encompasses three biological reserves (REBIO), four national forests (FLONA) and three national parks (PARNA) in its Brazilian part. Such results demonstrate the importance of deciphering the basic structure of the Amazonian diversity in order to conserve it efficiently.

Acknowledgements

This work has benefited from an 'Investissement d'Avenir' grant managed by *Agence Nationale de la Recherche* (CEBA, ref.ANR-10-LABX-25-01), France. We would like to thank the following people for their help on the field: Daniel Baudin, Sébastien Cally, Elodie Courtois, Andy Lorenzini, Benoît Villette. We thank Pierre Solbès at Laboratoire Évolution et Diversité Biologique (Toulouse, France) for support with the EDB-cCacl cluster.

References

Aleixo, A. (2004) Historical diversification of a terra-firme forest bird superspecies: a phylogeographic perspective on the role of different hypotheses of Amazonian diversification. *Evolution*, **58**, 1303–1317.

Aleixo, A. (2006) Historical diversification of floodplain forest specialist species in the Amazon: a case study with two species of the avian genus Xiphorhynchus (Aves: Dendrocolaptidae). *Biological Journal of the Linnean Society*, **89**, 383–395.

Anderson, L.O. (2012) Biome-Scale Forest Properties in Amazonia Based on Field and Satellite Observations. *Remote Sensing*, **4**.

Antonelli, A., Quijada-Mascareñas, A., Crawford, A.J., Bates, J.M., Velazco, P.M. & Wüster, W. (2010) Molecular studies and phylogeography of Amazonian tetrapods and their relation to geological and climatic models. *In Hoorn, C., Wesselingh, F.: Amazonia, Landscape and Species Evolution, 1st edition. Blackwell publishing*, 386–404.

Antonelli, A. & Sanmartín, I. (2011) Why are there so many plant species in the Neotropics? *Taxon*, **60**, 403–414.

Bates, H.W. (1863) *The naturalist on the River Amazons, a record of adventures, habits of animals, sketches of Brazilian and Indian life and aspects of nature under the Equator during eleven years of travel*, John Murray, London.

Bates, J.M., Haffer, J. & Grismer, E. (2004) Avian mitochondrial DNA sequence divergence across a headwater stream of the Rio Tapajós, a major Amazonian river. *Journal of Ornithology*, **145**, 199–205.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning*, **3**, 993–1022.

Carnaval, A.C. & Bates, J.M. (2007) Amphibian DNA shows marked genetic strucure and tracks Pleistocene climate change in Northeastern Brazil. *Evolution*, **61**, 2942–2957.

Claramunt, S., Derryberry, E.P., Remsen, J. V & Brumfield, R.T. (2011) High dispersal ability inhibits speciation in a continental radiation of passerine birds. *Proceedings of the Royal Society B: Biological Sciences*.

Colinvaux, P.A., De Oliveira, P.E. & Bush, M.B. (2000) Amazonian and Neotropical plant communities on glacial time-scales: The failure of the aridity and refuge hypotheses. *Quaternary Science Reviews*, **19**, 141–169.

Cortés-Ortiz, L., Bermingham, E., Rico, C., Rodríguez-Luna, E., Sampaio, I. & Ruiz-García, M. (2003) Molecular systematics and biogeography of the Neotropical monkey genus, Alouatta. *Molecular Phylogenetics and Evolution*, **26**, 64–81.

Cracraft, J. (1985) Historical Biogeography and Patterns of Differentiation within the South American Avifauna: Areas of Endemism. *Ornithological Monographs*, 49–84.

Dinerstein, E., Olson, D.M., Graham, D.J., Webster, A.L., Primm, S.A., Bookbinder, M.P. & Ledec, G. (1995) *A Conservation Assessment of the Terrestrial Ecoregions of Latin America and the Caribbean*, Washigton (DC): World Bank.

Duellman, W.E. & Trueb, L. (1994) *Biology of Amphibians*, John Hopkins University Press.

Emerson, B.C., Cicconardi, F., Fanciulli, P.P. & Shaw, P.J.A. (2011) Phylogeny, phylogeography, phylobetadiversity and the molecular analysis of biological communities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**.

Espíndola, A., Ruffley, M., Smith, M.L., Carstens, B.C., Tank, D.C. & Sullivan, J. (2016) Identifying cryptic diversity with predictive phylogeography. *Proceedings of the Royal Society B: Biological Sciences*, **283**.

Feeley, K.J. & Silman, M.R. (2016) Disappearing climates will limit the efficacy of Amazonian protected areas. *Diversity and Distributions*, **22**, 1081–1084.

Feeley, K.J. & Silman, M.R. (2011) The data void in modeling current and future distributions of tropical species. *Global Change Biology*, **17**, 626–630.

Fernandes, A.M., Wink, M. & Aleixo, A. (2012) Phylogeography of the chestnut-tailed antbird (Myrmeciza hemimelaena) clarifies the role of rivers in Amazonian biogeography. *Journal of Biogeography*, **39**, 1524–1535.

Ferrão, M., Colatreli, O., de Fraga, R., Kaefer, I.L., Moravec, J. & Lima, A.P. (2016) High species richness of Scinax treefrogs (Hylidae) in a threatened Amazonian landscape revealed by an integrative approach. *PLoS ONE*, **11**, e0165679.

Ficetola, G.F., Rondinini, C., Bonardi, A., Katariya, V., Padoa-Schioppa, E. & Angulo, A. (2014) An evaluation of the robustness of global amphibian range maps. *Journal of Biogeography*, **41**, 211–221.

Foden, W.B., Butchart, S.H.M., Stuart, S.N., Vié, J.-C., Akçakaya, H.R., Angulo, A., DeVantier, L.M., Gutsche, A., Turak, E., Cao, L., Donner, S.D., Katariya, V., Bernard, R., Holland, R.A., Hughes, A.F., O'Hanlon, S.E., Garnett, S.T., Şekercioğlu, Ç.H. & Mace, G.M. (2013) Identifying the World's Most Climate Change Vulnerable Species: A Systematic Trait-Based Assessment of all Birds, Amphibians and Corals. *PLoS ONE*, **8**, e65427.

Fouquet, A., Courtois, E.A., Baudain, D., Lima, J.D., Souza, S.M., Noonan, B.P. & Rodrigues, M.T. (2015a) The trans-riverine genetic structure of 28 Amazonian frog species is dependent on life history. *Journal of Tropical Ecology*, **31**, 361–373.

Fouquet, A., Gilles, A., Vences, M., Marty, C., Blanc, M. & Gemmell, N.J. (2007a) Underestimation of species richness in neotropical frogs revealed by mtDNA analyses. *PlosOne*, **2**, e1109.

Fouquet, A., Ledoux, J.-B., Dubut, V., Noonan, B.P. & Scotti, I. (2012a) The interplay of dispersal limitation, rivers, and historical events shapes the genetic structure of an Amazonian frog. *Biological Journal of the Linnean Society*, **106**, 356–373.

Fouquet, A., Loebmann, D., Castroviejo-Fisher, S., Padial, J.M., Orrico, V.G.D., Lyra, M.L., Roberto, I.J., Kok, P.J.R., Haddad, C.F.B. & Rodrigues, M.T. (2012b) From Amazonia to the Atlantic forest: Molecular phylogeny of Phyzelaphryninae frogs reveals unexpected diversity and a striking biogeographic pattern emphasizing conservation challenges. *Molecular Phylogenetics and Evolution*, **65**, 547–561.

Fouquet, A., Martinez, Q., Courtois, E.A., Dewynter, M., Pineau, K., Gaucher, P., Blanc, M., Marty, C. & Kok, P.J.R. (2013) A new species of the genus Pristimantis (Amphibia, Craugastoridae) associated with the moderately elevated massifs of French Guiana. *Zootaxa*, **3750**, 569–586.

Fouquet, A., Martinez, Q., Zeidler, L., Courtois, E.A., Gaucher, P., Blanc, M., Lima, J.D., Souza, S.M., Rodrigues, M.T. & Kok, P.J.R. (2016) Cryptic diversity in the Hypsiboas semilineatus species group (Amphibia, Anura) with the description of a new species from the eastern Guiana Shield. *Zootaxa*, **4084**, 79–104.

Fouquet, A., Noonan, B.P., Rodrigues, M.T., Pech, N., Gilles, A. & Gemmell, N.J. (2012c) Multiple quaternary refugia in the Eastern Guiana Shield revealed by comparative phylogeography of 12 frog species. *Systematic Biology*, **61**, 461–489.

Fouquet, A., Orrico, V.G.D., Ernst, R., Blanc, M., Martinez, Q., Vacher, J.-P., Rodrigues, M.T.,

Ouboter, P., Jairam, R. & Ron, S. (2015b) A new Dendropsophus Fitzinger, 1843 (Anura: Hylidae) of the parviceps group from the lowlands of the Guiana Shield. *Zootaxa*, **4052**, 39–64.

Fouquet, A., Recoder, R., Teixeira Jr., M., Cassimiro, J., Amaro, R.C., Camacho, A., Damasceno, R., Carnaval, A.C., Moritz, C. & Rodrigues, M.T. (2012d) Molecular phylogeny and morphometric analyses reveal deep divergence between Amazonia and Atlantic Forest species of Dendrophryniscus. *Molecular Phylogenetics and Evolution*, **62**, 826–838.

Fouquet, A., Vences, M., Salducci, M.D., Meyer, A., Marty, C., Blanc, M. & Gilles, A. (2007b) Revealing cryptic diversity using molecular phylogenetics and phylogeography in frogs of the Scinax ruber and Rhinella margaritifera species groups. *Molecular Phylogenetics and Evolution*, **43**, 567–582.

Funk, W.C., Caminer, M. & Ron, S.R. (2012) High levels of cryptic species diversity uncovered in Amazonian frogs. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 1806–1814.

Gehara, M., Crawford, A.J., Orrico, V.G.D., Rodríguez, A., Lötters, S., Fouquet, A., Barrientos, L.S., Brusquetti, F., De la Riva, I., Ernst, R., Urrutia, G.G., Glaw, F., Guayasamin, J.M., Hölting, M., Jansen, M., Kok, P.J.R., Kwet, A., Lingnau, R., Lyra, M., Moravec, J., Pombal Jr, J.P., Rojas-Runjaic, F.J.M., Schulze, A., Señaris, J.C., Solé, M., Rodrigues, M.T., Twomey, E., Haddad, C.F.B., Vences, M. & Köhler, J. (2014) High levels of diversity uncovered in a widespread nominal taxon: continental phylogeography of the neotropical tree frog Dendropsophus minutus. *PLoS ONE*, **9**, e103958.

Geurgas, S.R. & Rodrigues, M.T. (2010) The hidden diversity of Coleodactylus amazonicus (Sphaerodactylinae, Gekkota) revealed by molecular data. *Molecular Phylogenetics and Evolution*, **54**, 583–593.

Goldstein, P.Z. & DeSalle, R. (2011) Integrating DNA barcode data and taxonomic practice: Determination, discovery, and description. *BioEssays*, **33**, 135–147.

Grün, B. & Hornik, K. (2011) topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software; Vol 1, Issue 13* .

Haffer, J. (1974) *Avian speciation in Tropical South America*, Nuttall Ornithological Club, 14, Cambridge, Massachusetts.

Haffer, J. (2008) Hypotheses to explain the origin of species in Amazonia. *Brazilian Journal of Biology*, **68**, 917–947.

Hall, J.P.W. & Harvey, D.J. (2002) The phylogeography of Amazonia revisited: new evidence from riodinid butterflies. *Evolution*, **56**, 1489–1497.

Hayes, F.E. & Sewlal, J.-A.N. (2004) The Amazon River as a dispersal barrier to passerine birds: effects of river width, habitat and taxonomy. *Journal of Biogeography*, **31**, 1809–1818.

Hickerson, M.J., Stahl, E.A. & Lessios, H.A. (2006) Test for Simultaneous Divergence using Approximate Bayesian Computation. *Evolution*, **60**, 2435–2453.

Hollowell, T. & Reynolds, R.P. (2005) Checklist of the terrestrial vertebrates of the Guiana Shield. *Bulletin of the Biological Society of Washington*.

Hoorn, C. & Wesselingh, F.P. (2010) *Introduction: Amazonia, landscape and species evolution*. *Amazonia: landscape and species evolution*, pp. 1–6. Wiley-Blackwell.

Hoorn, C., Wesselingh, F.P., ter Steege, H., Bermudez, M.A., Mora, A., Sevink, J., Sanmartín, I., Sanchez-Meseguer, A., Anderson, C.L., Figueiredo, J.P., Jaramillo, C., Riff, D., Negri, F.R., Hooghiemstra, H., Lundberg, J., Stadler, T., Särkinen, T. & Antonelli, A. (2010) Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. *Science*, **330**, 927–931.

Hubbell, S.P., He, F., Condit, R., Borda-de-Água, L., Kellner, J. & ter Steege, H. (2008) How many tree species are there in the Amazon and how many of them will go extinct? *Proceedings of the National Academy of Sciences*, **105**, 11498–11504.

Hubert, N. & Hanner, R. (2015) DNA Barcoding, species delineation and taxonomy: a historical perspective. *DNA Barcodes*, **3**, 44–58.

Hughes, C.E., Pennington, R.T. & Antonelli, A. (2013) Neotropical Plant Evolution: Assembling the Big Picture. *Botanical Journal of the Linnean Society*, **171**, 1–18.

Jackson, N.D., Austin, C.C., Haffer, J., Capparella, A., Haffer, J., Gascon, C., Malcolm, J., Patton, J., Silva, M. da, Bogart, J., Hayes, F., Sewlal, J., Colwell, R., Slatkin, M., Peres, C., Patton, J., daSilva, M., McLuckie, A., Lamb, T., Schwalbe, C., McCord, R., Fouquet, A., Ledoux, J., Dubut, V., Noonan, B., Scott, I., Brice, J., Stølum, H., Akin, J., Mather, C., Brooks, G., Fitch, H., Achen, P., Jackson, N., Austin, C., Jackson, N., Austin, C., Soltis, D., Morris, A., McLachlan, J., Manos, P., Soltis, P., Pyron, R., Burbrink, F., Excoffier, L., Smouse, P., Quattro, J., Jackson, N., Glenn, T., Hagen, C., Austin, C., Irwin, D., Kocher, T., Wilson, A., Austin, C., Spataro, M., Peterson, S., Jordon, J., McVay, J., DeWoody, J., Schupp, J., Kenefic, L., Busch, J., Murfitt, L., Hoffman, J., Amos, W., Pompanon, F., Bonin, A., Bellemain, E., Taberlet, P., Weir, B., Cockerham, C., Kalinowski, S., Raymond, M., Rousset, F., Stamatakis, A., Pritchard, J., Stephens, M., Donnelly, P., Jost, L., Hedrick, P., Excoffier, L., Hedrick, P., Slatkin, M., Balloux, F., Lugon-Moulin, N., Paetkau, D., Waits, L., Clarkson, P., Craighead, L., Strobeck, C., Gaggiotti, O., Lange, O., Rassmann, K., Gliddon, C., Crawford, N., Heller, R., Siegismund, H., Faubet, P., Gaggiotti, O., Barton, N., Slatkin, M., Pemberton, J., Slate, J., Bancroft, D., Barrett, J., Dakin, E., Avise, J., Evanno, G., Regnaut, S., Goudet, J., Brandley, M., Guiher, T., Pyron, R., Winne, C., Burbrink, F., O'Donnell, R., Mock, K., Austin, J., Lougheed, S., Boag, P., Fontanella, F., Feldman, C., Siddall, M., Burbrink, F., Guiher, T., Burbrink, F., Starkey, D., Shaffer, H., Burke, R., Forstner, M., Iverson, J., Zamudio, K., Savage, W., Makowsky, R., Chesser, J., Rissler, L., Niemiller, M., Fitzpatrick, B., Miller, B., Li, J., Yeung, C., Tsai, P., Lin, R., Yeh, C., Postma, E., Noordwijk, A. van, Gavrilets, S., Tatarenkov, A., Healey, C., Avise, J., Gavrilets, S., Li, H., Vose, M., Jackson, S., Webb, R., Anderson, K., Overpeck, J., Webb, T., Haywood, A., Valdes, P., Sellwood, B., Kaplan, J., Dowsett, H., Saucier, R., Estoup, A., Jarne, P., Cornuet, J., O'Reilly, P., Canino, M., Bailey, K., Bentzen, P., Frazier, D., Kesel, R., Blum, M., Guccione, M., Wysocki, D., Robnett, P., Rutledge, E., Smith, L., Baker, J., Killgore, K. & Kasul, R. (2013) Testing the Role of Meander Cutoff in Promoting Gene Flow across a Riverine Barrier in Ground Skinks (Scincella lateralis). *PLoS ONE*, **8**, e62812.

Jansen, M., Bloch, R., Schulze, A. & Pfenninger, M. (2011) Integrative inventory of Bolivia's lowland anurans reveals hidden diversity. *Zoologica Scripta*, **40**, 567–583.

Jenkins, C.N., Alves, M.A.S., Uezu, A. & Vale, M.M. (2015) Patterns of Vertebrate Diversity and Protection in Brazil. *PLoS ONE*, **10**, e0145064.

Jenkins, C.N., Pimm, S.L. & Joppa, L.N. (2013) Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences*, **110**, E2602–E2610.

Ji, Y., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A., Kitching, R., Dolman, P.M., Woodcock, P., Edwards, F.A., Larsen, T.H., Hsu, W.W., Benedick, S., Hamer, K.C., Wilcove, D.S., Bruce, C., Wang, X., Levi, T., Lott, M., Emerson, B.C. & Yu, D.W. (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Katoh, K. & Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, **30**, 772–

780.

Krishna Krishnamurthy, P. & Francis, R.A. (2012) A critical review on the utility of DNA barcoding in biodiversity conservation. *Biodiversity and Conservation*, **21**, 1901–1919.

Kumar, S., Stecher, G. & Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, **33**, 1870–1874.

Lim, B.K. (2012) Preliminary assessment of Neotropical mammal DNA barcodes: an underestimation of biodiversity. *The Open Zoology Journal*, **5**, 10–17.

Lujan, N.K. & Armbruster, J.W. (2011) *The Guiana Shield*. *Historical Biogeography of Neotropical Freshwater Fishes*, pp. 211–224. The Regents of the University of California.

Mayle, F.E. & Power, M.J. (2008) Impact of a drier Early–Mid-Holocene climate upon Amazonian forests. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 1829–1838.

Moraes, L.J.C.L., Pavan, D., Barros, M.C. & Ribas, C.C. (2016) The combined influence of riverine barriers and flooding gradients on biogeographical patterns for amphibians and squamates in south-eastern Amazonia. *Journal of Biogeography*, **43**, 2113–2124.

Morrone, J.J. (2005) Biogeographic areas and transition zones of Latin America and the Caribbean islands based on panbiogeographic and cladistic analyses of the entomofauna. *Annual Review of Entomology*, **51**, 467–494.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B. & Kent, J. (2000) Biodiversity hotspots for conservation priorities. *Nature*, **403**, 853–858.

Naka, L.N. (2011) Avian distribution patterns in the Guiana Shield: implications for the delimitation of Amazonian areas of endemism. *Journal of Biogeography*, **38**, 681–696.

Naka, L.N., Bechtoldt, C.L., Henriques L. Magalli Pinto & Brumfield, R.T. (2012) The Role of Physical Barriers in the Location of Avian Suture Zones in the Guiana Shield, Northern Amazonia. *The American Naturalist*, **179**, E115–E132.

Nelson, B.W., Ferreira, C.A.C., da Silva, M.F. & Kawasaki, M.L. (1990) Endemism centres, refugia and botanical collection density in Brazilian Amazonia. *Nature*, **345**, 714–716.

de Oliveira, D.P., de Carvalho, V.T. & Hrbek, T. (2016) Cryptic diversity in the lizard genus Plica (Squamata): phylogenetic diversity and Amazonian biogeography. *Zoologica Scripta*, **45**, 630–641.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., D'amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P. & Kassem, K.R. (2001) Terrestrial Ecoregions of the World: A New Map of Life on Earth: A new global map of terrestrial ecoregions provides an innovative tool for conserving biodiversity. *BioScience*, **51**, 933–938.

Ortega-Andrade, H.M., Rojas-Soto, O.R., Valencia, J.H., Espinosa de los Monteros, A., Morrone, J.J., Ron, S.R. & Cannatella, D.C. (2015) Insights from Integrative Systematics Reveal Cryptic Diversity in Pristimantis Frogs (Anura: Craugastoridae) from the Upper Amazon Basin. *PLoS ONE*, **10**, e0143392.

Pebesma, E. J. (2004) Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, **30**, 683-691.

Pigot, A.L. & Tobias, J.A. (2015) Dispersal and the transition to sympatry in vertebrates. *Proceedings of the Royal Society B: Biological Sciences*, **282**, 20141929.

Pimm, S.L., Jenkins, C.N., Abell, R., Brooks, T.M., Gittleman, J.L., Joppa, L.N., Raven, P.H., Roberts, C.M. & Sexton, J.O. (2014) The biodiversity of species and their rates of extinction,

distribution, and protection. *Science*, **344**.

Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. (2012) ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology*, **21**, 1864–1877.

R Development Core Team (2016) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing Vienna Austria*, **0**, {ISBN} 3-900051-07-0.

Ribas, C.C., Aleixo, A., Nogueira, A.C.R., Miyaki, C.Y. & Cracraft, J. (2012) A palaeobiogeographic model for biotic diversification within Amazonia over the past three million years. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 681 LP-689.

Señaris, J.C. & MacCulloch, R.D. (2005) *Amphibians*. *Checklist of the Terrestrial Vertebrates of the Guiana Shield* (ed. by T. Hollowell) and R.P. Reynolds), pp. 9–23. Bulletin of the Biological Society of Washington no. 13.

Da Silva, J.M.C., Rylands, A.B. & Da Fonseca, G.A.B. (2005) The fate of the Amazonian areas of endemism. *Conservation Biology*, **19**, 689–694.

Sioli, H. (1984) *The Amazon: Limnology and Landscape Ecology of a Mighty Tropical River and its Basin*, W. Junk, Dordrecht, The Netherlands.

Valle, D., Baiser, B., Woodall, C.W. & Chazdon, R. (2014) Decomposing biodiversity data using the Latent Dirichlet Allocation model, a probabilistic multivariate statistical method. *Ecology letters*, **17**, 1591–1601.

Vedovato, L.B., Fonseca, M.G., Arai, E., Anderson, L.O. & Aragão, L.E.O.C. (2016) The extent of 2014 forest fragmentation in the Brazilian Amazon. *Regional Environmental Change*, 1–6.

Vences, M., Thomas, M., van der Meijden, A., Chiari, Y. & Vieites, D.R. (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, **2**, 1–12.

Vilhena, D.A. & Antonelli, A. (2015) A network approach for identifying and delimiting biogeographical regions. *Nature Communications*, **6**, 6848.

Wallace, A.R. (1852) On the monkeys of the Amazon. *Proceedings of the Zoological Society of London*, **20**, 107–110.

Wells, K.D. (2010) *The Ecology and Behavior of Amphibians*, University of Chicago Press, Chicago.

Wynn, A. & Heyer, W.R. (2001) Do geographically widespread species of tropical amphibians exist? An estimate of genetic relatedness within the neotropical frog Leptodactylus fuscus (Schneider, 1799) (Anura, Leptodactylidae). *Tropical Zoology*, **14**, 255–285.

Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Zeisset, I. & Beebee, T.J.C. (2008) Amphibian phylogeography: a model for understanding historical aspects of species distributions. *Heredity*, **101**, 109–119.

Zizka, A., Steege, H. Ter, Pessoa, M.D.C.R. & Antonelli, A. (2016) Finding needles in the haystack: Where to look for rare species in the American tropics. *Ecography*.

Author: Guilhem Sommeria-Klein

Title: From models to data: understanding biodiversity patterns from environmental DNA data

Supervisors: Jérôme Chave, Hélène Morlon

Abstract: Integrative patterns of biodiversity, such as the distribution of taxa abundances and the spatial turnover of taxonomic composition, have been under scrutiny from ecologists for a long time, as they offer insight into the general rules governing the assembly of organisms into ecological communities. Thank to recent progress in high-throughput DNA sequencing, these patterns can now be measured in a fast and standardized fashion through the sequencing of DNA sampled from the environment (e.g. soil or water), instead of relying on tedious fieldwork and rare naturalist expertise. They can also be measured for the whole tree of life, including the vast and previously unexplored diversity of microorganisms. Taking full advantage of this new type of data is challenging however: DNA-based surveys are indirect, and suffer as such from many potential biases; they also produce large and complex datasets compared to classical censuses. The first goal of this thesis is to investigate how statistical tools and models classically used in ecology or coming from other fields can be adapted to DNA-based data so as to better understand the assembly of ecological communities. The second goal is to apply these approaches to soil DNA data from the Amazonian forest, the Earth's most diverse land ecosystem.

Two broad types of mechanisms are classically invoked to explain the assembly of ecological communities: 'neutral' processes, i.e. the random birth, death and dispersal of organisms, and 'niche' processes, i.e. the interaction of the organisms with their environment and with each other according to their phenotype. Disentangling the relative importance of these two types of mechanisms in shaping taxonomic composition is a key ecological question, with many implications from estimating global diversity to conservation issues. In the first chapter, this question is addressed across the tree of life by applying the classical analytic tools of community ecology to soil DNA samples collected from various forest plots in French Guiana.

The second chapter focuses on the neutral aspect of community assembly. A mathematical model incorporating the key elements of neutral community assembly has been proposed by S.P. Hubbell in 2001, making it possible to infer quantitative measures of dispersal and of regional diversity from the local distribution of taxa abundances. In this chapter, the biases introduced when reconstructing the taxa abundance distribution from environmental DNA data are discussed, and their impact on the estimation of the dispersal and regional diversity parameters is quantified.

The third chapter focuses on how non-random differences in taxonomic composition across a group of samples, resulting from various community assembly processes, can be efficiently detected, represented and interpreted. A method originally designed to model the different topics emerging from a set of text documents is applied here to soil DNA data sampled along a grid over a large forest plot in French Guiana. Spatial patterns of soil microorganism diversity are successfully captured, and related to fine variations in environmental conditions across the plot.

Finally, the implications of the thesis findings are discussed. In particular, the potential of topic modelling for the modelling of DNA-based biodiversity data is stressed.

Keywords: spatial biodiversity patterns, species abundance distribution, beta-diversity, environmental DNA, metabarcoding, soil biodiversity, tropical forest, French Guiana, statistical modeling of biodiversity, neutral theory of biodiversity, topic modeling

Auteur : Guilhem Sommeria-Klein
Titre : Des modèle aux données: comprendre la structure de la biodiversité à partir de l'ADN environnemental
Directeurs de thèse : Jérôme Chave, Hélène Morlon
Lieu et date de soutenance : Université Paul Sabatier, Toulouse, le 14 septembre 2017

Résumé: La distribution de l'abondance des espèces en un site, et la similarité de la composition taxonomique d'un site à l'autre, sont deux mesures de la biodiversité ayant servi de longue date de base empirique aux écologues pour tenter d'établir les règles générales gouvernant l'assemblage des communautés d'organismes. Pour ce type de mesures intégratives, le séquençage haut-débit d'ADN prélevé dans l'environnement (« ADN environnemental ») représente une alternative récente et prometteuse aux observations naturalistes traditionnelles. Cette approche présente l'avantage d'être rapide et standardisée, et donne accès à un large éventail de taxons microbiens jusqu'alors indétectables. Toutefois, ces jeux de données de grande taille à la structure complexe sont difficiles à analyser, et le caractère indirect des observations complique leur interprétation. Le premier objectif de cette thèse est d'identifier les modèles statistiques permettant d'exploiter ce nouveau type de données pour mieux comprendre l'assemblage des communautés. Le deuxième objectif est de tester les approches retenues sur des données de biodiversité du sol en forêt amazonienne, collectées en Guyane française.

Deux grands types de processus sont invoqués pour expliquer l'assemblage des communautés d'organismes : les processus "neutres", indépendants de l'espèce considérée, que sont la naissance, la mort et la dispersion des organismes, et les processus liés à la niche écologique occupée par les organismes, c'est-à-dire les interactions avec l'environnement et entre organismes. Démêler l'importance relative de ces deux types de processus dans l'assemblage des communautés est une question fondamentale en écologie ayant de nombreuses implications, notamment pour l'estimation de la biodiversité et la conservation. Le premier chapitre aborde cette question à travers la comparaison d'échantillons d'ADN environnemental prélevés dans le sol de diverses parcelles forestières en Guyane française, via les outils classiques d'analyse statistique en écologie des communautés.

Le deuxième chapitre se concentre sur les processus neutres d'assemblages des communautés. S.P. Hubbell a proposé en 2001 un modèle décrivant ces processus de façon probabiliste, et pouvant être utilisé pour quantifier la capacité de dispersion des organismes ainsi que leur diversité à l'échelle régionale simplement à partir de la distribution d'abondance des espèces observée en un site. Dans ce chapitre, les biais liés à l'utilisation de l'ADN environnemental pour reconstituer la distribution d'abondance des espèces sont discutés, et sont quantifiés au regard de l'estimation des paramètres de dispersion et de diversité régionale.

Le troisième chapitre se concentre sur la manière dont les différences non-aléatoires de composition taxonomique entre sites échantillonnés, résultant des divers processus d'assemblage des communautés, peuvent être détectées, représentées et interprétés. Un modèle statistique conçu à l'origine pour classifier les documents à partir des thèmes qu'ils abordent est ici appliqué à des échantillons de sol prélevés selon une grille régulière au sein d'une grande parcelle forestière. La structure spatiale de la composition taxonomique des microorganismes est caractérisée avec succès et reliée aux variations fines des conditions environnementales au sein de la parcelle.

Les implications des résultats de la thèse sont enfin discutées. L'accent est mis en particulier sur le potentiel des modèles thématique (« topic models ») pour la modélisation des données de biodiversité issues de l'ADN environnemental.

Mots-clés : structure spatiale de la biodiversité, distribution d'abondance d'espèces, diversité beta, ADN environnemental, metabarcoding, biodiversité du sol, forêt tropicale, Guyane française, modélisation statistique de la biodiversité, théorie neutre de la biodiversité, topic modeling

Discipline administrative : Ecologie
Intitulé et adresse du laboratoire : Laboratoire Evolution & Diversité Biologique (EDB)
UMR 5174 (CNRS/UPS/IRD), Université Paul Sabatier, Bâtiment 4R1
118 route de Narbonne, 31062 Toulouse cedex 9, France.