ECOLOGY LETTERS

Ecology Letters, (2015) **18**: 974–982 doi: 10.1111/ele.12478

LETTER

Microbial ecosystems are dominated by specialist taxa

Mahendra Mariadassou*, Samuel Pichon, and Dieter Ebert

Universität Basel, Zoologisches Institut Vesalgasse 1, 4051 Basel, Switzerland

**Correspondence and present address: MaIAGE-INRA, UR 1404, Bât. 233, Domaine de Vilvert, 78352 Jouy-en-Josas, France. E-mail: mahendra.mariadassou@jouy.inra.fr*

Abstract

Abundance and specificity are two key characteristics of species distribution and biodiversity. Theories of species assembly aim to reproduce the empirical joint patterns of specificity and abundance, with the goal to explain patterns of biodiversity across habitats. The specialist-generalist paradigm predicts that specialists should have a local advantage over generalists and thus be more abundant. We developed a specificity index to analyse abundance–specificity relationships in microbial ecosystems. By analysing microbiota spanning 23 habitats from three very different data sets covering a wide range of sequencing depths and environmental conditions, we find that habitats are consistently dominated by specialist taxa, resulting in a strong, positive correlation between abundance and specificity. This finding is consistent over several levels of taxonomic aggregation and robust to errors in abundance measures. The relationship explains why shallow sequencing captures similar β-diversity as deep sequencing, and can be sufficient to capture the habitat-specific functions of microbial communities.

Keywords

Abundance, beta-diversity, communities, microbial ecology, microbiota, specialist taxa, specificity.

Ecology Letters (2015) **18**: 974–982

INTRODUCTION

The vast majority of microbial ecosystems, are characterised by highly skewed abundance-rank distributions: a few taxa account for the majority of individuals while most taxa are represented by only a few individuals (Connolly *et al.* 2014). In a spatial setting, the specificity, defined here as a measure for the unevenness with which a taxon occurs in different habitats, becomes important as abundance-rank distributions differ among habitats. The extremes of the specificity continuum are (1) taxa found with equal abundances in many habitats (generalists) and (2) taxa always and only found in one habitat (specialists). Extreme specialists are indicator species, with strong ecological preferences, that are specific to a given habitat (Dufrene & Legendre 1997). A long-standing view in ecology states that the differences between specialists and generalists arise from disparity in their resource utilisation: specialists have a narrow resource utilisation range and high peak performance, whereas generalists have a broader range but lower peak performance (Peers *et al.* 2012). If high performance is linked to high local competitive ability and migration does not greatly influence local population dynamics, specialists should have higher local abundance than generalists, which are found in several habitats. This pattern has for example been observed in the malaria parasite: species with narrow host range are associated with higher peak parasitaemia than species with a broad one (Garamszegi 2006).

Alternatively, specificity may be linked to abundance and the role of species in ecosystem functioning. In the case of host-associated microbial communities, hosts can recruit, either passively or actively (Kremer *et al.* 2013), their symbionts based on the functions they provide (Burke *et al.* 2011a). The impaired development of axenic animals also shows that microbiota provide essential functions to their hosts (Brummel *et al.* 2004; Rawls *et al.* 2004; Sison-Mangus *et al.* 2015). Recruited microbes are expected to thrive and be locally abundant. If furthermore, the recruited functions are costly to maintain outside of the host, they are likely specific, resulting in a correlation between local abundance and specificity.

A positive abundance–specificity relationship has important practical implications for our understanding of biodiversity. The total diversity of a landscape, γ-diversity, is traditionally partitioned between the average within-site diversity (α-diversity) and the among-site diversity (β-diversity). Several quantitative diversity indices accounting for abundance exist for community data (Graham & Fine 2008). They have in common that communities that share most or all of their abundant taxa are less β-diverse than communities that share none or a few. In the context of microbial ecology, where communities are often dominated by a few abundant taxa, we speculate that the β-diversity of distinct habitats is driven by a positive correlation between abundance and specificity of microbial taxa. As a consequence of this, the β-diversity of these communities should be easy to recover with small sampling effort. In case of microbial data, this means shallow sequencing depth by next generation sequencing approaches.

Previous studies have noted a positive relationship between abundance and occupancy. Habitat specialist taxa tend to be rare when habitats are simply defined by locality (Székely & Langenheder 2014) but become dominant when habitats are defined along an abiotic gradient (Fortunato *et al.* 2013; Logares *et al.* 2013), especially at extremes of the gradient. Nemergut *et al.* (2011) showed that detectable taxa are generally confined to single assemblage but did not report information about their local abundance. Other findings suggest that this relationship may be widespread. For instance, a relationship between specificity and abundance has recently been hinted at in butterfly gut microbiota, where communities at different developmental stages were characterised by different stage-specific abundant taxa (Hammer *et al.* 2014). The different stages are characterised by different diets, stressing that habitats can be influenced by diverse biotic and abiotic factors. The same relationship was apparent in sponge microbial communities (Moitinho-Silva *et al.* 2014; Reveillaud *et al.* 2014), but the trend was not discussed. A combination of biological (transient, dead or dormant taxa) and technical (contamination, clustering methodology) reasons lead to the increasing classification of locally abundant species as unspecific when sampling depth was increased.

Here, we investigate the abundance–specificity relationship of microbial communities, aiming to test if we can generalise it in a wide range of environments. We develop a local specificity index that separates specificity from abundance and test how it correlates with relative abundance in different microbiota habitats, defined by environment type rather than locality. We introduce a permutation test to assess the significance of this relationship. We apply our method to three data sets of microbiota data, one from zooplankton (this study), one from humans (Costello *et al.* 2009) and the last one from environmental samples (Caporaso *et al.* 2011). All data sets, chosen to reflect a wide array of environments, habitats, sampling conditions and sequencing depths, revealed strong, positive relationships between abundance and specificity. We discuss the implications of this finding for biodiversity surveys and functional metagenomics as well as mechanisms that can lead to specificity and to high local abundance of specific species.

MATERIAL AND METHODS

Data sets

Sample collection and DNA sequencing for the zooplankton data set

The *Daphnia*, sediments and freshwater samples were isolated from the Ägelsee pond, near Frauenfeld in Switzerland (47°558′ N, 8°862′ E) in June 2012. *Daphnia* are aquatic crustaceans of 0.5–5 mm body length. In our study pond, three species are sympatric: *Daphnia magna, D. pulex* and *D. longispina*. But we focused on *D. magna* as it outnumbers the others 10 to 1. *Daphnia* samples were collected using plankton net with 200 μm mesh size and kept in pond water during transport to the laboratory. Within 5 h, *D. magna* specimens were separated from other plankton species, washed in autoclaved medium and individually frozen at −20 °C in buffer for later DNA extraction. Samples from the top layer (5 cm) of the pond sediments were collected with a hand-held dreg, drained of standing water and stored 24 h at 4 °C in the dark before DNA extraction. Water samples were collected at the *Daphnia* sampling site and microbes successively filtered through 52-, 19- and 7-μm filters. The additional 3-, 1- and 0.2-μm filters were used for DNA extraction. Total genomic DNA was extracted using the CTAB method for the *Daphnia* and water samples and the PowerSoil kit (MoBio, Carlsbad, CA, USA), following the manufacturer's instructions, for the sediment samples.

The V3–V5 region of the 16S rRNA gene (ca. 640 nucleotides) was PCR amplified using the following forward and reverse primers: 5′-ACACGGYCCARACTCCTAC-3′ (positions 327–345) and 5′-GTGGWTTAATTCGAWGCAA-3′ (positions 951–969). Amplicons from the different samples were pooled at equimolar ratio for multiplexing. Amplicon libraries were sequenced on a GS FLX instrument using LibL Titanium chemistry (454 Life Sciences, Roche) at Microsynth (Balgach, Switzerland).

Bioinformatics pipeline for zooplankton data set

QIIME 1.6 pipeline (Caporaso *et al.* 2010a) was used for bioinformatics analyses. Reads were filtered based on their length (> 500, < 580) and average Phred score (> 30). They were then chimera filtered using *de novo* chimera detection and clustered into operational taxonomic units (OTUs) at 97% sequence identity level using USEARCH (Edgar 2010) with QIIME default parameters. Taxonomic assignment was done using the RDP classifier (Wang *et al.* 2007) with the Greengenes database, version 12_10 (McDonald *et al.* 2011). The most abundant sequence of OTU cluster was chosen as its representative and sequences were aligned with PyNAST (Caporaso *et al.* 2010b). A phylogenetic tree was then built using FastTree (Price *et al.* 2010) for computation of UniFrac distances (Lozupone *et al.* 2007). OTUs represented by a single read were filtered out from the count table.

Human microbiota data set

The human microbiome data set is described in and was collected by Costello *et al.* (2009): filtered reads were downloaded from the European Nucleotide Archive (www.ebi.ac.uk/ena/data/view/ERA000159) and OTU picking was performed using the analysis pipeline detailed in Costello *et al.* (2009). Singletons were filtered out.

Global patterns data set

The 'Global Patterns' data set of environmental samples was obtained from Caporaso *et al.* (2011) and is directly available as OTU count tables from the R phyloseq package (McMurdie & Holmes 2013). Singletons were filtered out.

Specificity index and permutation test

Here, we consider a quantitative measure for the specificity of a species (or OTU) in a focal habitat, based on a reinterpretation of the indicator values of Dufrene & Legendre (1997). A high value indicates that this species is found predominantly in this focal habitat. We consider a microbiota count table $M = (a_{ij})$ where a_{ij} is the count, or relative abundance, of species i in sample j. We use OTUs as proxy for species. The microbiota samples originate from H different habitats (e.g. different locations, different hosts) and there are S^h samples from habitat h. We note S_i^h the number of samples from habitat h where species i is present. The local specificity index Λ_i^h of species i in habitat h is defined as:

$$\Lambda_i^h = A_i^h \times B_i^h \tag{1}$$

where

$$A_i^h = \frac{S_i^h}{S^h}, \quad B_i^h = \frac{\langle a_i \rangle^h}{\sum\limits_{h=1}^{H} \langle a_i \rangle^h} \quad \text{and} \quad \langle a_i \rangle^h = \frac{\sum\limits_{j=1}^{S^h} a_{ij}}{S^h} \tag{2}$$

A_i^h is the prevalence of species i in habitat h, i.e. the fraction of samples from habitat h where species i was found. In case of limited sampling, it should be understood as discoverability, especially for rare species. $\langle a_i \rangle^h$ is the average local abundance of species i in habitat h and B_i^h is the fraction of the share of habitat h in the total population of species i, i.e. summed across habitats. Λ_i^h ranges from 0 to 1. A value of 0 means that the species is never found in habitat h, whereas a value of 1 means that the species is always and only found in that habitat, i.e. the species is a perfect indicator of that habitat. Unlike strict definitions of specificity, that use only presence/absence information, our index uses the full abundance spectrum and is therefore robust to contamination and differences in sequencing depths. Λ_i^h is the same as INDVAL but for one change: INDVAL keeps only the maximum local specificity across habitats ($\mathrm{INDVAL(i)} = \max_h \Lambda_i^h$) to have one value per species, whereas we keep all values to have one value per pair (species, habitat).

Since B_i^h requires the computation of the total count of species i, the abundances must be on the same scale in different samples for the index to be meaningful. We use a double normalisation of abundances to avoid the undue effect of individual rich samples and sample rich habitats. Abundances are first transformed to relative abundance (sample-level normalisation) and then averaged within each habitat (habitat-level normalisation) to produce $\langle a_i \rangle^h$.

The variability of Λ_i^h is assessed by stratified bootstrap. A new data set is created from the original by sampling communities with replacement within each habitat (the strata). The specificity index is computed on this data set and the process is repeated 999 times to estimate the variability of Λ_i^h. The significance of Λ_i^h is also assessed by bootstrap resampling: original communities are resampled with replacement and randomly assigned to habitats (habitat permutation). The specificity index is computed on each bootstrapped data set and the process is repeated 999 times. These values represent the distribution of Λ_i^h under the null hypothesis of no association between local abundances and habitats. The traditional way of assessing the significance of Λ_i^h would be to compare it to this distribution. We are however interested in the significance of high specificities and therefore compare Λ_i^h to the null distribution of $\max_h \Lambda_i^h$ rather than Λ_i^h. In other words, the observed specificity is compared to the *maximum* randomised specificity across all habitats. This corrects for the multiplicity of habitats and means that a species is significantly specific only if more specific to a habitat than to any other one. Local specificities, permutation tests and standard error computations were performed using custom R code and the R phyloseq package (McMurdie & Holmes 2013).

RESULTS

Zooplankton

This data set includes 32 microbiota samples from one pond in Switzerland, but corresponding to three different habitats: microbiota associated with the planktonic crustacean *D. magna* (11 samples), microbes in the free water (15 samples) and microbes from the sediment surface of the pond (6 samples). These three habitats are interconnected: *D. magna* filter water for uptake of food, water and sediments are in close contact and *Daphnia* browse over the sediment surface to enrich their food (Horton *et al.* 1979; Ebert 2005). Furthermore, *Daphnia* resting eggs (ephippia) diapause in the sediments.

We used the zooplankton data set to derive abundance–specificity curves (Fig. 1) using our local specificity index. We find that on average the more abundant a taxon is in a habitat, the more specific it is to this habitat. Furthermore, the positive abundance–specificity relationship observed is significantly steeper than also positive random relationship (grey dots and dark grey curves in the three panels of Fig. 1). The latter arises because universal species with heterogeneous abundances across samples have a skewed abundance distribution with many small values and a few high ones. Chance grouping of high values in the same habitat during the randomisation scheme shifts the distribution of randomised specificities upwards compared to species with the same overall abundance but a more even distribution across samples. The upward trend of the specificity curve is nevertheless significantly steeper in the non-randomised than in the randomised data sets (Fig. 1): the most locally abundant species are highly specific to their habitat.

In parallel, we studied the diversity of samples using either Bray–Curtis distances, which account for differences in terms of species composition and abundances, or weighted UniFrac (wUF) distances, which also account for phylogenetic relatedness between species, combined with multidimensional scaling. The results (Fig. 2) show that the samples cluster neatly by habitats for the Bray–Curtis distances (Fig. 2, left) but in a less pronounced way for wUF distances (Fig. 2, right). The first axis distinguishes sediments from water and *Daphnia*. *Daphnia* and water samples then form a *Daphnia*-Water gradient along the second axis.

We evaluated the robustness of our findings with respect to copy number variation in SSU-rRNA genes across bacterial taxa (Vetrovsky & Baldrian 2013). Copy numbers of this gene vary from 1 to 15 in sequenced microbes and thus may influence the calculation of relative frequencies. Vetrovsky & Baldrian (2013) provided a copy number prediction scheme based on the average copy number at a given taxonomic level and estimated the average error to be 38% when predictions are made at the phylum level. We corrected observed species counts using either (1) average phylum-level (class in case of the abundant Proteobacteria) copy number or a (2) random copy number drawn from a truncated Gaussian, approximating the copy number distribution observed within each phylum (class in case of the abundant Proteobacteria). Our results did not change (Fig. S1). A further source of bias of amplicon-based microbiota studies is the classification of species as OTUs, which is built on a similarity threshold (typically 97% similarity). This threshold impacts specificity and abundance of microbial taxa (Youngblut *et al.* 2013): all things remain unchanged, higher thresholds lead to more OTUs, with lower abundances and higher specificities. However, using thresholds ranging from 91 to 99% identity does not change our finding of a positive relationship between abundance and specificity (Fig. S2).

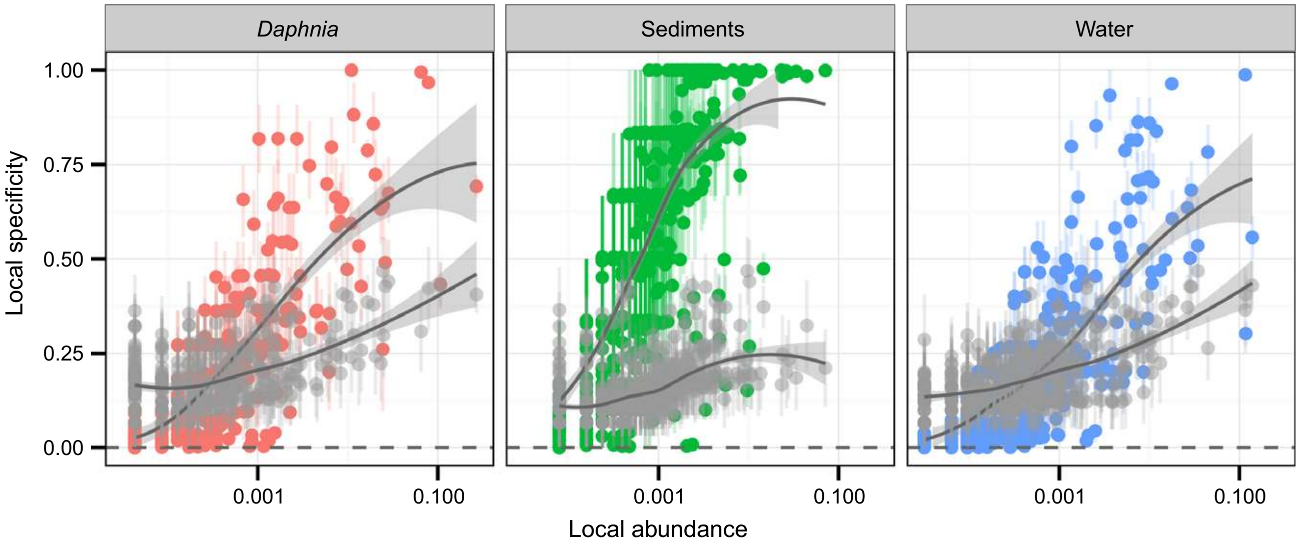

Figure 1 Positive relationship between abundance and specificity in the zooplankton data set. Coloured dots and error bars correspond to observed specificity values and interquartile variability. Grey dots (resp. error bars) correspond to expected local specificity (interquartile range) under the null distribution where samples are randomly assigned to body habitats (see Permutation Test for full details). Dark grey curves are loess regressions of the original and randomised specificity values against abundance in the habitat, with associated confidence bands.

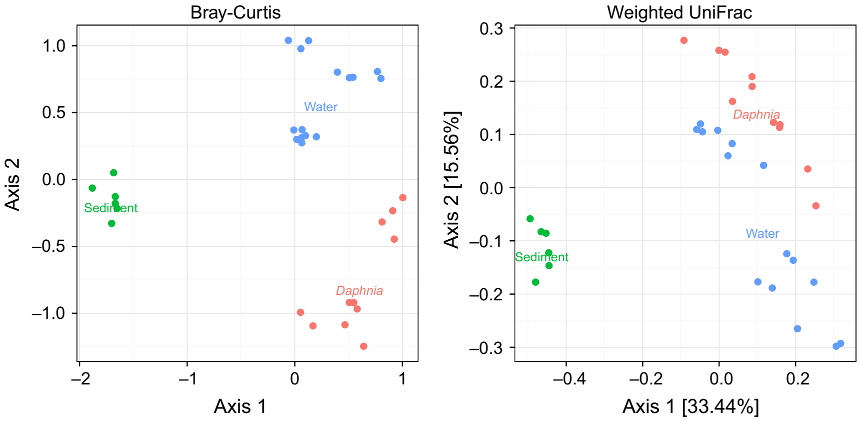

Figure 2 Multidimensional scaling ordination plot of Bray–Curtis (left) and weighted UniFrac (right) distances of freshwater samples. Water and *Daphnia* samples are well separated from sediments and also distinct from each other using Bray–Curtis distances.

Human microbiome data set

The human body is host to diverse bacterial communities whose collective population largely outnumbers human cells (Costello *et al*. 2009). We analyse the publicly available data set built by Costello *et al*. (2009), who sampled microbial communities across 26 body sites on seven subjects at two time points. Sampling sites were grouped by body habitat as in the original study (Costello *et al*. 2009): Skin, External Auditory Canal (EAC), Gut, Oral Cavity, Hair, Nostril and Hair. Skin is the most diverse habitat with 18 sites ranging from sole of foot to forehead, while other environments included much less sampling sites (minimum 2). This data set is interesting for our purpose as the habitats are very diverse but in close proximity and therefore allow for exchange of bacteria. Furthermore, the human microbiota has been extensively studied and is well characterised. Previous studies show that body habitats differ in their overall composition, based on ordination plots of UniFrac distances (Costello *et al*. 2009).

All body habitats exhibit a positive and significant abundance–specificity relationship, with the exception of the skin (Fig. 3). Skin is considered here as a single body habitat as in the original study but really encompasses a variety of habitats with different environmental variables. For example, parts of the skin such as armpits, soles of feet, hand palm and index finger differ in key environmental variables such as humidity, temperature and light exposure. An alternative grouping of skin sampling sites that respect body symmetry (see Supporting information for details), splits skin into six habitats. Using this classification, we find a stronger abundance–specificity relationship (Fig. S3). The finding of the abundance–specificity relationship is also stable across time: the analysis performed on each time point individually shows the same positive trend (Fig. S4) and the most abundant taxa are conserved across time (Fig. S5).

Global patterns

The global patterns (GP) data set consist of 25 environmental samples gathered by Caporaso *et al.* (2011) with an average depth of about a million reads per sample. The samples are distributed across eight different habitats, with two to four replicates per habitat: human faeces (n = 3), hand palm skin (n = 3) and dorsal tongue surface (n = 3), faecal samples from human twins (n = 2), soil (n = 3), freshwater and freshwater sediments (n = 5), ocean water (n = 3) and marine sediments (n = 3). We left out the artificially assembled mock communities as they are unlikely to capture natural abundance–specificity relationships.

Like in the previous data sets, all habitats exhibit a strong abundance–specificity relationship: abundant species have on average a high local specificity. In this data set we note that many low frequency (< 1e-4) taxa have high local specificity (Fig. 4), but many more have low specificity. A rarefaction analysis of the samples (Fig. S6) confirms that the relationship is preserved at much shallower sampling depths (5000 reads per samples). It also shows that species with high local specificity (> 0.9) are easy to capture: 5000 reads capture on average 40% of the specific species accounting for 96% per cent of the specific population.

DISCUSSION

The specificity of species from microbial communities to a given habitat type is positively correlated with their local abundance. This relationship is robust to various technical and biological aspects of microbiota assessment and analysis, such as differences in 16S rRNA gene copy numbers, OTU definition threshold and sampling depth. The diverse data sets used to test this relationship reveal that even when habitats have frequent opportunities for microbial exchange, microbial compositions remain distinct. The rich global pattern data set further shows that abundant taxa are highly specific to their habitat even though habitat-specific taxa are found everywhere across the abundance scale. Our communities differ clearly by habitat in terms of specific and abundant species, even when those habitats interact with each other through known overlaps and migration from a regional pool of taxa, like in the zooplankton data set. Local microbiota may share some taxa, like the planktonic freshwater bacteria from the genus *Limnohabitans* that accounts on average for 15% of both *Daphnia* and water samples but we found only one instance of shared abundant taxa. The fraction of shared abundant taxa varies with the differentiation between the

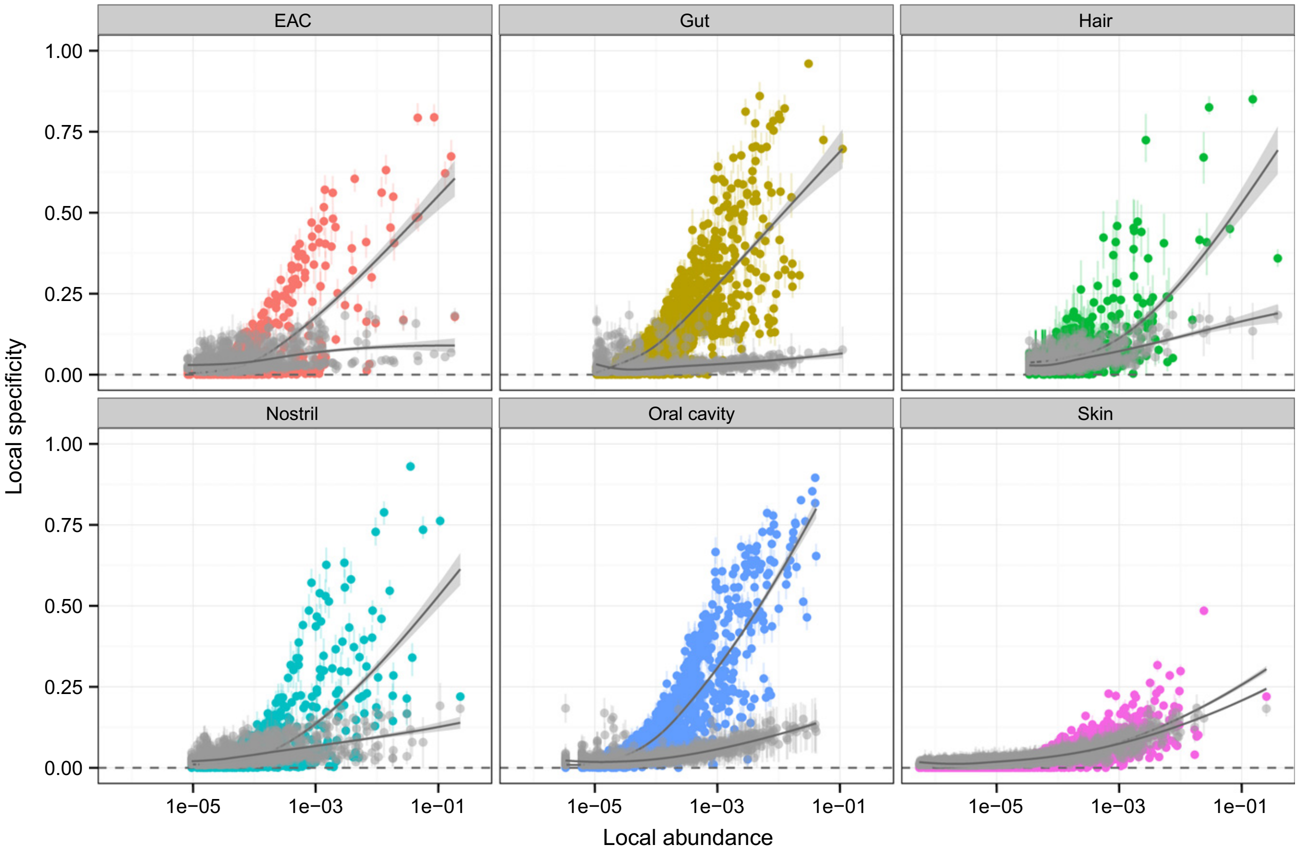

Figure 3 Positive relationship between local abundance and specificity in the human microbiome (Costello *et al.* 2009). Details of the legend are identical to Figure 1. Note that 'skin' encompasses 18 body sites corresponding to very different environmental variables and does not form a coherent habitat.

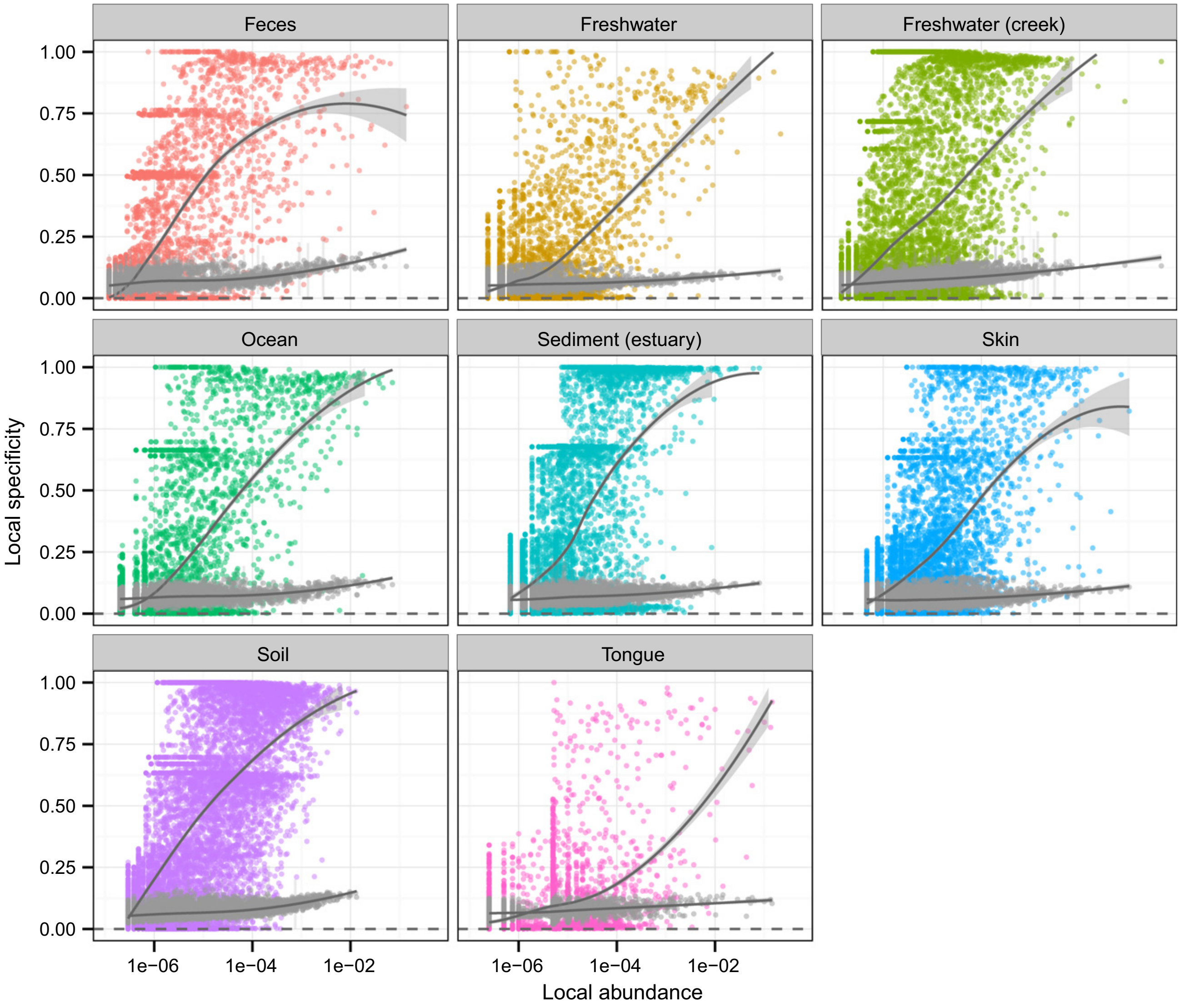

Figure 4 Positive relationship between local abundance–specificity in environmental habitats from the global pattern study (Caporaso *et al.* 2011). Details of the legend are identical to Fig. 1 but because of the large number of taxa, error bars were omitted. Unlike the Human microbiome data set, skin samples are limited to hand palms.

habitats being compared. In the absence of differentiation between habitats, the Neutral Community Model (NCM) predicts that taxa with high mean abundance should have high occupancy and variance inversely proportional to the migration rate (Sloan *et al.* 2006). This abundance-occupancy trend has been reported in Nemergut *et al.* (2011) but was not observed in another large-scale soil study (Barberán *et al.* 2012). However, if habitats are well differentiated, as in Barberán *et al.* (2012) or this study, NCM does not rule out a mix of (1) habitat specialists with high abundances and occupancies in their habitat but low ones in other habitats and (2) habitat generalists with low abundances and occupancies in all habitats. In this case, differentiation leads to habitat-exclusive abundant taxa (Gravel *et al.* 2006).

The monotonic increase in the abundance–specificity curves at the habitat level suggests that habitat filtering plays an important role in shaping local communities and that locally abundant taxa are likely to be specific to their habitat. Our finding means that β-diversity is driven by taxa that are both specific and locally abundant. As a consequence, β-diversity can be estimated without a large sampling effort. It explains why wUF and other weighted (semi)metrics (Graham & Fine 2008) or quantitative β-diversity measures are unexpectedly good at separating different habitats even at relatively low sampling depth (~ 2000 reads per sample) and why low-depth sampling captures the same relationship among samples as does deep sampling, as noted by Caporaso *et al.* (2011) on the GP data set. Our analysis also revealed that taxa specific to a habitat do not need to be strongly phylogenetically differentiated. Figure 2 shows that water and *Daphnia* specific taxa are phylogenetically close, but taxonomically different; the difference may lie in their ecological function, as was previously suggested in this setting (Qi *et al.* 2009) and in other studies (Burke *et al.* 2011a; Jiang *et al.* 2012).

Previous studies provide examples of abundance–specificity relationships (although not formally analysed) and suggest diverse mechanisms through which this relationship may come about. For example, Barberán *et al.* (2012) report that specialists, especially in desert and Antarctic soils, are more abundant than predicted based on their persistence across habitats. Species with active hydrocarbon degradation capabilities are not detectable in sediments, but become very abundant (6–18%) after oil spill (Mason *et al.* 2014). Habitat specialists are locally abundant at the extremes of salinity gradients (Logares *et al.* 2013) and coastal marine environment are reproducibly dominated by taxa with limited distributions (Fortunato *et al.* 2013). In these examples, specificity may be achieved by high growth rates of specialists. Host-associated communities also provide interesting examples: sponge species share most of their OTUs but harbour species-specific members that make up to 30% of the communities (Reveillaud *et al.* 2014). Moitinho-Silva *et al.* (2014) suggest that ‘dominant OTUs do account for sponge communities’. Likewise resident communities of *Daphnia* (this study) and algal surfaces (Burke *et al.* 2011b) are substantially made up of bacteria that are rare in the surrounding waters. In those examples, specificity may result from species sorting along a nutrient gradient but specificity can also result from active selection from the host as in the squid-*Vibrio* system (Kremer *et al.* 2013). Finally, chemico-physical barriers may limit competition, prevent invasion and therefore favour resident taxa. Jones & McMahon (2009) show that immigration of atmospheric taxa into lakes that are otherwise closed systems is essentially erased by habitat filtering: the studied lakes are dominated by abundant and specific species. Similarly, bile production is thought to protect the human gut from foreign microbe species. Alterations caused by liver cirrhosis can lead to invasion and displacement of up to 40% of the resident community by a few commensal oral species that are absent from healthy communities (Qin *et al.* 2014).

Community assembly

Community assembly involves four distinct kinds of processes: selection, drift, speciation and migration (Vellend 2010). They all have been shown to play a role in microbial community assembly (Costello *et al.* 2012). Our findings suggest that even when habitats are in close spatial proximity and species migration is possible, selection plays a comparatively central role in determining the composition of local communities, as suggested by the Baas-Becking hypothesis, and may mitigate the effects of migration (Gravel *et al.* 2006; Jones & McMahon 2009). It is also well known that assemblages are typically dominated by a few abundant taxa while most taxa are rare and that neutral assembly can lead to such a skew (Sloan *et al.* 2006). Our observations may be a consequence of (1) habitat differentiation, (2) selection for specialist taxa across habitats (for example in the form of habitat filtering or high growth rate in the favoured habitat of a taxon), and (3) neutral assembly processes within habitat. Selection implies that habitat-level relative abundances would be skewed in favour of specialists that would then be abundant through the neutral assembly of individual communities. The skew should increase with increased migration and increased differentiation between habitats, as observed in this study.

Deterministic components of assembly processes have been suggested elsewhere (Vellend 2010; Stegen *et al.* 2012). Connolly *et al.* (2014) deduced, from studying abundance-rank curves of over 1000 communities from the marine biosphere, that lognormal abundances are a better fit than the neutral Poisson-gamma distribution. They suggest differences between species (e.g. in terms of growth rates, niche sizes, etc.) as mechanisms leading to lognormal abundances. Recent work (Harris *et al.* 2014) in statistical ecology looking at the fit of neutral models (Hubbell 2001) to human gut microbiota reach similar conclusions: neutrality is rejected from abundance data at all but low taxonomic ranks (genus). Overall, habitat filtering selects specialist taxa, probably based on markedly different metabolic and functional roles (Fierer *et al.* 2007; Philippot *et al.* 2010) and among those, some taxa with selective advantage become dominant in local communities.

Limits and potential confounding factors

Quantitative metagenomics is a powerful approach to study microbial communities because it enables us to obtain data sets with very deep sampling in short time periods and high spatial resolution. However, it is subject to a variety of biases that may influence the inference obtained from the analysis. Standard methodology assumes that each microbial taxon has the same number of 16S rDNA genes, although it is known that species differ from 1 to 15 copies of this gene (Vetrovsky & Baldrian 2013). Our analyses suggest that our results do not suffer from a serious bias caused by copy number variation. Assumptions about the definition of the OTU as a surrogate for a species definition may also influence the interpretation of microbiota studies because it affects both the species classification and their abundance. Again, our results are robust with regard to this assumption.

A further bias arises from the complex interplay between niche and habitat. Niche is defined by a volume in a space of mostly unknown biotic and abiotic environmental variables (Hutchinson 1957). Habitats may be comprised of a single niche, or may include multiple niches. For example, the human skin is diverse and covers many environmental variables. Not surprisingly, the skin habitat does not show a strong abundance–specificity relationship. After a custom classification of skin sampling sites based on hierarchical clustering, body geography and symmetry, we recovered more narrowly defined habitats and stronger abundance–specificity relationships. In conducting a custom classification we took care to avoid circularity and to artificially increase the abundance–specificity trend by clustering the samples using only presence/absence data, not abundance or prevalence data. Additionally, the skin habitat in the GP data set consists only of hand palms and exhibits a strong association from the start. The same observation applies for the butterfly gut microbiota study of Hammer *et al.* (2014). The gut at different developmental stages corresponds to different diets. It is dominated by different taxa, suggesting that host diet is a key environmental variable for niches of gut microbiota (see also Delsuc *et al.* 2014).

Our local specificity index captures relative specificity rather than endemism. Our double normalisation, standard in microbiota studies, assumes (1) that each habitat has a carrying capacity shared by all species and (2) that all habitats contribute equally to the regional pool of taxa. When these assumptions are violated we need to be careful with the interpretation of the results. For example, in a continent-island scenario, a species can be locally very specific to a small habitat (an island) and yet be mostly found in the much larger habitat (the continent) (see Supporting information). Our index therefore works only for relative specificity rather than endemism. An important consequence is that the abundance–specificity relationship is influenced by all habitats included in the study: it is stronger if habitats are very distinct, like in the GPs, and weaker but still significant for more similar habitats, like in the Human Microbiome.

Finally, our method counts all zeroes as true absence when computing prevalence in a given habitat. Replacing zeroes by probability of absence, estimated from zero-inflated distributions such as Gaussian or Negative Binomial (McMurdie & Holmes 2014) should alleviate this problem. This limitation induces a downward bias of specificity values for rare species but does not weaken the abundance–specificity relationship, as confirmed by the rarefaction analysis of the GP data set (Fig. S6). Furthermore, the high depths of the GP data set almost guarantee discovery of taxa with even low frequency (> 1e-5) and confirm that the relationship is not an artefact of rare taxa having low specificity simply because of discoverability issues.

CONCLUSION

The main finding of our analysis is the on average high specificity of abundant microbe species in microbiota samples from well differentiated habitats. It means that (1) shallow sampling is sufficient to recover compositional differences and (2) locally abundant species, many of which are habitat specific, are likely to perform habitat-specific ecological functions. This finding has implications for other studies. By reducing the sampling effort per site, it will be possible to increase the replication effort and the number of habitats studied. For example, we suggest that in studies of β-diversity surveys, one should aim for moderate sequencing depth, but for a large sampling width. On the other hand, the existence of specific species across the abundance scale means that α-diversity surveys capture most specific species with shallow sampling, but cannot spare deep sequencing as some species are both rare and specific.

Furthermore, our study suggests that abundant species are on average good predictors of environmental conditions and that the search for microbial indicators should start with abundant species. An exciting extension in this direction would be possible if transcriptional activity is positively correlated with abundance or skewed towards abundant taxa (Moitinho-Silva *et al.* 2014). In this case, shallow sampling might be sufficient to identify the genes of key biological processes provided by microbial ecosystems (Burke *et al.* 2011a; Jiang *et al.* 2012; Qin *et al.* 2014).

ACKNOWLEDGEMENTS

We thank Jürgen Hottinger for help in the laboratory, members of the microbiota study group at the Zoological Institute for fruitful discussions and Aurélie Cotillard for feedback on earlier versions of the manuscript. We are grateful to the INRA MIGALE bioinformatics platform for providing computational resources. The work was supported by an ERC Advanced Investigator grant to DE. We also thank three anonymous reviewers and the editor James P. Grover for comments that greatly improved the first version of this manuscript.

AUTHORSHIP

MM, SP and DE designed the study. SP collected the data. MM performed the modelling work and analysed the data. MM wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

REFERENCES

Barberán, A., Bates, S.T., Casamayor, E.O. & Fierer, N. (2012). Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.*, 6, 343–351.

Brummel, T., Ching, A., Seroude, L., Simon, A.F. & Benzer, S. (2004). Drosophila lifespan enhancement by exogenous bacteria. *Proc. Natl. Acad. Sci. USA*, 101, 12974–12979.

Burke, C., Steinberg, P., Rusch, D., Kjelleberg, S. & Thomas, T. (2011a). Bacterial community assembly based on functional genes rather than species. *Proc. Natl. Acad. Sci. USA*, 108, 14288–14293.

Burke, C., Thomas, T., Lewis, M., Steinberg, P. & Kjelleberg, S. (2011b). Composition, uniqueness and variability of the epiphytic bacterial community of the green alga Ulva australis. *ISME J.*, 5, 590–600.

Caporaso, G.J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. *et al.* (2010a). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods*, 7, 335–336.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L. & Knight, R. (2010b). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26, 266–267.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J. *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc. Natl. Acad. Sci. USA*, 108, 4516–4522.

Connolly, S.R., MacNeil, M.A., Caley, M.J., Knowlton, N., Cripps, E., Hisano, M. *et al.* (2014). Commonness and rarity in the marine biosphere. *Proc. Natl. Acad. Sci. USA*, 111, 8524–8529.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I. & Knight, R. (2009). Bacterial community variation in human body habitats across space and time. *Science*, 326, 1694–1697.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J.M. & Relman, D.A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science*, 336, 1255–1262.

Delsuc, F., Metcalf, J.L., Wegener Parfrey, L., Song, S.J., González, A. & Knight, R. (2014). Convergence of gut microbiomes in myrmecophagous mammals. *Mol. Ecol.*, 23, 1301–1317.

Dufrene, M. & Legendre, P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecol. Monogr.*, 67, 345–366.

Ebert, D. (2005). Ecology, Epidemiology, and Evolution of Parasitism in Daphnia. National Center for Biotechnology Information (US). http://www.ncbi.nlm.nih.gov/books/NBK2036/

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.

Fierer, N., Bradford, M.A. & Jackson, R.B. (2007). Toward an ecological classification of soil bacteria. *Ecology*, 88, 1354–1364.

Fortunato, C.S., Eiler, A., Herfort, L., Needoba, J.A., Peterson, T.D. & Crump, B.C. (2013). Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *ISME J.*, 7, 1899–1911.

Garamszegi, L.Z. (2006). The evolution of virulence and host specialization in malaria parasites of primates. *Ecol. Lett.*, 9, 933–940.

Graham, C.H. & Fine, P.V.A. (2008). Phylogenetic beta diversity: linking ecological and evolutionary processes across space in time. *Ecol. Lett.*, 11, 1265–1277.

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

Hammer, T.J., McMillan, W.O. & Fierer, N. (2014). Metamorphosis of a butterfly-associated bacterial community. *PLoS ONE*, 9, e86995.

Harris, K., Parsons, T.L., Ijaz, U.Z., Lahti, L., Holmes, I. & Quince, C. (2014). Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process. arXiv:1410:4038 [q-bio.PE].

Horton, P.A., Rowan, M., Webster, K.E. & Peters, R.H. (1979). Browsing and grazing by cladoceran filter feeders. *Can. J. Zool.*, 57, 206–212.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton.

Hutchinson, G.E. (1957). Concluding remarks. *Cold Spring Harb. Symp. Quant. Biol.*, 22, 415–427.

Jiang, X., Langille, M.G.I., Neches, R.Y., Elliot, M., Levin, S.A., Eisen, J.A. *et al.* (2012). Functional biogeography of ocean microbes revealed through non-negative matrix factorization. *PLoS ONE*, 7, e43866.

Jones, S.E. & McMahon, K.D. (2009). Species-sorting may explain an apparent minimal effect of immigration on freshwater bacterial community dynamics. *Environ. Microbiol.*, 11, 905–913.

Kremer, N., Philipp, E.E.R., Carpentier, M.-C., Brennan, C.A., Kraemer, L., Altura, M.A. *et al.* (2013). Initial symbiont contact orchestrates host-organ-wide transcriptional changes that prime tissue colonization. *Cell Host Microbe*, 14, 183–194.

Logares, R., Lindström, E.S., Langenheder, S., Logue, J.B., Paterson, H., Laybourn-Parry, J. *et al.* (2013). Biogeography of bacterial communities exposed to progressive long-term environmental change. *ISME J.*, 7, 937–948.

Lozupone, C.A., Hamady, M., Kelley, S.T. & Knight, R. (2007). Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. *Appl. Environ. Microbiol.*, 73, 1576–1585.

Mason, O.U., Scott, N.M., Gonzalez, A., Robbins-Pianka, A., Bælum, J., Kimbrel, J. *et al.* (2014). Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill. *ISME J.*, 8, 1464–1475.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A. *et al.* (2011). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.*, 6, 610–618.

McMurdie, P.J. & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*, 8, e61217.

McMurdie, P.J. & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.*, 10, e1003531.

Moitinho-Silva, L., Bayer, K., Cannistraci, C.V., Giles, E.C., Ryu, T., Seridi, L. *et al.* (2014). Specificity and transcriptional activity of microbiota associated with low and high microbial abundance sponges from the Red Sea. *Mol. Ecol.*, 23, 1348–1363.

Nemergut, D.R., Costello, E.K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S.K. *et al.* (2011). Global patterns in the biogeography of bacterial taxa. *Environ. Microbiol.*, 13, 135–144.

Peers, M.J.L., Thornton, D.H. & Murray, D.L. (2012). Reconsidering the specialist-generalist paradigm in niche breadth dynamics: resource gradient selection by Canada lynx and bobcat. *PLoS ONE*, 7, e51488.

Philippot, L., Andersson, S.G.E., Battin, T.J., Prosser, J.I., Schimel, J.P., Whitman, W.B. *et al.* (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev. Microbiol.*, 8, 523–529.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2010). FastTree 2: approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5, e9490.

Qi, W., Nong, G., Preston, J.F., Ben-Ami, F. & Ebert, D. (2009). Comparative metagenomics of Daphnia symbionts. *BMC Genom.*, 10, 172.

Qin, N., Yang, F., Li, A., Prifti, E., Chen, Y., Shao, L. *et al.* (2014). Alterations of the human gut microbiome in liver cirrhosis. *Nature*, 513, 59–64.

Rawls, J.F., Samuel, B.S. & Gordon, J.I. (2004). Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microbiota. *Proc. Natl. Acad. Sci. USA*, 101, 4596–4601.

Reveillaud, J., Maignien, L., Eren, A.M., Huber, J.A., Apprill, A., Sogin, M.L. *et al.* (2014). Host-specificity among abundant and rare taxa in the sponge microbiome. *ISME J.*, 8, 1198–1209.

Sison-Mangus, M.P., Mushegian, A.A. & Ebert, D. (2015). Water fleas require microbiota for survival, growth and reproduction. *ISME J.*, 9, 59–67.

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. *Environ. Microbiol.*, 8, 732–740.

Stegen, J.C., Lin, X., Konopka, A.E. & Fredrickson, J.K. (2012). Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J.*, 6, 1653–1664.

Székely, A.J. & Langenheder, S. (2014). The importance of species sorting differs between habitat generalists and specialists in bacterial communities. *FEMS Microbiol. Ecol.*, 87, 102–112.

Vellend, M. (2010). Conceptual synthesis in community ecology. *Q. Rev. Biol.*, 85, 183–206.

Vetrovsky, T. & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE*, 8, e57923.

Wang, Q., Garrity, G.M., Tiedje, J.M. & Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.*, 73, 5261–5267.

Youngblut, N.D., Shade, A., Read, J.S., McMahon, K.D. & Whitaker, R.J. (2013). Lineage-specific responses of microbial communities to environmental change. *Appl. Environ. Microbiol.*, 79, 39–47.

SUPPORTING INFORMATION

Additional Supporting Information may be downloaded via the online version of this article at Wiley Online Library (www.ecologyletters.com).

Editor, James P. Grover
Manuscript received 5 March 2015
First decision made 14 April 2015
Second decision made 19 June 2015
Manuscript accepted 29 June 2015