npg

ORIGINAL ARTICLE

Using network analysis to explore co-occurrence patterns in soil microbial communities

Albert Barberán[1], Scott T Bates[2], Emilio O Casamayor[1] and Noah Fierer[2,3]

[1]*Department of Continental Ecology–Biogeodynamics and Biodiversity Group, Centre d'Estudis Avançats de Blanes (CEAB-CSIC), Blanes, Spain;* [2]*Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO, USA and* [3]*Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA*

Exploring large environmental datasets generated by high-throughput DNA sequencing technologies requires new analytical approaches to move beyond the basic inventory descriptions of the composition and diversity of natural microbial communities. In order to investigate potential interactions between microbial taxa, network analysis of significant taxon co-occurrence patterns may help to decipher the structure of complex microbial communities across spatial or temporal gradients. Here, we calculated associations between microbial taxa and applied network analysis approaches to a 16S rRNA gene barcoded pyrosequencing dataset containing >160 000 bacterial and archaeal sequences from 151 soil samples from a broad range of ecosystem types. We described the topology of the resulting network and defined operational taxonomic unit categories based on abundance and occupancy (that is, habitat generalists and habitat specialists). Co-occurrence patterns were readily revealed, including general non-random association, common life history strategies at broad taxonomic levels and unexpected relationships between community members. Overall, we demonstrated the potential of exploring inter-taxa correlations to gain a more integrated understanding of microbial community structure and the ecological rules guiding community assembly.

Introduction

Studies of complex microbial communities have advanced considerably in recent years , in part, due to methodological advances such as high-throughput DNA sequencing technologies that yield detailed information on the composition of microbial communities (Sogin *et al.*, 2006). The sequence data are typically derived from sequencing a portion of the small-subunit rRNA gene (Pace, 1997) and a wide variety of techniques can be applied to the analysis of the sequence data in order to describe the composition of microbial communities, their diversity and how communities may change across space, time, or experimental treatments. However, most of the analytical techniques focus on single properties of the communities being studied. For example, studies describing and comparing the structure of microbial communities often focus on the total

numbers of taxa or unique lineages found in individual samples (that is, alpha-diversity), the relative abundances of individual taxa or lineages and the extent of phylogenetic or taxonomic overlap between communities or community categories (that is, beta-diversity). Alpha-diversity measures (for example, richness and coverage estimators, rarefaction curves) yield estimates of microbial diversity and its limits in different environments (Hughes *et al.*, 2001; Curtis *et al.*, 2002; Sogin *et al.*, 2006). Likewise, multivariate statistical techniques such as clustering and ordination have allowed microbial ecologists to describe beta-diversity patterns, revealing how biotic and abiotic variables control microbial community composition. For example, analyses of beta-diversity patterns have revealed how microbial communities are structured across a wide range of natural habitats (Lozupone and Knight, 2007; Auguet *et al.*, 2010; Barberán and Casamayor, 2010, 2011), the spatial and temporal variability of microbial communities on and in the human body (Fierer *et al.*, 2008; Costello *et al.*, 2009), and the factors structuring soil bacterial communities (Lauber *et al.*, 2009). In contrast, there has been far less attention focused on using sequence data to explore the direct or indirect interactions between microbial taxa

344

coexisting in environmental samples. Documenting these interactions between taxa (that is, co-occurrence patterns) across complex and diverse communities may help to ascertain the functional roles or environmental niches occupied by uncultured microorganisms (Ruan et al., 2006; Fuhrman and Steele, 2008; Chaffron et al., 2010). With the ever-increasing accumulation of sequence data from microbial communities, we now have the challenge to begin exploring these interactions, and to extend community analyses beyond the exploration of alpha- and beta-diversity patterns that represent the bulk of most sequence-based microbial community analyses.

Network analysis tools and network thinking (Proulx et al., 2005) have been widely used by biologists, mathematicians, social scientists, and computer scientists to explore interactions between entities, whether those entities are individuals in a school (Moody, 2001), species in a food web (Krause et al., 2003), nodes on a computer network (Pastor-Satorras and Vespignani, 2001), or proteins in metabolic pathways (Guimera and Amaral, 2005). Network analysis is used to explore the mathematical, statistical and structural properties of a set of items (nodes) and the connections between them (edges; Newman, 2003). With a few notable exceptions (for example, Ruan et al., 2006; Chaffron et al., 2010; Freilich et al., 2010), network analysis has not been widely applied to exploring co-occurrence patterns between microbial taxa in complex communities. To detect robust associations between microorganisms within and between habitats using network analysis, it is essential to have fairly detailed information on the microbial taxa found across relatively large numbers of samples, as without sufficiently large sample sets it will be difficult to determine whether or not co-occurrence patterns are statistically significant. Ideally sample sets should cover spatial or temporal gradients in environmental conditions in order for there to be a sufficient variability in taxon abundances to resolve co-occurrence patterns. As shown in recent studies that have used barcoded pyrosequencing to survey microbial communities in large numbers of samples (for example Fierer et al., 2008; Costello et al., 2009; Galand et al., 2009; Lauber et al., 2009), it is now possible to generate microbial datasets that can take full advantage of network analysis approaches and we can apply them to even highly diverse communities, like those found in soils, to explore co-occurrence patterns.

Network analysis of taxon co-occurrence patterns offers new insight into the structure of complex microbial communities, insight that complements and expands on the information provided by the more standard suite of analytical approaches. First, inter-taxa associations may help reveal the niche spaces shared by community members (even members of different domains of life, such as Bacteria and Archaea) or, perhaps, more direct symbioses between community members. Such information is particularly valuable in environments, such as soil, where the basic ecology and life history strategies of many microbial taxa remain unknown (Janssen, 2006). Exploring co-occurrence patterns between soil microorganisms can help identify potential biotic interactions, habitat affinities, or shared physiologies that could guide more focused studies or experimental settings. More generally, network analysis represents an approach for exploring and identifying patterns in large, complex datasets, patterns that may be more difficult to detect using the standard alpha/beta diversity metrics widely used in microbial ecology (Proulx et al., 2005).

Here we used network analyses to explore associations between prokaryotic taxa in soil, one of the most complex and taxon-rich microbial habitats on Earth. We analyzed over 160 000 bacterial and archaeal 16S rRNA gene sequences from 151 soil samples from a wide variety of ecosystem types in order to demonstrate the utility of network analyses and address the following questions: (i) Do soil microorganisms tend to co-occur more than expected by chance? (ii) Can the lack of agreement between observed and random intra-phyla co-occurrence be used as a proxy of niche differentiation? and (iii) Which taxa are generalists (broadly distributed across soil habitats) or specialists (restricted to certain habitats but locally abundant) and how these ecological categories shape network structure?

Materials and methods

Soil description and molecular methods
The dataset consisted of 151 soil samples distributed across North and South America, and Antarctica. The collected soils came from a broad range of ecosystems, climates and soil types (Supplementary Table 1). Soil collection protocol and methods for edaphic and environmental properties have been described previously (Fierer and Jackson, 2006; Bates et al., 2010).

Preparation of extracted DNA for pyrosequencing followed the protocol described in detail in Fierer et al., 2008 and Bates et al., 2010. In brief, a region of the 16S rRNA gene ($\sim$250 nucleotides) was amplified with the primers F515 (5′-GTGCCAGCMGCCGCGG TAA-3′) and R806 (5′-GGACTACVSGGGTATCT AAT-3′) that should amplify nearly all bacteria and archaea with few biases against particular groups (Bates et al., 2010). The resulting barcoded PCR product was normalized in equimolar amounts and sequenced on a Roche GS-FLX 454 automated pyrosequencer (Roche Applied Science, Branford, CT, USA) at the Environmental Genomics Core Facility (Engencore) at the University of South Carolina.

Sequence processing
Raw sequence data generated from pyrosequencing were processed in QIIME (Caporaso et al., 2010).

Briefly, sequences were quality trimmed and clustered into operational taxonomic units (OTUs) using a 90% identity threshold with uclust (Edgar, 2010). A 90% identity threshold, which corresponds approximately to the taxonomic level of Family for bacteria (Konstantinidis and Tiedje, 2007), was used to generate consistent OTUs with high abundances for subsequent analyses based on correlations and to circumvent potential taxonomic misclassifications due to sequencing anomalies. If we were to use the more standard 'species-level' OTU cutoff (97% sequence identity), the resulting OTU table would be far larger making data visualizations and analysis more difficult. At the 90% identity level, the final OTU table consisted of 160 469 sequences (average of 1063 sequences per sample) distributed into 4088 OTUs, of those 2798 were represented by more than 1 sequence. Taxonomic assignment was carried out with the RDP Classifier (Wang *et al.*, 2007), and manually curated by BLAST searches against the GenBank non-redundant nucleotide database (nt).

Data analysis
Non-random co-occurrence patterns were tested with the checkerboard score (*C*-score) under a null model preserving site frequencies (Stone and Roberts, 1990). A checkerboard unit is a 2×2 matrix where both OTUs occur once but on different sites. For network inference, we calculated all possible Spearman's rank correlations between OTUs with more than five sequences (1577 OTUs). This previous filtering step removed poorly represented OTUs and reduced network complexity, facilitating the determination of the core soil community. We considered a valid co-occurrence event to be a robust correlation if the Spearman's correlation coefficient (ρ) was both >0.6 and statistically significant (*P*-value <0.01; Junker and Schreiber, 2008). The nodes in the reconstructed network represent the OTUs at 90% identity, whereas the edges (that is, connections) correspond to a strong and significant correlation between nodes (see Supplementary File for the resulting network in GRAPHML format). In order to describe the topology of the resulting network, a set of measures (that is, average node connectivity, average path length, diameter, cumulative degree distribution, clustering coefficient and modularity) were calculated (Newman, 2003). All statistical analyses were carried out in the R environment (http://www.r-project.org) using vegan (Oksanen *et al.*, 2007) and igraph (Csárdi and Nepusz, 2006) packages. Networks were explored and visualized with the interactive platform gephi (Bastian *et al.*, 2009).

Results and discussion

General co-occurrence patterns
Soils are heterogeneous environments that harbor enormously diverse prokaryotic communities

(Torsvik *et al.*, 1990; Curtis *et al.*, 2002). Previous work has explored soil microbial diversity from various perspectives, including the estimation of species richness levels in individual samples (for example, Fierer *et al.*, 2007; Roesch *et al.*, 2007; Youssef and Elshahed, 2009), assessment of the abiotic variables that control the diversity and composition of communities (for example, McCaig *et al.*, 2001; Fierer and Jackson, 2006; Lauber *et al.*, 2009), and the assessment of how specific abiotic factors influence specific taxa (for example, Jones *et al.*, 2009). The relationships between microbial taxa also shape the structure of microbial communities (Prosser *et al.*, 2007), and thus, it can be expected that non-random co-occurrence patterns and significant inter-taxa relationships should occur.

After quality filtering and OTU clustering at the 90% identity, we obtained 2798 OTUs represented by more than one sequence distributed across the 151 soil samples included in this study (Supplementary Table 1). In order to assess non-random co-occurrence patterns, we first used an ecological measure based on checkerboard units (*C*-score; Stone and Roberts, 1990). Overall, we observed non-random co-occurrence pattern using the whole dataset (*C*-score $= 46.56$, *P*-value <0.01). Restricting the analysis to only those OTUs showing significant relationships (the ones appearing in Figure 1), the measure increased to *C*-score $= 185.03$ and *P*-value <0.01. A recent meta-analysis showed similar patterns of co-occurrence for microorganisms and macroorganisms suggesting that non-random community assembly may be a general characteristic across all life domains (Horner-Devine *et al.*, 2007). This finding that there are significant non-random co-occurrence patterns is not surprising, given that we have known for some time that many bacterial taxa exhibit predictable biogeographical patterns (Prosser *et al.*, 2007). Likewise, documenting non-random co-occurrence patterns is far different from actually identifying the causal mechanisms structuring the communities. However, non-random assembly patterns do indicate the dominance of deterministic processes including competitive interactions, non-overlapping niches or historical effects in shaping community composition (Horner-Devine *et al.*, 2007). Overall, the approach allowed us to conclude that soil microorganisms tended to co-occur more than expected by chance.

Network description
Once we established that the soil microbial assemblage patterns were certainly non-random, we further explored co-occurrence patterns using network inference based on strong and significant correlations (using non-parametric Spearman's; Junker and Schreiber, 2008). Correlation networks of co-occurring microorganisms permit the visual summary of lots of information (Chaffron *et al.*, 2010) and have been successfully applied to discern
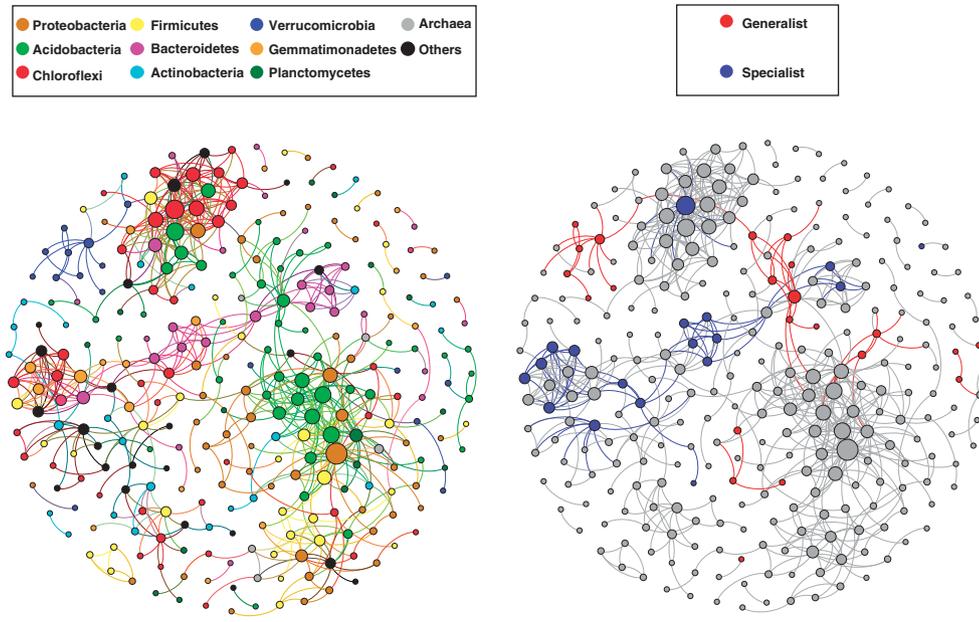
Figure 1 Network of co-occurring 90% cutoff OTUs based on correlation analysis. A connection stands for a strong (Spearman's $\rho > 0.6$) and significant (P-value <0.01) correlation. The size of each node is proportional to the number of connections (that is, degree). Left panel: OTUs colored by taxonomy. Right panel: OTUs colored by abundance and occupancy (generalists and specialists).

associations between marine microorganisms and their environment (Ruan *et al.*, 2006).

The resulting soil microbial network (Figure 1 and Supplementary Figure 1) consisted of 296 nodes (OTUs) and 679 edges (average degree or node connectivity 4.59; see Supplementary Figure 1 for a cumulative degree distribution). Some topological properties commonly used in network analysis were calculated to describe the complex pattern of inter-relationships between OTUs (Newman, 2003). The average network distance between all pairs of nodes (average path length) was 5.53 edges with a diameter (longest distance) of 18 edges. The clustering coefficient (that is, how nodes are embedded in their neighborhood and, thus, the degree to which they tend to cluster together) was 0.33 and the modularity index was 0.77 (values >0.4 suggest that the network has a modular structure; Newman, 2006). Overall, the soil microbial network was comprised of highly connected OTUs (~5 edges per node) structured among densely connected groups of nodes (that is, modules) and forming a clustered topology (as expected for real-world networks that are more significantly clustered than random graphs). These structural properties offer the potential for quick and easiest comparisons among complex datasets from different ecosystem types in order to explore how the general traits of a certain habitat type may influence the assembly of microbial communities.

The structural analysis also showed that OTUs from the same phyla tended to co-occur more (33%) than expected when considering observed phyla frequencies and random association (11%; Figure 1, left panel). Thus, the magnitude of the lack of agreement between the observed intra-phyla percentage of significant co-occurrences, on the one hand, and the expected assembling under random association, on the other hand, may be used as an index of niche preferences or synergetic relationships. This index may also capture differences among habitats (for example, aquatic vs soil) that may shed light on the ecological rules guiding microbial community composition. In all likelihood, most of these co-occurrence patterns are derived from taxa sharing similar ecological niches, not direct symbioses, however, these data alone do not allow us to separate these two possibilities. Some of the co-occurrence patterns reveal or confirm interesting ecological patterns for taxa that have not been well studied. For instance, members of the verrucomicrobial phylum tended to co-occur more (1.8%) than expected by chance (0.3%) suggesting that, despite being abundant and ubiquitous in soils (Bergmann *et al.*, 2011), they share a specific (and as yet undefined) niche (Figure 4). Another abundant and cosmopolitan phylum that also showed higher incidences of co-occurrence than expected by random association was Acidobacteria (2.4% at random while 9.4% observed; Figure 4). In this case, the pattern is most likely driven by the previously observed phenomenon that soil pH largely governs the distributions of many soil acidobacterial taxa (Jones *et al.*, 2009). Other examples were the Chloroflexi and *Deinococcus* groups, which were neither abundant nor generalists/cosmopolitan (Figure 4), but instead, appeared to be mostly restricted to desert soils. Several OTUs shared the same habitat preferences and thus appeared to be very interconnected (1.9% at random while 6.2%
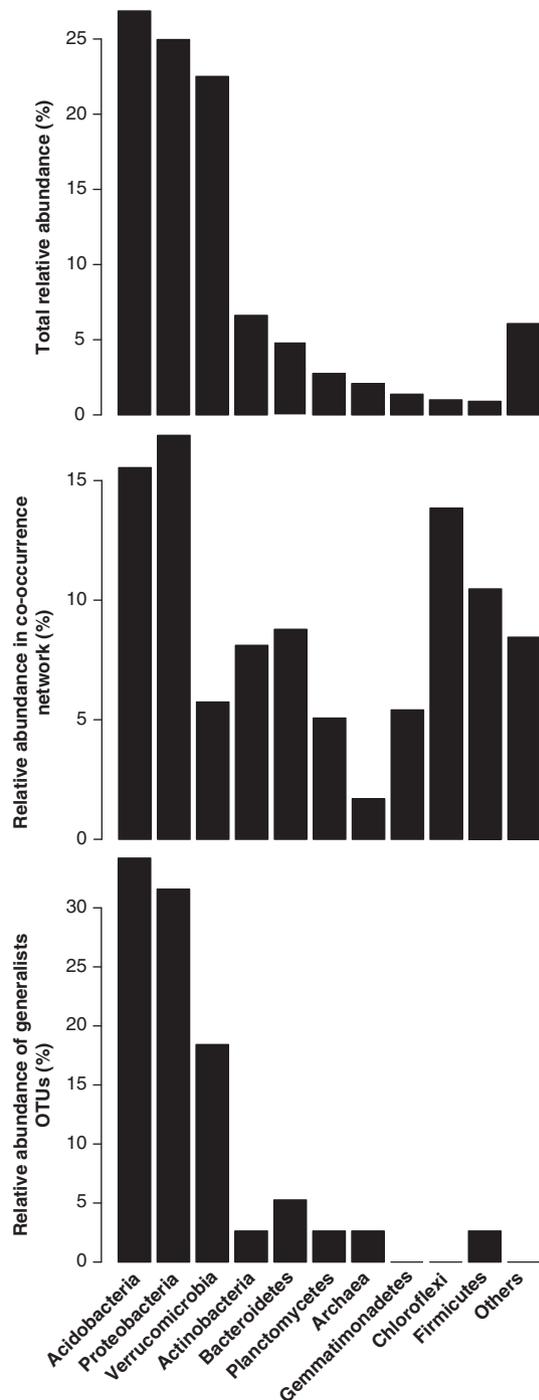
observed for Chloroflexi; 0.01% at random while 1% observed for *Deinococcus*). The degree of disagreement between observed and random co-occurrence may therefore provide further insights in the niche differentiation for the different populations sharing a common phylogeny at different levels of relatedness. Overall, these findings suggest that environmental filtering effects and niche differentiation are evident at broad taxonomic levels, as noted elsewhere (Philippot *et al.*, 2010).

Habitat generalists and specialists

Each of the OTUs represented by more than one sequence was drawn in the abundance vs occupancy plot (Figure 2) to split the set of taxa into two general categories: soil generalists, on the one hand (that is, broadly distributed microbial taxa, which we operationally define here as present in >80 of the 151 soils) and soil specialists (operationally defined here as those that were locally abundant, representing >2% of the sequences in individual libraries, but only found in <10 soil samples), on the other hand. Under this criterion, ~2% of the OTUs fell into the generalist category while ~1% fell into the specialist category. Despite using in the present work a high-throughput DNA sequencing method with higher sequencing depth than traditional methodologies, we may have missed very low abundance taxa that potentially could be habitat generalists. However, dividing taxa from ecological communities into these two categories, which we admittedly defined somewhat arbitrarily, is useful for defining ecological categories/strategies that offer additional information from those defined by phylogeny, taxonomy, or functional capacity (Magurran and Henderson, 2003). Recently, partitioning microbial taxa based on abundance and occupancy has been proved useful in the analysis of clinical samples (van der Gast *et al.*, 2011). In general ecology, positive relationships between mean abundance and occupancy have been observed at many spatial scales (Guo *et al.*, 2000; van der Gast *et al.*, 2011). However, we did not observe such trend in our dataset. Although most of the soil samples analyzed in the present work had their origins in temperate and fertile soils, the environmental variability covered in this study (that is, different habitats and a broad spatial scale) probably altered this relationship. For example, specialist bacterial OTUs inhabiting extreme environments such as deserts or Antarctica soils had a higher abundance than expected regarding their persistence in the overall sampling range.

Overall, we observed a consistent separation in the co-occurring network analysis between generalists and specialists (see significant correlations in Figure 1, right panel, against all the remaining OTUs, and a detail in Figure 3 adding taxonomic information). Specialists OTUs (locally abundant in a few samples) were composed of a diverse range of

Figure 2 Abundance (*y* axis) and occupancy (*x* axis) plot for the 90% cutoff OTUs. Habitat generalists OTUs (in red) defined as appearing in >80 soil samples. Habitat specialists OTUs (in blue) defined as locally abundant (>18 sequences) and appearance in <10 soils.

Figure 3 Network of co-occurring generalists and specialists 90% cutoff OTUs based on correlation analysis. A connection stands for a strong (Spearman's $\rho > 0.6$) and significant (*P*-value <0.01) correlation. The size of each node is proportional to the number of connections (that is, degree). Labels according to taxonomic affiliation: Ac, Acidobacteria. A.R, Alphaproteobacteria; Rhizobiales. A.Rh, Alphaproteobacteria; Rhodobacterales. A.S, Alphaproteobacteria; Sphingomonadales. Ba.F, Bacteroidetes; Flavobacteria. Ba.S, Bacteroidetes; Sphingobacteria. Ch, Chloroflexi. Cr, Crenarchaeota. Cy, Cyanobacteria; D, Deltaproteobacteria. De, *Deinococcus*. G, Gemmatimonadetes. Ga, Gammaproteobacteria. V, Verrucomicrobia.

348

Figure 4 Relative abundance of different microbial taxonomic groups. Top panel: number of sequences in all soil samples. Middle panel: number of significant co-occurrent OTUs (nodes from Figure 2). Low panel: number of cosmopolitan OTUs.

phylogenetic groups not common in soils from most biomes (that is, Chloroflexi, *Deinococcus*, Gemmatimonadetes; Figure 3). Generalists OTUs (distributed broadly), in turn, were typical soil members from the Acidobacteria, Proteobacteria (especially of the Alpha subclass) and Verrucomicrobia groups (see Figure 4, top and low panels, and Janssen, 2006

for a recent review). The different taxonomic composition and range of distribution probably influenced the network structure observed (Figure 3), indicating that these two ecological categories shaped differently the network structure. The generalists network was less connected and more compartmentalized (19 significant co-occurrences and five compartments) than the specialists network (29 significant co-occurrences and two compartments) probably because of the highest habitat variability covered by the former and the presence in restricted environments by the latter. Thus, the two somehow arbitrary ecological categories that we established allowed us to capture additional information on the community assembling structure as previously shown for macroorganisms (Pandit *et al.*, 2009).

Finally, the significant correlations found between the generalists OTUs and the whole dataset are shown in Table 1. The listed correlations do not include co-occurrent events among members of the same taxonomic group. Generalists OTUs belonged to abundant phyla such as Verrucomicrobia, Acidobacteria, Proteobacteria and Bacteroidetes. Interestingly, generalists OTUs not classified as typical and abundant soil members (such as Deltaproteobacteria and Crenarchaeota) co-occurred with other ubiquitous members. The case of the crenarchaeotal OTU (closely related to *Candidatus Nitrososphaera gargensis* and previously described as ubiquitous in soil; Bates *et al.*, 2010) is particularly remarkable due to our poor understanding of the niches occupied by this taxon, even though it has been proposed that related Crenarchaeota may have an important role in the nitrogen cycle as ammonia oxidizers (Leininger *et al.*, 2006). This particular crenarchaeotal OTU co-occurred with sequences closely related to methane oxidizers (Table 1). This observation may be a first step to revise the expected functional role of soil Crenarchaeota in the nitrification process because of the high sequence similarity of the enzymes that catalyze ammonia oxidation (ammonia monooxygenase) and methane oxidation (particulate methane monooxygenase; Holmes *et al.*, 1995). This is an example of the potential that the approach proposed in the present work has to gain knowledge on elusive but ecologically relevant microorganisms.

Final remarks

With this work we have demonstrated the utility of including network analysis approaches in the repertory of statistical approaches so far available to microbial ecologists. By employing network analyses to a large soil microbial dataset generated by pyrosequencing, the process of exploring the complex set of data was more feasible and interesting unseen patterns emerged, including non-random association, deterministic processes at different taxonomic levels and unexpected relationships

Table 1 Taxonomy of generalists OTUs (red nodes in Figure 3) and their significant (*P*-value <0.01) co-occurrent OTUs

Cosmopolitan	Co-occurrent	Spearman's ρ	Number of co-occurring soils	Closest cultured relative	% identity
Deltaproteobacteria	Verrucomicrobia	0.61	53	*Spartobacteria bacterium* Gsoil 144	91
	Gammaproteobacteria	0.61	80	*Steroidobacter* sp. ZUMI 37	100
	Acidobacteria	0.65	93	*Acidobacteriaceae bacterium* Gsoil 1619	90
	Acidobacteria	0.74	102	*Acidobacteria bacterium* IGE-003	93
	Firmicutes	0.61	69	*Moorella thermoacetica*	84
Crenarchaeota	Alphaproteobacteria; Rhodospirillales	0.62	64	*Azospirillum amazonense* strain CBAmc	89
	Alphaproteobacteria; Rhizobiales	0.60	68	*Balneimonas* sp. sptzw09	100
	Bacteroidetes; Sphingobacteria	0.63	82	*Flavosolibacter* sp. HU1-JC5	96
	Alphaproteobacteria; Sphingomonadales	0.71	104	*Sphingomonadaceae bacterium* Gsoil 690	100
Verrucomicrobia	Acidobacteria	0.71	105	*Acidobacteriaceae bacterium* Gsoil 1619	91
	Firmicutes	0.64	97	Unknown sulfate-reducing bacterium clone A8	86
	Acidobacteria	0.68	130	*Bacterium Ellin6075*	97
	Acidobacteria	0.64	70	*Acidobacteria bacterium* IGE-017	91
Acidobacteria	Verrucomicrobia	0.68	130	*Verrucomicrobia bacterium* WSF2-44	95
	Actinobacteria	0.62	39	*Rubrobacter xylanophilus* DSM 9941	91
	Gammaproteobacteria	0.66	83	*Steroidobacter* sp. ZUMI 37	100
	Deltaproteobacteria	0.65	93	*Desulfonatronum* sp. ASO4–1	87
	Crenarchaeota	0.63	78	*Candidatus Nitrososphaera gargensis*	93
	Bacteroidetes; Sphingobacteria	0.63	56	*Sphingobacteriales bacterium* TP524	95
	Gammaproteobacteria	0.62	74	Gamma proteobacterium CH43	95
Alphaproteobacteria; Rhizobiales	Betaproteobacteria; Burkholderiales	0.61	79	*Burkholderia glathei* strain YUST-DW12	100
	Gammaproteobacteria	0.64	124	*Proteobacterium Ellin*181	100
Alphaproteobacteria; Sphingomonadales	Crenarchaeota	0.71	104	*Candidatus Nitrososphaera gargensis*	93
	Bacteroidetes; Sphingobacteria	0.74	89	*Flavosolibacter* sp. HU1-JC5	96
Gammaproteobacteria	Deltaproteobacteria	0.61	33	*Thermodesulforhabdus* sp. nov. M40/2 CIV-3.2	87
	Alphaproteobacteria; Rhizobiales	0.64	124	*Rhodoplanes* sp. 303	97
Bacteroidetes; Sphingobacteria	Gemmatimonadetes	0.61	74	*Gemmatimonas aurantiaca* T-27	88

Abbreviation: OTUs, operational taxonomic units.
OTUs belonging to Firmicutes, Planctomycetes, Actinobacteria, and Burkholderiales and Rhodocyclales from the phylum Betaproteobacteria had no significant co-occurrent matches.

between community members. Different ecological rules guiding microbial community composition may be reflected in diverse network structure among habitats (for example, aquatic vs soil, or early colonized vs late-successional ecosystems) that deserves further research. The next logical step is to go beyond merely describing the patterns revealed by the network analysis and design more focused experiments, or the study of specific environmental gradients and community shifts over time, in order to understand the mechanisms producing patterns of community coexistence, that is, what finally determines which and how many species live together in a community.

350

and members of the Rob Knight lab in UC, particularly Greg Caporaso for computational support and the development of QIIME. AB is supported by the Spanish FPU predoctoral scholarship program and EOC by grants PIRENA CGL2009-13318 and CONSOLIDER-INGENIO 2010 GRACCIE CSD2007-00067 from the Spanish Ministerio de Ciencia e Innovación (MICINN).

References

Auguet JC, Barberán A, Casamayor EO. (2010). Global ecological patterns in uncultured Archaea. *ISME J* **4**: 182–190.

Barberán A, Casamayor EO. (2010). Global phylogenetic community structure and beta-diversity patterns of surface bacterioplankton metacommunities. *Aquat Microb Ecol* **59**: 1–10.

Barberán A, Casamayor EO. (2011). Euxinic freshwater hypolimnia promote bacterial endemicity in continental areas. *Microb Ecol* **61**: 465–472.

Bastian M, Heymann S, Jacomy M. (2009). *Gephi: An Open Source Software For Exploring and Manipulating Networks.* In International AAAI conference on weblogs and social media: San Jose, California.

Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N. (2010). Examining the global distribution of dominant archaeal populations in soil. *ISME J* **5**: 908–917.

Bergmann GT, Bates ST, Eilers KG, Lauber CL, Caporaso JG, Walters WA et al. (2011). The under-recognized dominance of Verrucomicrobia in soil bacterial communities. *Soil Biol Biochem* **43**: 1450–1455.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Chaffron S, Rehrauer H, Pernthaler J, von Mering C. (2010). A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res* **20**: 947–959.

Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. (2009). Bacterial community variation in human body habitats across space and time. *Science* **326**: 1694–1697.

Csárdi G, Nepusz T. (2006). The igraph software package for complex network research. *InterJ Complex Syst* **1695**: 2006.

Curtis TP, Sloan WT, Scannell JW. (2002). Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**: 10494–10499.

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

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626.

Fierer N, Breitbart M, Nulton J, Salamon P, Lozupone C, Jones R et al. (2007). Metagenomic and small-subunit rrna analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol* **73**: 7059–7066.

Fierer N, Hamady M, Lauber CL, Knight R. (2008). The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proc Natl Acad Sci USA* **105**: 17994–17999.

Freilich S, Kreimer A, Meilijson I, Gophna U, Sharan R, Ruppin E. (2010). The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res* **38**: 3857–3868.

Fuhrman JA, Steele JA. (2008). Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquati Microb Ecol* **53**: 69–81.

Galand PE, Casamayor EO, Kirchman DL, Potvin M, Lovejoy C. (2009). Unique archaeal assemblages in the arctic ocean unveiled by massively parallel tag sequencing. *ISME J* **3**: 860–869.

Guimera R, Amaral LAN. (2005). Functional cartography of complex metabolic networks. *Nature* **433**: 895–900.

Guo Q, Brown JH, Valone TJ. (2000). Abundance and distribution of desert annuals: are spatial and temporal patterns related? *J Ecol* **88**: 551–560.

Holmes AJ, Costello A, Lindstrom ME, Murrell JC. (1995). Evidence that particulate methane monooxygenase and ammonia monooxygenase may be evolutionarily related. *FEMS Microbio Lett* **132**: 203–208.

Horner-Devine MC, Silver JM, Leibold MA, Bohannan BJ, Colwell RK, Fuhrman JA et al. (2007). A comparison of taxon co-occurrence patterns for macro- and micro-organisms. *Ecology* **88**: 1345–1353.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. (2001). Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* **67**: 4399–4406.

Janssen PH. (2006). Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl Environ Microbiol* **72**: 1719–1728.

Jones RT, Robeson MS, Lauber CL, Hamady M, Knight R, Fierer N. (2009). A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J* **3**: 442–453.

Junker BH, Schreiber F. (2008). Correlation Networks. In: *Analysis of Biological Networks.* Wiley-Interscience.

Konstantinidis KT, Tiedje JM. (2007). Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Curr Opin Microbiol* **10**: 504–509.

Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW. (2003). Compartments revealed in food-web structure. *Nature* **426**: 282–285.

Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75**: 5111–5120.

Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW et al. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**: 806–809.

Lozupone CA, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436.

Magurran AE, Henderson PA. (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature* **422**: 714–716.

McCaig AE, Glover LA, Prosser JI. (2001). Numerical analysis of grassland bacterial community structure under different land management regimens by using 16S ribosomal DNA sequence data and denaturing gradient gel electrophoresis banding patterns. *Appl Environ Microbiol* **67**: 4554–4559.

Moody J. (2001). Race, school integration, and friendship segregation in America. *Am J Sociol* **107**: 679–716.

Newman MEJ. (2003). The structure and function of complex networks. *SIAM Rev* **45**: 167–256.

Newman MEJ. (2006). Modularity and community structure in networks. *Proc Natl Acad Sci USA* **103**: 8577.

Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P *et al.* (2007). Vegan: community ecology package. *R package version* **1**: 8.

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

Pandit SN, Kolasa J, Cottenie K. (2009). Contrasts between habitat generalists and specialists: an empirical extension to the basic metacommunity framework. *Ecology* **90**: 2253–2262.

Pastor-Satorras R, Vespignani A. (2001). Epidemic spreading in scale-free networks. *Phys Rev Lett* **86**: 3200–3203.

Philippot L, Andersson SG, Battin TJ, Prosser JI, Schimel JP, Whitman WB *et al.* (2010). The ecological coherence of high bacterial taxonomic ranks. *Nat Rev Microbiol* **8**: 523–529.

Prosser JI, Bohannan BJ, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP *et al.* (2007). The role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5**: 384–392.

Proulx SR, Promislow DE, Phillips PC. (2005). Network thinking in ecology and evolution. *Trends Ecol Evol* **20**: 345–353.

Roesch LFW, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD *et al.* (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J* **1**: 283–290.

Ruan Q, Dutta D, Schwalbach MS, Steele JA, Fuhrman JA, Sun F. (2006). Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors. *Bioinformatics* **22**: 2532.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stone L, Roberts A. (1990). The checkerboard score and species distributions. *Oecologia* **85**: 74–79.

Torsvik V, Goksoyr J, Daae FL. (1990). High diversity in DNA of soil bacteria. *Appl Environ Microbiol* **56**: 782–787.

van der Gast C, Walker AW, Stressmann FA, Rogers GB, Scott P, Daniels TW *et al.* (2011). Partitioning core and satellite taxa from within cystic fibrosis lung bacterial communities. *ISME J* **5**: 780–791.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261.

Youssef NH, Elshahed MS. (2009). Diversity rankings among bacterial lineages in soil. *ISME J* **3**: 305–313.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)