Received: 12 June 2018 | Revised: 5 December 2019 | Accepted: 6 December 2019

DOI: 10.1111/1755-0998.13126

Check for updates

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Investigating microbial associations from sequencing survey data with co-correspondence analysis

Benjamin Alric[1] | Cajo J. F. ter Braak[2] | Yves Desdevises[3] | Hugo Lebredonchel[3] | Stéphane Dray[1]

[1]CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Villeurbanne, France

[2]Biometris, Wageningen University and Research, Wageningen, The Netherlands

[3]CNRS, UMR 7232, BIOM, Biologie Intégrative des Organismes Marins, Observatoire Océanologique, Sorbonne Université, Banyuls sur Mer, France

Correspondence
Benjamin Alric, Irstea, UR Riverly, Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, 5 rue de la Doua CS 20244, F-69625, Villeurbanne, France.
Email: benjamin.alric@irstea.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: DECOVIR ANR-12-BSV7-0009; EcoNet, Grant/Award Number: ANR-18-CE02-0010-01

Abstract

Microbial communities, which drive major ecosystem functions, consist of a wide range of interacting species. Understanding how microbial communities are structured and the processes underlying this is crucial to interpreting ecosystem responses to global change but is challenging as microbial interactions cannot usually be directly observed. Multiple efforts are currently focused to combine next-generation sequencing (NGS) techniques with refined statistical analysis (e.g., network analysis, multivariate analysis) to characterize the structures of microbial communities. However, most of these approaches consider a single table of sequencing data measured for several samples. Technological advances now make it possible to collect NGS data on different taxonomic groups simultaneously for the same samples, allowing us to analyse a pair of tables. Here, an analytical framework based on co-correspondence analysis (CoCA) is proposed to study the distributions, assemblages and interactions between two microbial communities. We show the ability of this approach to highlight the relationships between two microbial communities, using two data sets exhibiting various types of interactions. CoCA identified strong association patterns between autotrophic and heterotrophic microbial eukaryote assemblages, on the one hand, and between microalgae and viruses, on the other. We demonstrate also how CoCA can be used, complementary to network analysis, to reorder co-occurrence networks and thus investigate the presence of patterns in ecological networks.

KEYWORDS

co-correspondence analysis, co-occurrence network, Mamiellophyceae, microbial eukaryotes, next-generation sequencing, *Prasinovirus*

1 | INTRODUCTION

Microbial communities are highly diverse (Rappé & Giovannoni, 2003) and drive major ecosystem functions (e.g., carbon sequestration, climate regulation, gas regulation, nutrient cycling; Ducklow, 2008; Falkowski, Fenchel, & Delong, 2008; Hutchins & Fu, 2017). Understanding how these systems are structured and identifying the underlying processes is crucial to predicting communities and ecosystem responses to global change (Fuhrman, 2009; Graham et al., 2016). Biotic interactions across taxonomic groups (e.g., predation, parasitism, mutualism or competition) are of broad interest because they are expected to influence the structure and composition of communities (Wisz et al., 2013). Unfortunately, our understanding of the underlying assemblage rules of microbial communities

remains limited (Cordero & Datta, 2016; Little, Robinson, Peterson, Raffa, & Handelsman, 2008).

The emergence of high-throughput sequencing techniques (next-generation sequencing; NGS) gave access to the diversity of whole microbial communities, including the noncultivable fraction (Handelsman, 2004; Zimmerman, Izard, Klatt, Zhou, & Aronson, 2014). With the large amount of data generated in a single NGS experiment, powerful statistical methods are needed to assess and explain structural patterns in such complex data sets (Bálint et al., 2016; Paliy & Shankar, 2016). A common approach is to combine NGS techniques with network analysis to represent and characterize interactions between partners in microbial communities (Cardona, Weisenhorn, Henry, & Gilbert, 2016; Vacher et al., 2016). Various computational methods have been developed to infer co-occurrence networks from NGS data sets (e.g., CoNet: Faust et al., 2012; SparCC: Friedman & Alm, 2012; rebacca: Ban, An, & Jiang, 2015; CCLasso: Fang, Huang, Zhao, & Deng, 2015; spiec-easi: Kurtz et al., 2015). Another popular approach is to use ordination methods to extract information from NGS data and describe the variations in community composition among samples (Paliy & Shankar, 2016). Ordination methods arrange objects in a multidimensional space using directly the original raw data table (e.g., principal component analysis [PCA], correspondence analysis [CA]), or after computing a distance matrix (e.g., nonmetric multidimensional scaling, principal coordinate analysis) (Legendre & Legendre, 2012).

These different approaches are based on a single table composed of read counts for each operational taxonomic unit (OTU) measured for several samples. Technological advances now make it possible to acquire NGS data on different taxonomic groups simultaneously for the same samples (Fierer et al., 2007) and lead to analysis of a pair of tables (i.e., OTU composition for the same sampling sites for two different taxonomic groups). To analyse such a pair of tables, common practice consists of merging the two tables into a single one and then applying network analysis (e.g., Banerjee et al., 2016; Kueneman et al., 2016; Ma et al., 2016) and/or multivariate analysis (Bergelson, Mittelstrass, & Horton, 2019; Cannon et al., 2017). However, this data aggregation is unsuitable especially when NGS data sets, which are a function of sequencing depth (Ni, Yan, & Yu, 2013), are standardized by dividing read counts by the total number of reads in each sample. In this case, the normalization step and further analysis are very sensitive to the difference in number of OTUs and associated counts in each taxonomic group. Hence, it is important to use techniques allowing for the analysis of a pair of NGS data tables while preserving the original structure of the data. In addition, these approaches must be able to mitigate the statistical bias stemming from the high dimensionality (i.e., a number of samples substantially lower than the number of variables), sparsity (i.e., a high proportion of zero counts) and the compositional nature (i.e., nonindependence of relative abundances induced by the row-sum normalization) that characterize NGS data (Li, 2015). For network analysis, the spiec-easi method has been adapted to infer associations among microorganisms in a cross-domain analysis (Tipton et al., 2018). For multivariate analyses, several two-table methods exist and have been presented to microbial ecologists in methodological reviews (Buttigieg & Ramette, 2014; Paliy & Shankar, 2016; Ramette, 2007). However, these studies focused on asymmetric methods (e.g., canonical correspondence analysis and redundancy analysis) that aim to explain the composition of microbial communities by a limited number of environmental predictors. Unfortunately, these methods are not adapted to link two NGS data tables as they require that there are fewer predictor variables than samples (Dray, Chessel, & Thioulouse, 2003) and thus are not able to deal with the high dimensionality of NGS data. Moreover, these methods compute linear combinations of the predictor variables, which is not suitable if the table of predictors contains community data that display unimodal structure and/or are sum-normalized (as the NGS data).

This study aims to propose an analytical framework based on co-correspondence analysis (CoCA; ter Braak & Schaffers, 2004), a two-table coupling method developed in community ecology, to study the distributions and assemblages between two sets of communities. This framework is based on correspondence analysis, a method that effectively handles proportional data that contain many zeroes (Gauch, Whittaker, & Wentworth, 1977; Greenacre, 2009; Jackson, 1997), like NGS data (Paulson, Stine, Bravo, & Pop, 2013). We show how this method allows us to extract information about the costructure among two microbial communities to estimate the congruence between them. Finally, we show that the outputs of the method can be used to reorder inferred co-occurrence networks to enhance the visualization of microbial associations and the understanding of assemblage patterns within networks. Hence, our approach echoes the current debate regarding the practices used to create network visualizations which are both aesthetically appealing and have high information content (for a review see Pocock et al., 2016). We illustrate our approach using two real data sets, one on autotrophic and heterotrophic microbial eukaryotes in shallow freshwater systems and another on microalgae and viruses in marine systems.

2 | MATERIALS AND METHODS

2.1 | Studying cross-taxon congruence by CoCA

Analysing the relationships between two communities (i.e., costructure) raises a number of issues, especially concerning the ability of multivariate methods to link two tables containing abundance data and to deal with the particular case of NGS data that have high dimensionality and contain many zeroes. We consider two tables $\mathbf{X}=[x_{ij}]$ and $\mathbf{Y}=[y_{ik}]$ containing the relative abundances of p and q OTUs (columns) in n samples (rows). Simple multivariate methods such as PCA or CA are useful to summarize the main structures of each table separately. These methods provide graphical representations in which samples and OTUs are arranged along axes allowing us to identify the OTUs that best discriminate the samples based on their composition. Whereas PCA is based on the diagonalization of a correlation matrix and thus assumes linear relationships between OTUs and axes, CA relies on the principle of weighted averaging and thus assumes unimodal response of OTUs along axes (see Appendix

S1). In contrast to PCA, CA aims to differentiate samples based on their relative composition in OTUs. Hence, it is adapted to NGS data (compositional data that contain many zeroes) but not designed to identify changes in absolute abundance (see Appendix S1).

CoCA extends CA to analyse a pair of tables and aims to identify which structures are common to both data sets (ter Braak & Schaffers, 2004). In practice, this method seeks for two sets of OTU scores to compute two sample scores (one for each table) by weighted averaging (equations 6 and 8 in Appendix S1): a sample is located at the centre of the OTUs it contains. The two sample scores allow us to best summarize the covariation between the two sets of OTUs as they have maximal covariance (equation 4 in Appendix S1). This procedure allows us to retain the weighted averaging properties of CA and estimate the costructure between **Y** and **X** (ter Braak & Schaffers, 2004). CoCA preserves another fundamental property of CA, namely the use of the chi-square (χ^2) distance, which is particularly adapted to NGS data as it properly handles zero values (and in particular double absences) and thus is not hampered by zero inflation (Legendre & Legendre, 2012). Also, unlike other methods such as canonical correspondence analysis or redundancy analysis, CoCA is based on covariance and thus allows us to deal with tables in which the number of samples is less than the number of OTUs and thus the high dimensionality of NGS data.

CoCA can be implemented either in a predictive or a symmetric framework according to whether it is assumed that one set of OTUs (**X**) explains the other (**Y**), or not. Here, CoCA in its predictive form (pCoCA) is used for the microalgae–virus data set while the symmetric form (sCoCA) is used for the microbial eukaryote data set. In the first case, pCoCA is chosen because we wish to investigate the composition of virus communities assuming that the occurrence of viruses depends mainly on whether microalgal hosts are present or absent at a particular sample location. In the second case, sCoCA is chosen as we simply wish to study the relationships between autotrophic and heterotrophic microbial eukaryotes without assuming that the composition of a community predicts the second. In its predictive form, CoCA is based on partial least-squares regression analysis (PLS) in a SIMPLS version (ter Braak & Schaffers, 2004) to deal with high dimensionality and subsequent collinearity in the table of explanatory variables. PLS searches for a linear regression model from a set of orthogonal components (called latent factors) built from collinear explanatory variables with the constraint that these components maximize the covariance with the response variables (Martens, 2001). In its symmetric form, CoCA fits in the framework of co-inertia analysis (COIA, Dolédec & Chessel, 1994) that is not affected by the problem of collinearity Dray et al., 2003. COIA relates two data tables in a symmetric way, by providing two sets of sample scores of maximal covariance Dray et al., 2003.

2.2 | Ordination of the structure and assemblage of interacting communities

From sCoCA, ordination diagrams can be made in the usual way by jointly plotting the OTU and sample scores of each community for the first axes of the analysis (biplot; ter Braak & Schaffers, 2004). These biplots satisfy the principles of weighted averaging so that an OTU is located at the centre of the samples in which it occurs whereas a sample is at the centre of the OTUs that it contains. From pCoCA, the fit of the response community (viruses) to the predictive community (microalgal hosts) as well as the variations in the composition of communities can also be displayed in ordination diagrams (ter Braak & Schaffers, 2004). For instance, the joint plot of sample scores of hosts (table **X**) with OTU scores of viruses (table **Y**) displays the fit of the virus OTUs from the host communities. Microalgae (hosts) can be superimposed using the loadings (i.e., the coefficients of variables on the first components of the PLS) of predictor species (i.e., hosts) so that both types of OTUs are jointly displayed. For all these diagrams, axes are optimized to maximize the link between assemblages (covariance) rather than to depict associations within individual tables.

2.3 | Real data application

We illustrate the use of CoCA to study the congruence between microbial communities from NGS data, employing the data of Simon et al. (2015) on the synecology of microbial eukaryotes in shallow freshwater systems, and a data set acquired during our research programme (ANR programme DECOVIR-12-BSV7-0009) on the monitoring of microalgae and viruses in marine systems.

2.3.1 | Case study 1: Microbial eukaryotes

Surface water was sampled monthly from April 2011 to April 2013 in five small shallow freshwater systems (four ponds: Etang des Vallées [EV], La Claye [LC], Mare Gabard [GB], Saint Robert [SR]; and one brook: Ru Sainte Anne [RSA]), located in the Natural Regional Park of the Chevreuse Valley (south of Paris, France). These systems were characterized by different local environmental conditions. Raw genomic sequences were obtained from 18S rDNA fragments, encompassing the V4 hypervariable region, applying 454 pyrosequencing and filtered to remove potential spurious sequences using a local pipeline (Simon et al., 2014). Sequences from all samples were then processed together and clustered into OTUs at a 0.98 similarity cutoff using CD hit (Fu, Niu, Zhu, Wu, & Li, 2012), and singletons were eliminated, before assigning OTUs to taxonomic groups based on sequence similarity to the PR2 database (Guillou et al., 2013). From this overall OTU table, we used the method of Simon et al. (2015) to select the most abundant OTUs. Finally, we subdivided the data set by grouping OTUs into two functional groups (autotrophic and heterotrophic) based on literature information (Genitsaris, Monchy, Breton, Lecuyer, & Christaki, 2016; Simon et al., 2015 and references within), and we obtained one table for autotrophic microbial eukaryotes (n = 108, p = 122) and another table for microbial heterotrophic eukaryotes (n = 108, p = 104).

2.3.2 | Case study 2: Microalgae–virus system

The data set coming from our research programme contains an NGS-based eukaryotic microalgae community table (photosynthetic picoeukaryotes in the class Mamiellophyceae) and an NGS-based virus community table (viruses infecting this class of eukaryotic phytoplankton and belonging to the genera *Prasinovirus* of the family Phycodnaviridae). The data have been acquired across four sites in the northwest Mediterranean Sea (Gulf of Lion) and sampled monthly from March 2013 to April 2014. The Gulf of Lion is characterized by contrasted environments, including eutrophic lagoons connected to the sea, nutrient-rich coastal sites and oligotrophic open-sea locations. The sample locations included two sites in Leucate lagoon (one coastal site [LB] and another site [LA] at the level of the Grau, that is, connected to the sea), a coastal site (SA, marine station included in the French marine monitoring network SOMLIT) and an open-sea site (MA, marine station included in the monitoring network MOOSE). The characterization of *Prasinovirus* was based on analysing the partial sequence of the DNA polymerase gene (*PolB*) amplified using two primer sets (Chen & Suttle, 1995; Clerissi et al., 2014). For Mamiellophyceae, the sequence of the V9 region of the 18S rDNA was amplified using primers defined by Amaral-Zettler, McCliment, Ducklow, and Huse (2009). The genomic sequences of *PolB* and the V9 region were amplified and sequenced using an Illumina MiSeq platform (GeT-PlaGe, INRA). Sequences were processed and clustered into OTUs at a 0.99 similarity cutoff and singletons were removed using MOTHUR version 1.35.1 (Schloss et al., 2009) for Mamiellophyceae and USEARCH version 7 (Edgar, 2010) combined with MUSCLE software (Edgar, 2004) for *Prasinovirus*. Sequences were then compared against the PR2 database (Guillou et al., 2013) and the NCBI database for Mamiellophyceae and *Prasinovirus*, respectively, in order to assign OTUs to taxonomic groups based on similarity. Finally, we focused specifically on OTUs assigned to the family Mamiellales of the class Mamiellophyceae, and notably the genera *Bathycoccus*, *Micromonas* and *Ostreococcus* which usually dominate this class in the Gulf of Lion and more generally the picoeukaryotic fraction in other ecosystems (Wu, Huang, & Zhong, 2013; Zhu, Massana, Not, Marie, & Vaulot, 2005). Subsequently, the *Prasinovirus* data set was limited to OTUs assigned as *Bathycoccus* viruses (BpVs), *Micromonas* viruses (MpVs) and *Ostreococcus* viruses (OtVs). The dominant microalgae and virus OTUs (i.e., 0.1% of mean relative abundance for at least two samples) were selected to obtain one microalgae table (n = 31, p = 67) and one virus table (n = 31, q = 98).

2.4 | Statistical analysis

Tables **X** and **Y** of autotrophic and heterotrophic microbial eukaryotes, respectively, were subjected to sCoCA. To test the significance of the global covariation between the two tables, a Monte-Carlo permutation procedure with 9,999 permutations was used. In each permutation, sCoCA (by considering all axes) is reapplied to obtain a value of the covariance between table **Y** and row-permuted **X** (so that samples are randomized while preserving the relative abundance of individuals). Note that the choice of the table to be reordered is not important here because we used the symmetric form of the CoCA. A null distribution was estimated from covariance calculated for the permuted data. The observed covariance is then compared to the distribution obtained under the null hypothesis. The positions of the samples on ordination axis of each table are then correlated to show the overall level of covariation between them. For the microalgae–virus data set, both tables were subjected to pCoCA with the SIMPLS algorithm. In their seminal paper, ter Braak and Schaffers (2004) suggest that the number of axes used to summarize the data can be selected by a "leave-one-out" cross-validation procedure to maximize the cross-validatory fit (%), which measures how well the table **X** (microalgae in our case) predicts the response table **Y** (viruses in our case). Working just with these significant axes provides a measure of association between the tables by removing random noise and keeping only the major dimensions of ecological variability. Finally, we combine CoCA and network analysis so that nodes in co-occurrence networks are reordered according to the OTU scores from the CoCA, and thus from the costructure between communities. A novel extension of the SPIEC-EASI method (Tipton et al., 2018) was used to infer the cross-group co-occurrence networks between two data sets. We used the neighbourhood (MB) setting and selected the optimal sparsity parameter based on the Stability Approach to Regularization Selection (StARS) (Liu, Roeder, & Wasserman, 2010). The StARS variability threshold was set to 0.05 for networks built from the two data sets. All statistical analyses were performed with the R software (R Core Team, 2019) and using the *cocorresp* package (Simpson, 2009) for CoCAs and the *SpiecEasi* package (Kurtz et al., 2015) for co-occurrence networks. A supporting information contains an R script and example data (from case study 1 and case study 2) allowing users to reproduce the analysis and apply them on their own data sets.

3 | RESULTS

3.1 | Case study 1: Microbial eukaryotes

The common variance between the two microbial groups computed from the sCoCA explained, significantly (p = .001), 13.21% of the total variation of autotrophic microbial eukaryotes and 15.22% of heterotrophic microbial eukaryotes. Of the common variance, 39.51% was accounted by the first three axes of the sCoCA (sCoCA axis 1:18.95%, sCoCA axis 2:10.55%, sCoCA axis 3:10.01%). The first three ordination axes of the autotrophic eukaryotes were highly correlated with the first three ordination axes of the heterotrophic eukaryotes (correlations being 0.95, 0.92 and 0.92), demonstrating a high degree of similarity in change between autotrophic and heterotrophic microbial eukaryote assemblages. Communities of autotrophic and heterotrophic microbial eukaryotes covaried along a brook/pond gradient on the first axis (from left to right), and an interpond variability on the second axis (Figure 1). Marked differences in the composition of the two communities are visible in joint plots. In the brook system (i.e., RSA),

the heterotrophic microbial eukaryote community is mainly composed of fungi, Marine Stramenopiles (MAST), Labyrinthulida and Telonema, whereas in pond systems (i.e., EV, LC, MG, SR) Ciliophora, Biocosoecida, Katablepharida and Choanoflagellida dominated the communities (Figure 1a). Differences between pond systems are explained by higher relative abundance of Biocosoecida and Katablepharida in EVs and LCs and higher relative abundance of Ciliophora in SRs. OTU scores of autotrophic microbial eukaryotes indicated that the patterns in the heterotrophic communities are associated with a structure in the autotrophic community (Figure 1b). In the brook system, the autotrophic microbial eukaryote community exhibits high relative abundances of specific OTUs of Bacillariophyceae, Chrysophyte and Cryptophyta. In pond systems, the autotrophic communities mainly comprised other specific OTUs of Chlorophyta, Chrysophyte and Cryptophyta. Note that Dinophyta and Streptophyta were found exclusively in pond systems (in particular in MG and SR, respectively).

The community organization of microbial eukaryotes was highlighted from a cross-group co-occurrence network between autotrophic and heterotrophic individuals. Among 226 dominant autotrophic and heterotrophic OTUs, 209 displayed 296 associations (Figure 2). From these associations between OTUs in the co-occurrence network, more positive (98.3%) than negative associations were inferred. All negative associations occurred between OTUs assigned to Ciliophora for heterotrophic microorganisms and to Chlorophyta, Chrysophyte and Cryptophyta for autotrophic microorganisms. No clear association patterns can be identified in the co-occurrence network from raw tables of autotrophic and heterotrophic microbial eukaryotes (Figure 2a). When the co-occurrence network is reordered according to OTU scores on the first sCoCA axis, two modules can be distinguished (Figure 2b). The first module (i.e., top right corner) consists of OTUs exhibiting higher relative abundances in the brook system (i.e., RSA), while OTUs that comprise the second module (i.e., bottom left corner) dominate pond systems (i.e., EV, LC, MG and SR) (Figure 2b–d). Heterotrophic OTUs exhibited major associations with autotrophic OTUs belonging to the same module, with only 1.7% of associations between OTUs from a distinct module. A striking pattern is that Chrysophyte is the autotrophic group that contributes most to associations in the two modules (module 1:71%, module 2:41%), whereas for the heterotrophic group it is fungi in module 1 (47%) and Ciliophora in module 2 (59%). In pond systems (i.e., module 2), surprisingly, fungi are involved in very few associations (7%).

3.2 | Case study 2: Microalgae–virus system

The cross-validation procedure identified the best pCoCA model based on the first two significant axes (pCoCA axis 1: p = .001, pCoCA axis 2: p = .001), in which the Mamiellophyceae community predicted 32.02% of the variation in the *Prasinovirus* community. The first two axes accounted for 37.47% (24.26% and 13.21% for axis 1 and 2 respectively) and 44.82% (28.01% and 16.81%) of the variation in the structure of the Mamiellophyceae and *Prasinovirus* community respectively. The biplots indicated that the two communities covaried along a lagoon (LA samples)/open-sea (MA samples) gradient on the first axis (from left to right), while a temporal gradient for site LA (intrasite variability) could be identified along the second axis (Figure 3). OtVs have a higher prevalence in the lagoon samples (especially LAs) and coastal samples (SAs) in which *Ostreococcus* exhibited a high density (Figure 3a,b). Conversely, open-sea samples (MAs) were dominated by *Bathycoccus*, which supported virus assemblages dominated by BpVs. *Micromonas* showed a wider distribution, with a relative contribution of its OTUs in both lagoon samples, coastal samples and open-sea samples, associated with a similar repartition of MpVs (Figure 3a,b).

Based on the cross-group analysis of co-occurrence networks, 62 associations were identified between the 67 major OTUs assigned to one of the three groups of Mamiellophyceae (i.e., *Bathycoccus*, *Micromonas* and *Ostreococcus*) and the major 98 OTUs assigned to *Prasinovirus* (i.e., BpVs, MpVs and OtVs) (Figure 4a). Reordering the co-occurrence network according to the OTU scores on the first pCoCA axis highlighted a structure in the co-occurrence network (Figure 4b). The network topology suggests that the identity of OTUs contained in co-occurring groups of viruses and microalgae are related to their respective prevalence along the lagoon/open-sea gradient. Virus OTUs mostly present in lagoon samples have significant associations primarily with microalgae OTUs that had a higher prevalence in lagoon samples (top right corner, Figure 4b–d). Similarly, virus OTUs dominating the open-sea samples were mainly associated with microalgae OTUs from open-sea samples (bottom left corner, Figure 4b–d). Among the associations contained in the co-occurrence network, 46 were identified between OTUs belonging to an expected host–virus system (i.e., associations *Bathycoccus*/BpV, *Micromonas*/MpV and *Ostreococcus*/OtV) while the other 16 significant associations were found between OTUs belonging to different host–virus systems. On average, 62.5%, 74.2% and 82.6% of associations found for OTUs of BpVs, MpVs and OtVs respectively were with OTUs assigned to their respective host group. Within the associations, some single Mamiellophyceae OTUs were associated with many *Prasinovirus* OTUs and vice versa. On the other hand, dyads (i.e., specific associations) were identified in *Bathycoccus*/BpV, *Micromonas*/MpV and *Ostreococcus*/OtV systems. A few negative associations inferred from the observation that those OTUs do not co-occur were found in network (Figure 4b). Interestingly, nine negative associations from 10 in total total involved virus OTUs and microalgae OTUs belonging to different host–virus systems.

4 | DISCUSSION

Critical review and guidance papers on the analysis of NGS-based community data (Buttigieg & Ramette, 2014; Paliy & Shankar, 2016; Ramette, 2007) do not mention any direct quantitative method for predicting the composition of one community from another. CoCA (ter Braak & Schaffers, 2004) fills this gap. At the level of case study 1, a symmetric form of CoCA indicated that heterotrophic microbial eukaryote assemblages in shallow freshwater ecosystems were

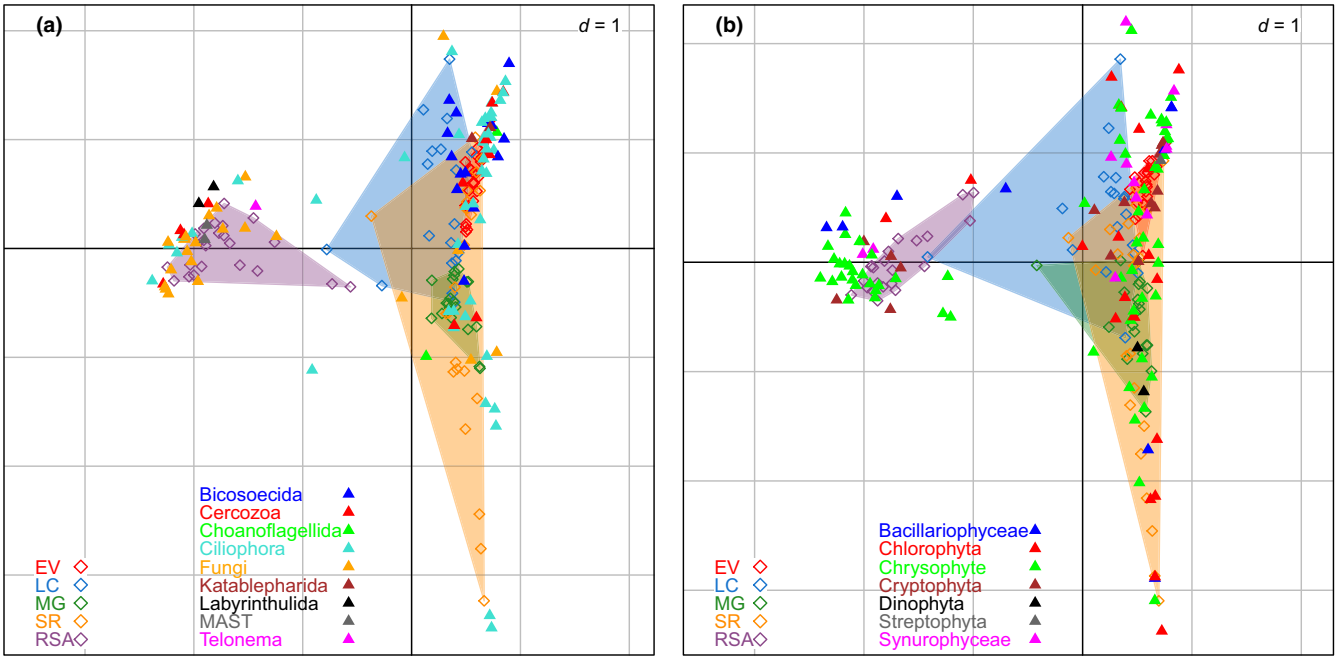

FIGURE 1 Ordination biplots of case study 1, representing the positions of sites (open diamond) and OTUs (solid triangle) on the first two axes of the symmetric co-correspondence analysis. Representation of (a) the heterotrophic microbial eukaryote data and (b) the autotrophic microbial eukaryote data. EV, Etang des Vallées; GB, Mare Gabard; LC, La Claye; RSA, Ru Sainte Anne ; SR, Saint Robert. Heterotrophic and autotrophic microbial eukaryote OTUs are coloured according to the phylogenetic group they belong to. "d" indicates the mesh of the grid. [Colour figure can be viewed at wileyonlinelibrary.com]

strongly associated with patterns of autotrophic microbial eukaryotes occurrence with links that can be taxon-specific (Figure 1). Our results also show that the composition of the heterotrophic microbial eukaryote community was dominated by fungi in the brook system compared to ponds. This is in line with recent observations, based on the estimation of ergosterol level, of a generally higher fungal biomass in rivers than in ponds (Baldy, Chauvet, Charcosset, & Gessner, 2002). A higher fungal abundance might potentially be linked to incoming resources from runoff, because in brooks the most important source of imported material is usually deciduous leaves, the decomposition of which to a large extent involved fungi (Bärlocher, 1985; Webster & Benfield, 1986). The composition of the heterotrophic microbial eukaryote community characterizing the brook system is associated with a specific composition of microbial autotrophs. This suggests that heterotroph community composition exerts a control on the composition of the autotroph community, and that microbial autotrophs can be drivers of microbial heterotrophs.

Regarding case study 2, the predictive form of CoCA indicates different distribution patterns among the three groups of *Prasinovirus* along the lagoon/open-sea gradient (Figure 3). Note the importance of the dimension reduction step in pCoCA that allows us to focus on ecological structures depicted on a limited number of axes and removes random variation from the data. The patterns in *Prasinovirus* assemblages, with a dominance of OtV in lagoon and coastal samples compared to offshore locations, the inverse distribution for BpV and MpV exhibiting a wider spatial distribution are in part a consequence of the presence of their respective hosts in lagoon, coastal and open-sea samples. Indeed, *Ostreococcus* is known to be abundant in lagoons (Subirana et al., 2013), a more eutrophic system, compared to *Bathycoccus*, which is found mainly in oligotrophic areas (Vaulot et al., 2012; Wu et al., 2013) such as offshore sites (i.e., MA). *Micromonas* is ubiquitous and particularly present in nutrient-rich environments (Not et al., 2004; Viprey, Guillou, Ferréol, & Vaulot, 2008). Our findings confirm also the data of Bellec, Grimsley, Derelle, Moreau, and Desdevises (2010) showing that OtVs are more abundant in lagoon than in the open sea.

Cross-taxon congruence description and evaluation (Gioria, Bacaro, & Feehan, 2011; Virtanen, Ilmonen, Paasivirta, & Muotka, 2009) provides a more comprehensive picture of the community similarity than the richness metrics conventionally used (Westgate, Barton, Lane, & Lindenmayer, 2014; Wolters, Bengtsson, & Zaitsev, 2006). Our results reinforce the need to use CoCA to study the cross-taxon congruence in microbial communities from NGS data. This is all the more important because the high co-correspondence between the two functional groups in microbial eukaryote community may be especially informative given the key ecological role of microbial eukaryotes (Caron, Countway, Jones, Kim, & Schnetzer, 2012). In addition, the study of cross-taxon congruence between Mamiellophyceae and *Prasinovirus* is of particular interest in marine ecosystems, warmed by climate change, where the expected gradual shift towards small primary producers could render the role of small eukaryotes more important than they are today (Morán, López-Urrutia, Calvo-Díaz, & Li, 2010). Microbial eukaryotes are recognized as a significant contributor across various geographical locations of picophytoplankton (Jardillier, Zubkov, Pearman, & Scanlan, 2010; Worden, Nolan, & Palenik, 2004), which accounts for > 50%

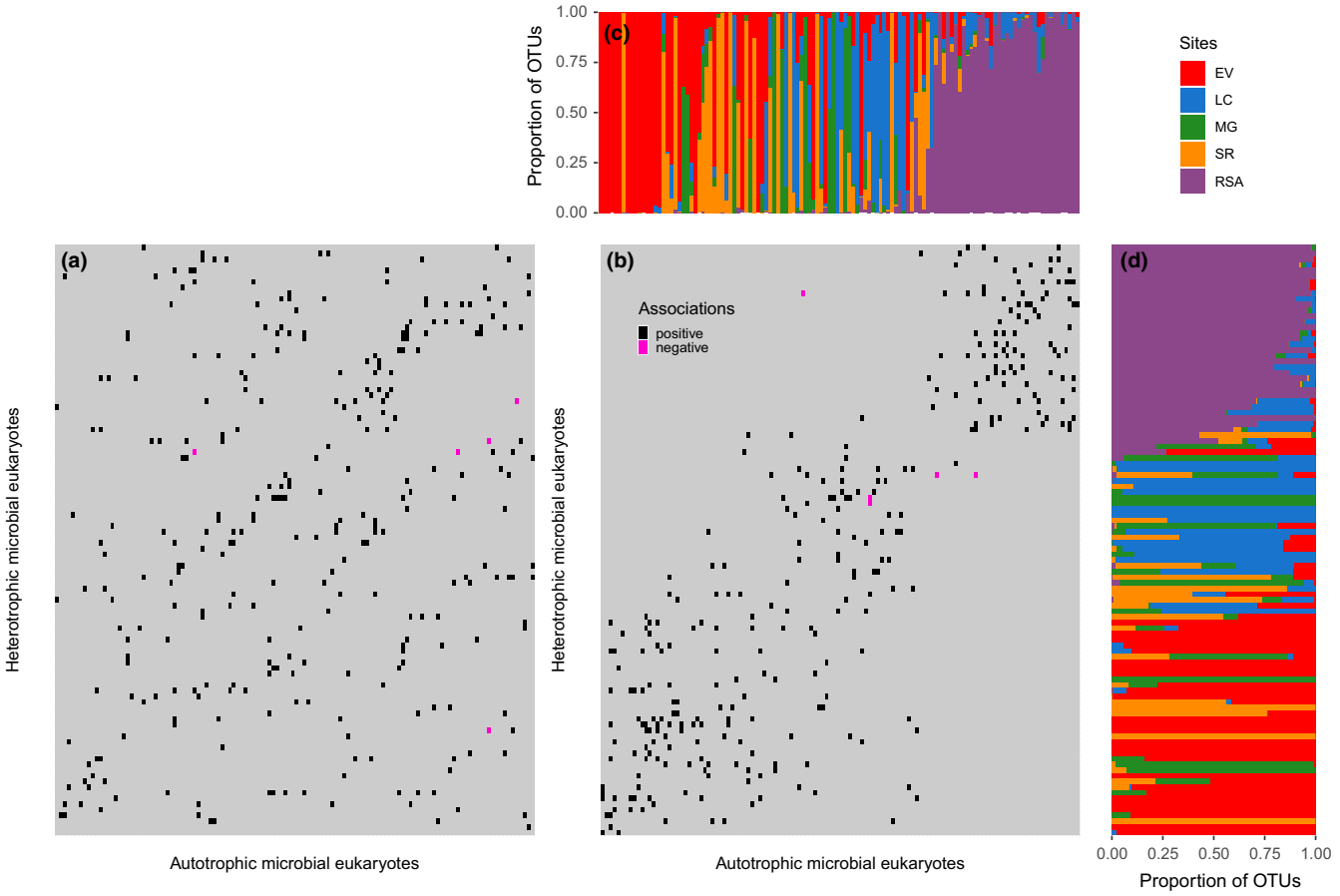

FIGURE 2 Heatmap of case study 1, representing the co-occurrence network between the heterotrophic and autotrophic microbial eukaryotes (a) before and (b) after reordering the position of each OTU in the network according to their scores on the first axis of symmetric co-correspondence analysis. Bar plots of the relative abundance of (c) autotrophic microbial eukaryotes and (d) heterotrophic microbial eukaryotes. Each OTU is represented by a vertical line partitioned into segments corresponding to its relative abundance in one of five sites. EV, Etang des Vallées; GB, Mare Gabard; LC, La Claye; RSA, Ru Sainte Anne; SR, Saint Robert. [Colour figure can be viewed at wileyonlinelibrary.com]

of phytoplankton biomass and productivity in marine ecosystems (Maranon et al., 2001; Teira et al., 2005).

Previous studies have suggested that biotic interactions are the most likely mechanisms underlying cross-taxon congruence at local scales (Jackson & Harvey, 1993; Johnson & Hering, 2010), although concordance is also expected from similar responses to environmental gradients (Bini, Vieira, Machado, & Machado Velho, 2007; Rooney & Bayley, 2012). An important implication is that the level of congruence can inform us about the structural pattern among interacting groups (Özkan et al., 2014). That being said, reordering co-occurrence networks, from the OTU scores on the first CoCA axis, allows us to identify structural patterns in co-occurrence networks of microbial eukaryote community (Figure 2b) and in the Mamiellophyceae/*Prasinovirus* system (Figure 4b). For example, the structure of the microbial eukaryote network is characterized by two modules underlying a brook/pond gradient in the composition of heterotrophic and autotrophic microbial eukaryote assemblages. These differences in OTU composition within the two modules suggest that the food-web structure is different between lotic and lentic ecosystems, and reinforce the differences previously observed at the bacterioplankton level (Portillo, Anderson, & Noah, 2012). Substantial effort has been made in the development of metrics to estimate and test the level of nestedness (e.g., Rodrigeuz-Gironés & Santamaría, 2006; Ulrich & Gotelli, 2007) and modularity (e.g., Barber, 2007; Dorman & Strauss, 2014) within interaction networks. The order of individuals of two groups in bipartite matrices affects the magnitude of metrics that represent deviations from an idealized state (e.g., perfect nestedness or modularity) (Almeida-Neto, Guimarães, Loyola, & Ulrich, 2008). It has been advocated that prior to analysis of the structure of networks, original bipartite matrices should be reordered to maximize the coherence of individual distributions in rows and columns, so that individuals with most similar links are close together (Borgatti & Everett, 1997; Leibold & Mikkelson, 2002). Together with our results, these findings validate our proposition to combine CoCA with network analysis to study structural patterns of microbial networks. In addition, in contrast to the expected view of the nestedness structure of phage–bacteria networks (Flores, Meyer, Valverde, Farr, & Weitz, 2011), the modular structure of the Mamiellophyceae/*Prasinovirus* network observed

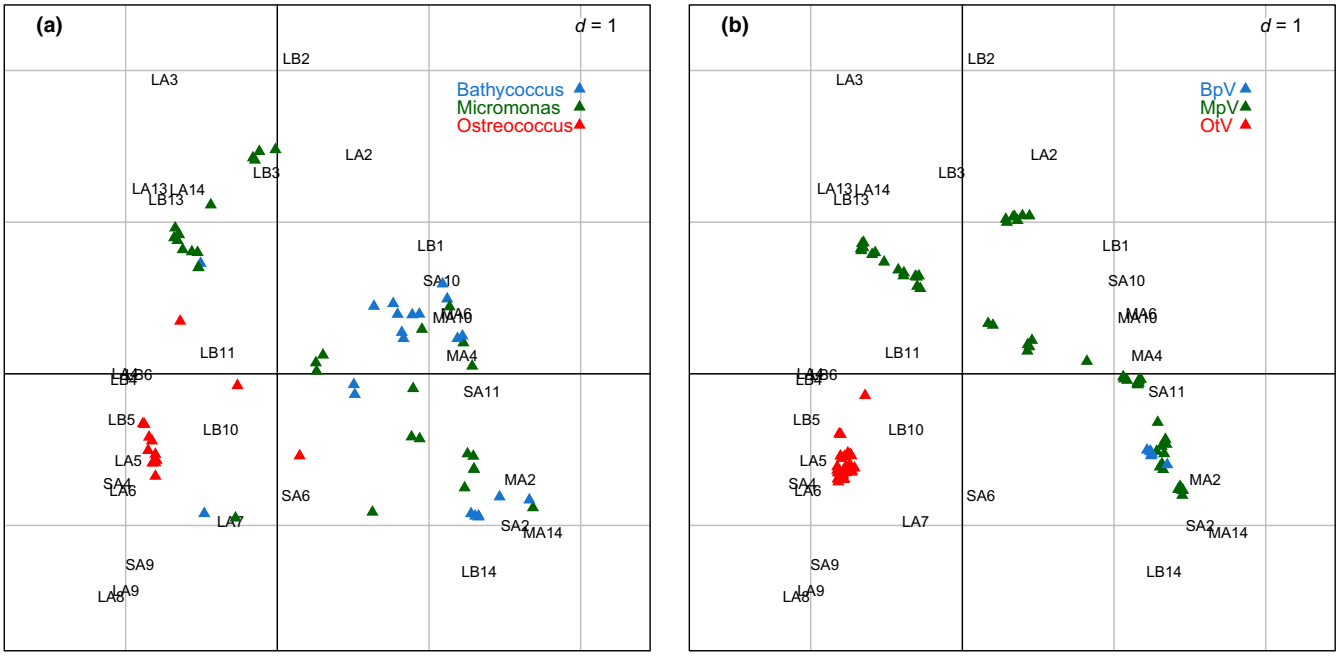

FIGURE 3 Ordination biplots of case study 2, representing the positions of sites (open diamond) and OTUs (solid triangle) on the first two axes of the predictive co-correspondence analysis. Representation of (a) Mamiellophyceae and (b) *Prasinovirus* data. LA: site at the level of the Grau in Leucate lagoon, LB: coastal site in Leucate lagoon, SA: coastal site, MA: open-sea site. Microalgae and virus OTUs are coloured according to the phylogenetic group they belong to. BpV, *Bathycoccus* viruses; MpV, *Micromonas* viruses; OtV, *Ostreococcus* viruses. "d" indicates the mesh of the grid. [Colour figure can be viewed at wileyonlinelibrary.com]

in the field underpinned the modularity patterns previously observed in phage–bacteria networks from cross-infection experiments (Flores, Valverde, & Weitz, 2013). Given that the structure of interaction networks is constrained by the coevolutionary processes between species (Peralta, 2016), this would mean that we need to take phylogenetic signals into account within co-occurrence networks (Derocles et al., 2018). In this context, it would be possible to disentangle the confounding effect of phylogeny from true biotic interactions by developing a partial analysis (ter Braak, Šmillauer, & Dray, 2018) in the context of CoCA to partial out the phylogenetic effect and focus on patterns of co-occurrence that are not related to phylogenetic signal.

All computational methods used to infer networks from NGS data sets produce species co-occurrence networks, where a link between two species represents a significant statistical association (positive or negative) between their abundance (or proportion). This raises a critical issue about the interpretation of inferred associations (Derocles et al., 2018), because co-occurrence networks differ from interaction networks constructed on observations of both the species and their interactions (Ings et al., 2009). For instance, all inferred associations between Mamiellophyceae and *Prasinovirus* belonging to the expected host–virus system were positive. These results are consistent with a previous study showing that when parasitism is captured as a significant link in a co-occurrence network, it is retrieved as a positive link despite the detrimental effect of the parasite on its host (Weiss et al., 2016). This might be explained because the co-presence of the host species and the parasite species is necessary for the interaction to occur. Another surprising result is that all negative associations (expected one) were between Mamiellophyceae and *Prasinovirus* belonging to different host–virus systems. Such negative associations may account for opposite abiotic requirements, because in our case, OTUs of concerned viruses and microalgae had inverse spatiotemporal dynamics. Positive associations were also found between individuals of different host–virus systems, which could be explained by the increase of a *Prasinovirus* population that removes, by infection, a major competitor of a co-occurring Mamiellophyceae host of another group. It is important to keep in mind that, although these associations between microalgae and *Prasinovirus* suggest that they interact, they do not necessarily mean that the co-occurring Mamiellophyceae are the virus hosts, even if they belong to the expected group of hosts (e.g., *Bathycoccus*, *Micromonas* or *Ostreococcus*). Our approach (CoCA combined with network analysis, or another method to infer associations) could be combined with other approaches, such as single cell genomics (Kalisky, Baliney, & Quake, 2011; Martinez-Garcia et al., 2012) or Epic-PCR (Spencer et al., 2016), to validate the predicted associations in interactions. The justification of the association sign between the two functional groups making up the microbial eukaryote community is also not straightforward, although they can be triggered by ecological interactions or by species abiotic requirements (Derocles et al., 2018).

In conclusion, the successful application of CoCA over two real data sets of microbial communities exhibiting various types of interactions reinforces that this method for studying the distributions, assemblages and interactions between two microbial

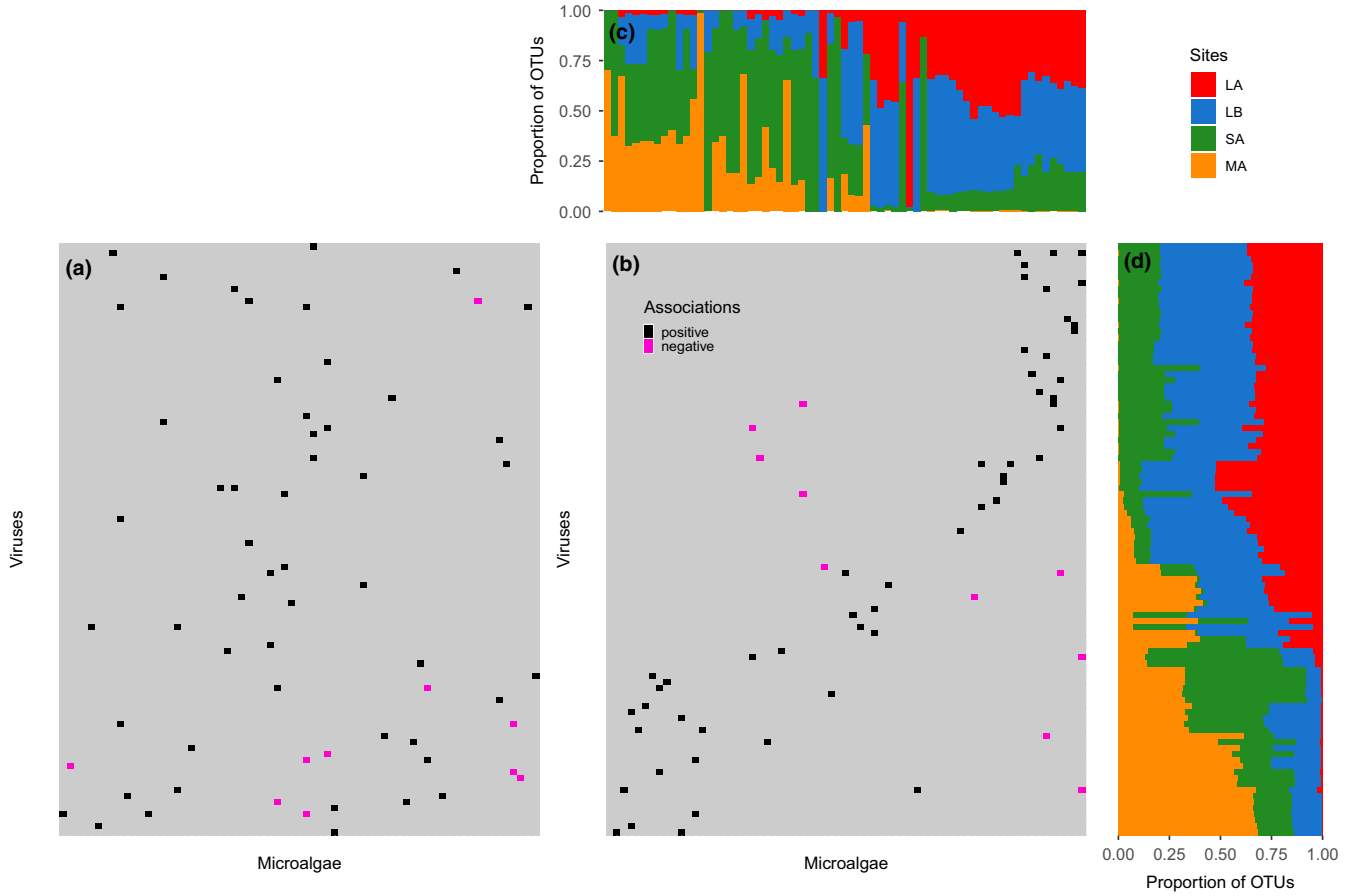

FIGURE 4 Heatmap of case study 2, representing the microalgae–virus (Mamiellophycaea/*Prasinovirus*) co-occurrence network (a) before and (b) after reordering the position of each OTU in the network according to their scores on the first axis of predictive co-correspondence analysis. Bar plots of the relative abundance of (c) microalgae and (d) viruses. Each OTU is represented by a vertical line partitioned into segments corresponding to its relative abundance in one of four sites. LA, site at the level of the Grau in Leucate lagoon; LB, coastal site in Leucate lagoon; SA, coastal site; MA, open-sea site. [Colour figure can be viewed at wileyonlinelibrary.com]

communities constitutes a highly valuable approach to understand the cross-taxon congruence between microorganisms. A useful consequence of cross-taxon congruence is that the distribution of well-known taxa may provide insight into the processes structuring the distribution of other taxa (e.g., Bilton, McAbendorth, Bedford, & Ramsay, 2006; Santi et al., 2010). This approach could be used to enhance our understanding of a major problem, namely the effect of phytoplankton blooms (in particular toxic groups such as cyanobacteria) on the microbial communities and in turn on the ecosystem functions (e.g., Xue et al., 2018; Yang et al., 2016). Our findings also demonstrate that the re-ordering of co-occurrence networks, according to the congruence information extracted from CoCA, allows us to investigate the presence of ecological signals in networks. The advantage of this approach is that the complexity of the network is considerably reduced by the nonrandom placement of nodes in the space in such a way as to improve the aesthetic quality of the representation and consequently its readability, as proposed in good practice of data visualization (Kjærgaard, 2015; Pocock et al., 2016; Spiegelhalter, Pearson, & Short, 2011). Interestingly, the applicability of our approach goes beyond the particular case of data sets with row-sum normalization (i.e., compositional data). Indeed, CoCA was originally designed to analyse abundance data and is thus able to deal with counts without the need to rarefy data, in accordance with the recent advice against rarefaction (McMurdie & Holmes, 2014). It paves the way for further studies to examine the cross-taxon congruence and structural pattern of co-occurrence networks in microbial communities and in turn their effects on ecosystem functioning.

ACKNOWLEDGEMENTS

We are grateful for financial support from the French National Research Agency (DECOVIR ANR-12-BSV7-0009, coordinator Y.D.). This work benefited also from the support of the EcoNet ANR-18-CE02-0010-01 programme to S.D. We also thank Ludwig Jardillier for access to genomic data on autotrophic and heterotrophic microbial eukaryotes.

AUTHOR CONTRIBUTIONS

B.A., Y.D. and S.D. designed the study; B.A. performed all the statistical analyses; C.J.F.B. contributed to mathematical development of the method; H.L. contributed to collecting and genotyping the biological materials of microalgae and viruses; B.A. wrote the first

draft of the manuscript; B.A., C.J.F.B., Y.D. and S.D. commented and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Data and R scripts to reproduce the different analyses of case study 1 and case study 2 are available in the online tutorial.

ORCID


Benjamin Alric https://orcid.org/0000-0003-2774-0546

REFERENCES

Almeida-Neto, M., Guimarães, P. R., Jr, Loyola, R. D., & Ulrich, W. (2008). A consistent metric for nestedness analysis in ecological systems: Reconciling concept and measurement. *Oikos*, *117*, 1127–1239. https://doi.org/10.1111/j.0030-1299.2008.16644.x

Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., & Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE*, *4*, e6372.

Baldy, V., Chauvet, E., Charcosset, J.-Y., & Gessner, M. O. (2002). Microbial dynamics associated with leaves decomposing in the mainstem and floodplain pond of a large river. *Aquatic Microbial Ecology*, *28*, 25–36. https://doi.org/10.3354/ame028025

Bálint, M., Bahram, M., Eren, A. M., Faust, K., Fuhrman, J. A., Lindahl, B., … Tedersoo, L. (2016). Millions of reads, thousands of taxa: Microbial community structure and associations analyzed via marker genes. *FEMS Microbial Reviews*, *40*, 686–700. https://doi.org/10.1093/femsre/fuw017

Ban, Y., An, L., & Jiang, H. (2015). Investigating microbial co-occurrence patterns based on metagenomics compositional data. *Bioinformatics*, *31*, 3322–3329.

Banerjee, S., Kirkby, C. A., Schmutter, D., Bissett, A., Kirkegaard, J. A., & Richardson, A. E. (2016). Network analysis reveals functional redundancy and keystone taxa amongst bacterial and fungal communities during organic matter decomposition in an arable soil. *Soil Biology and Biochemistry*, *97*, 188–198. https://doi.org/10.1016/j.soilbio.2016.03.017

Barber, M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review*, *76*, 066102. https://doi.org/10.1103/PhysRevE.76.066102

Bärlocher, F. (1985). The role of fungi in the nutrition of stream invertebrates. *Botanical Journal of the Linnean Society*, *91*, 83–94. https://doi.org/10.1111/j.1095-8339.1985.tb01137.x

Bellec, L., Grimsley, N., Derelle, E., Moreau, H., & Desdevises, Y. (2010). Abundance, spatial distribution and genetic diversity of *Ostreococcus tauri* viruses in two different environments. *Environmental Microbiology Reports*, *2*, 313–321.

Bergelson, J., Mittelstrass, J., & Horton, M. W. (2019). Characterizing both bacteria and fungi improves understanding of the *Arabidopsis* root microbiome. *Scientific Reports*, *9*, 24. https://doi.org/10.1038/s41598-018-37208-z

Bilton, D. T., McAbendorth, L., Bedford, A., & Ramsay, P. M. (2006). How wide to cast the net? Cross-taxon congruence of species richness, community similarity and indicator taxa ponds. *Freshwater Biology*, *51*, 578–590.

Bini, L. M., Vieira, L. C. G., Machado, J., & Machado Velho, L. F. (2007). Concordance of species patterns among micro-crustaceans, rotifers and testate amoebae in a shallow pond. *International Review of Hydrobiology*, *92*, 9–22.

Borgatti, S. P., & Everett, M. G. (1997). Network analysis of 2-mode data. *Social Networks*, *19*, 243–269. https://doi.org/10.1016/S0378-8733(96)00301-2

Buttigieg, P. L., & Ramette, A. (2014). A guide to statistical analysis in microbial ecology: A community-focused, living review of multivariate data analyses. *FEMS Microbiology Ecology*, *90*, 543–550. https://doi.org/10.1111/1574-6941.12437

Cannon, M. V., Craine, J., Hester, J., Shalkhauser, A., Chan, E. R., Logue, K., … Serre, D. (2017). Dynamic microbial populations along the Cuyahoga River. *PLoS ONE*, *12*, e0186290. https://doi.org/10.1371/journal.pone.0186290

Cardona, C., Weisenhorn, P., Henry, C., & Gilbert, J. A. (2016). Network-based metabolic analysis and microbial community modeling. *Current Opinion in Microbiology*, *31*, 124–131. https://doi.org/10.1016/j.mib.2016.03.008

Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y., & Schnetzer, A. (2012). Marine protistan diversity. *Annual Review of Marine Science*, *4*, 467–493. https://doi.org/10.1146/annurev-marine-120709-142802

Chen, F., & Suttle, C. A. (1995). Amplification of DNA polymerase gene fragments from viruses infecting microalgae. *Applied and Environmental Microbiology*, *61*, 1274–1278.

Clerissi, C., Grimsley, N., Ogata, H., Hingamp, P., Poulain, J., & Desdevises, Y. (2014). Unveiling of the diversity of Prasinoviruses (*Phycodnaviridae*) in marine samples by using high-throughput sequencing analyses of PCR-amplified DNA polymerase and major capsid protein genes. *Applied and Environmental Microbiology*, *80*, 3150–3160. https://doi.org/10.1128/AEM.00123-14

Cordero, O. X., & Datta, M. (2016). Microbial interactions and community assembly at microscales. *Current Opinion in Microbiology*, *31*, 227–234. https://doi.org/10.1016/j.mib.2016.03.015

Derocles, S. A. P., Bohan, D. A., Dumbrell, A. J., Kitson, J. J. N., Massol, F., Pauvert, C., … Evans, D. M. (2018). Biomonitoring for the 21st century: Integrating next-generation sequencing into ecological network analysis. *Advances in Ecological Research*, *58*, 3–62.

Dolédec, S., & Chessel, D. (1994). Co-inertia analysis: An alternative method for studying species-environment relationships. *Freshwater Biology*, *31*, 277–294. https://doi.org/10.1111/j.1365-2427.1994.tb01741.x

Dormann, C. F., & Strauss, R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, *5*, 90–98. https://doi.org/10.1111/2041-210X.12139

Dray, S., Chessel, D., & Thioulouse, J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, *84*, 3078–3089. https://doi.org/10.1890/03-0178

Ducklow, H. (2008). Microbial services: Challenges for microbial ecologists in a changing world. *Aquatic Microbial Ecology*, *53*, 13–19. https://doi.org/10.3354/ame01220

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*, 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*, 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, *320*, 1034–1038. https://doi.org/10.1126/science.1153213

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

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PloS Computational Biology*, *8*, e1002602. https://doi.org/10.1371/journal.pcbi.1002606

Fierer, N., Breitbart, M., Nulton, J., Salamon, P., Lozupone, C., Jones, R., … Jackson, R. B. (2007). Metagenomic and small-subunit rRNA analyses reveal the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*, *73*, 7059–7066. https://doi.org/10.1128/AEM.00358-07

Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., & Weitz, J. S. (2011). Statistical structure of host-phage interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*, E288–E297. https://doi.org/10.1073/pnas.1101595108

Flores, C. O., Valverde, S., & Weitz, J. S. (2013). Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages. *ISME Journal*, *7*, 520–532. https://doi.org/10.1038/ismej.2012.135

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, *8*, e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Fuhrman, J. A. (2009). Microbial community structure and its functional implications. *Nature*, *459*, 193–199. https://doi.org/10.1038/nature08058

Gauch, G. H. G., Jr, Whittaker, R. H., & Wentworth, T. R. (1977). A comparative study of reciprocal averaging and other ordination techniques. *Journal of Ecology*, *65*, 157–174. https://doi.org/10.2307/2259071

Genitsaris, S., Monchy, S., Breton, E., Lecuyer, E., & Christaki, U. (2016). Small-scale variability of protistan planktonic communities relative to environmental pressures and biotic interactions at two adjacent coastal stations. *Marine Ecology Progress Series*, *548*, 61–75. https://doi.org/10.3354/meps11647

Gioria, M., Bacaro, G., & Feehan, J. (2011). Evaluation an interpretating cross-taxon congruence: Potential pitfalls and solutions. *Acta Oecologia*, *37*, 187–194.

Graham, E. B., Knelman, J. E., Schindlbacher, A., Siciliano, S., Breulmann, M., Yannarell, A., … Nemergut, D. R. (2016). Microbes as engines of ecosystem function: When does community structure enhance predictions of ecosystem processes? *Frontiers in Microbiolgy*, *7*, 214. https://doi.org/10.3389/fmicb.2016.00214

Greenacre, M. (2009). Power transformation in correspondence analysis. *Computational Statistics and Data Analysis*, *53*, 3107–3116.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Christen, R. (2013). The protest Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*, D597–D604.

Handelsman, J. (2004). Metagenomics: Application of genomics to uncultured micro-organisms. *Microbiology and Molecular Biology Reviews*, *68*, 669–685. https://doi.org/10.1128/MMBR.68.4.669-685.2004

Hutchins, D. A., & Fu, F. (2017). Microorganisms and ocean global change. *Nature Microbiology*, *2*, 17058. https://doi.org/10.1038/nmicrobiol.2017.58

Ings, T. C., Montoya, J. M., Bascompte, J., Blüthgen, N., Brown, L., Dormann, C. F., … Woodward, G. (2009). Ecological network – beyond food webs. *Journal of Animal Ecology*, *78*, 253–269.

Jackson, D. A. (1997). Compositional data in community ecology: The paradigm or peril of proportions? *Ecology*, *78*, 929–940. https://doi.org/10.1890/0012-9658(1997)078[0929:CDICET]2.0.CO;2

Jackson, D. A., & Harvey, H. H. (1993). Fish and benthic invertebrates: Community concordance and community-environment relationships. *Canadian Journal of Fisheries and Aquatic Sciences*, *50*, 2641–2651. https://doi.org/10.1139/f93-287

Jardillier, L., Zubkov, M. V., Pearman, J., & Scanlan, D. J. (2010). Significant CO_2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME Journal*, *4*, 1180–1192. https://doi.org/10.1038/ismej.2010.36

Johnson, R. K., & Hering, D. (2010). Spatial congruency of benthic diatom, invertebrate, macrophyte, and fish assemblages in European streams. *Ecological Applications*, *20*, 978–992. https://doi.org/10.1890/08-1153.1

Kalisky, T., Baliney, P., & Quake, S. R. (2011). Genomics analysis at the single-cell level. *Annual Review of Genetics*, *45*, 431–445.

Kjærgaard, R. S. (2015). Data visualization: Mapping the topical space. *Nature*, *520*, 292–293. https://doi.org/10.1038/520292a

Kueneman, J. G., Woodhams, D. C., Treuren, W. V., Archer, H. M., Kniht, R., & McKenzie, V. J. (2016). Inhibitory bacteria reduce fungi on early life stages of endangered Colorado boreal toads (*Anaxyrus boreas*). *ISME Journal*, *10*, 934–944. https://doi.org/10.1038/ismej.2015.168

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, *11*, e1004226. https://doi.org/10.1371/journal.pcbi.1004226

Legendre, L., & Legendre, P. (2012). *Numerical ecology*. Amsterdam, NL: Elsevier.

Leibold, M. A., & Mikkelson, G. M. (2002). Coherence, species turnover, and boundary clumping: Elements of meta-community structure. *Oikos*, *97*, 237–250. https://doi.org/10.1034/j.1600-0706.2002.970210.x

Li, H. (2015). Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Applications*, *2*, 73–94.

Little, A. E., Robinson, C. J., Peterson, S. B., Raffa, K. F., & Handelsman, J. (2008). Rules of engagement: Interspecies interactions that regulate microbial communities. *Annual Review of Microbiology*, *62*, 375–401. https://doi.org/10.1146/annurev.micro.030608.101423

Liu, H., Roeder, K., & Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Advances in Neural Information Processing Systems*, *23*, 1432–1440.

Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., … Gilbert, J. A. (2016). Geographic patterns of co-occurrence network topological features for soil microbiota at continental scale in eastern China. *ISME Journal*, *10*, 1891–1901.

Maranon, E., Holligan, P. M., Barciela, R., Gonzalez, N., Mourino, B., Pazo, M. J., & Varela, M. (2001). Patterns of phytoplankton size structure and productivity in contrasting open-ocean environments. *Marine Ecology Progress Series*, *216*, 43–56. https://doi.org/10.3354/meps216043

Martens, H. (2001). Reliable and relevant modelling of real world data: A personal account of the development of PLS Regression. *Chemometrics and Intelligent Laboratory Systems*, *58*, 85–95. https://doi.org/10.1016/S0169-7439(01)00153-8

Martinez-Garcia, M., Brazel, D., Poulton, N. J., Swan, B. K., Gomez, M. L., Masland, D., … Stepanauskas, R. (2012). Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *ISME Journal*, *6*, 703–707. https://doi.org/10.1038/ismej.2011.126

McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: Why rarefying microbiome data is inadmissible. *PLoS Computational Biology*, *10*, e1003531. https://doi.org/10.1371/journal.pcbi.1003531

Morán, X. A., López-Urrutia, A., Calvo-Díaz, A., & Li, W. K. W. (2010). Increasing importance of small phytoplankton in a warmer ocean. *Global Change Biology*, *16*, 1137–1144. https://doi.org/10.1111/j.1365-2486.2009.01960.x

Ni, J., Yan, Q., & Yu, Y. (2013). How much metagenomic sequencing is enough to achieve a given goal? *Scientific Reports*, *3*, 1968. https://doi.org/10.1038/srep01968

Not, F., Latasa, M., Marie, D., Cariou, T., Vaulot, D., & Simon, N. (2004). A single species, *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the western English Channel. *Applied and Environmental Microbiology*, *70*, 4064–4072. https://doi.org/10.1128/AEM.70.7.4064-4072.2004

Özkan, K., Jeppesen, E., Davidson, T. A., Søndergaard, M., Lauridsen, T. L., Bjerring, R., … Svenning, J.-C. (2014). Cross-taxon congruence in lake plankton largely independent of environmental gradients. *Ecology*, *95*, 2778–2788. https://doi.org/10.1890/13-2141.1

Paliy, O., & Shankar, V. (2016). Application of multivariate statistical techniques in microbial ecology. *Molecular Ecology*, *25*, 1032–1057. https://doi.org/10.1111/mec.13536

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*, 1200–1200. https://doi.org/10.1038/nmeth.2658

Peralta, G. (2016). Merging evolutionary history into species interaction networks. *Functional Ecology*, *30*, 1917–1925.

Pocock, M. J. O., Evans, D. M., Fontaine, C., Harvey, M., Julliard, R., McLaughlin, O., … Bohan, D. A. (2016). The visualization of ecological networks, and their use as a tool for engagement, advocacy and management. *Advances in Ecological Research*, *54*, 41–85.

Portillo, M. C., Anderson, S. P., & Noah, F. (2012). Temporal variability in the diversity and composition of stream bacterioplankton communities. *Environmental Microbiology*, *14*, 2417–2428. https://doi.org/10.1111/j.1462-2920.2012.02785.x

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramette, A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology*, *62*, 142–160. https://doi.org/10.1111/j.1574-6941.2007.00375.x

Rappé, M. S., & Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, *57*, 369–394. https://doi.org/10.1146/annurev.micro.57.030502.090759

Rodrigeuz-Gironés, M. A., & Santamaría, L. (2006). A new algorithm to calculate the nestednes temperature of presence-absence matrices. *Journal of Biogeography*, *33*, 924–935.

Rooney, R. C., & Bayley, S. E. (2012). Community congruence of plants, invertebrates and birds in natural and constructed shallow open-water wetlands: Do we need to monitor multiple assemblages? *Ecological Indicators*, *20*(2012), 42–50. https://doi.org/10.1016/j.ecolind.2011.11.029

Santi, E., Maccherini, S., Rocchini, D., Bonini, I., Brunialti, G., Favilli, L., … Chiarucci, A. (2010). Simple to sample: Vascular plants as surrogate group in a nature reserve. *Journal for Nature Conservation*, *18*, 2–11. https://doi.org/10.1016/j.jnc.2009.02.003

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., … Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*, 7537–7541. https://doi.org/10.1128/AEM.01541-09

Simon, M., Jardillier, L., Deschamps, P., Moriera, D., Restoux, G., Bertolino, P., & López-García, P. (2014). Complex communities of small protists and unexpected occurrence of typical marine lineages in shallow freshwater systems. *Environmental Microbiology*, *17*, 3610–3627. https://doi.org/10.1111/1462-2920.12591

Simon, M., López-García, P., Deschamps, P., Moreira, D., Restoux, G., Bertolino, P., & Jardillier, L. (2015). Marked seasonality and high spatial variability of protist communities in shallow freshwater systems. *ISME Journal*, *9*, 1941–1953. https://doi.org/10.1038/ismej.2015.6

Simpson, G. L. (2009). *Cocorresp: Co-correspondence analysis ordination methods. (R package version 0.4-0)*. Retrieved from http://cran.r-project.org/package=analoguehttp://cran.r-project.org/cocorresp/package=cocorresp

Spencer, S. J., Tamminen, M. V., Preheim, S. P., Guo, M. T., Briggs, A. W., Brito, I. L., … Alm, E. J. (2016). Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME Journal*, *10*, 427–436. https://doi.org/10.1038/ismej.2015.124

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future. *Science*, *333*, 1393–1400. https://doi.org/10.1126/science.1191181

Subirana, L., Péquin, B., Michely, S., Escande, M.-L., Meilland, J., Derelle, E., … Grimsley, N. H. (2013). Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus* reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist*, *164*, 643–659. https://doi.org/10.1016/j.protis.2013.06.002

Teira, E., Mouriño, B., Marañón, E., Pérez, V., Pazó, M. J., Serret, P., … Fernández, E. (2005). Variability of chlorophyll and primary production in the Eastern North Atlantic Subtropical Gyre: Potential factors affecting phytoplankton activity. *Deep-Sea Research Part I*, *52*, 569–588. https://doi.org/10.1016/j.dsr.2004.11.007

ter Braak, C. J. F., & Schaffers, A. P. (2004). Co-correspondence analysis: A new ordination method to relate two community compositions. *Ecology*, *85*, 834–846. https://doi.org/10.1890/03-0021

ter Braak, C. J. F., Šmilauer, P., & Dray, S. (2018). Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecology Statistics*, *25*, 171–197. https://doi.org/10.1007/s10651-017-0395-x

Tipton, L., Müller, C. L., Kurtz, Z. D., Huang, L., Kleerup, E., Morris, A., … Ghedin, E. (2018). Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, *6*, 12. https://doi.org/10.1186/s40168-017-0393-0

Ulrich, W., & Gotelli, N. J. (2007). Null model analysis of species nestedness patterns. *Ecology*, *88*, 1824–1831. https://doi.org/10.1890/06-1208.1

Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., … Bohan, D. A. (2016). Learning ecological network from next-generation sequencing data. *Advances in Ecological Research*, *54*, 1–39.

Vaulot, D., Lepère, C., Toulza, E., De la Iglesia, R., Poulain, J., Gaboyer, F., … Piganeau, G. (2012). Metagenomes of the Picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS ONE*, *7*, e39648. https://doi.org/10.1371/journal.pone.0039648

Viprey, M., Guillou, L., Ferréol, M., & Vaulot, D. (2008). Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environmental Microbiology*, *10*, 1804–1822. https://doi.org/10.1111/j.1462-2920.2008.01602.x

Virtanen, R., Ilmonen, J., Paasivirta, L., & Muotka, T. (2009). Community concordance between bryophyte and insect assemblages in boreal springs: A broad-scale study in isolated habitats. *Freshwater Biology*, *54*, 1651–1662. https://doi.org/10.1111/j.1365-2427.2009.02212.x

Webster, J. R., & Benfield, E. F. (1986). Vascular plant breakdown in freshwater ecosystems. *Annual Review of Ecology and Systematics*, *17*, 567–594. https://doi.org/10.1146/annurev.es.17.110186.003031

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y. E., … Knight, R. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME Journal*, *10*, 1669–1681. https://doi.org/10.1038/ismej.2015.235

Westgate, M. J., Barton, P. S., Lane, P. W., & Lindenmayer, D. B. (2014). Global meta-analysis reveals low consistency of biodiversity congruence relationships. *Nature Communications*, *5*, 3899. https://doi.org/10.1038/ncomms4899

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., … 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. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Wolters, V., Bengtsson, J., & Zaitsev, A. S. (2006). Relationship among the species richness of different taxa. *Ecology*, *87*, 1886–1895. https://doi.org/10.1890/0012-9658(2006)87[1886:RATSRO]2.0.CO;2

Worden, A. Z., Nolan, J. K., & Palenik, B. (2004). Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography*, *49*, 168–179. https://doi.org/10.4319/lo.2004.49.1.0168

Wu, W., Huang, B., & Zhong, C. (2013). Photosynthetic picoeukaryote assemblages in the South China Sea from the Pearl River estuary to the SEATS station. *Aquatic Microbial Ecology*, *71*, 271–284.

Xue, Y., Chen, H., Yang, J. R., Liu, M., Huang, B., & Yang, J. (2018). Distinct patterns and processes of abundant and rare eukaryotic plankton communities following a reservoir cyanobacterial bloom. *ISME Journal*, *12*, 2263–2277. https://doi.org/10.1038/s41396-018-0159-0

Yang, C., Li, Y., Zhou, Y., Lei, X., Zheng, W., Tian, Y., … Zheng, T. (2016). A comprehensive insight into functional profiles of free-living microbial community responses to toxic *Akshiwo sanguinea* bloom. *Scientific Reports*, *6*, 34645.

Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeukaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology*, *52*, 79–92.

Zimmerman, N., Izard, J., Klatt, C., Zhou, J., & Aronson, E. (2014). The unseen world: Environmental microbial sequencing and identification methods for ecologists. *Frontiers in Ecology and the Environment*, *12*, 224–231. https://doi.org/10.1890/130055

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Alric B, ter Braak CJF, Desdevises Y, Lebredonchel H, Dray S. Investigating microbial associations from sequencing survey data with co-correspondence analysis. *Mol Ecol Resour.* 2020;20:468–480. https://doi.org/10.1111/1755-0998.13126