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

Benjamin Alric, Cajo J.F. ter Braak, Yves Desdevises, Hugo Lebredonchel, Stéphane Dray

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]*Université de Lyon, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Lyon1, F-69622 Villeurbanne, France*

[2]*Biometris, Wageningen University & Research, Box 16, 6700 AA Wageningen, The Netherlands*

[3]*Sorbonne Université, CNRS, UMR 7232, Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, F-66650 Banyuls sur Mer, France*

*Corresponding author: Benjamin ALRIC, Irstea, UR RiverLy, Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, 5 rue de la Doua CS 20244, F-69625, Villeurbanne, France. benjamin.alric@irstea.fr

†Current address: Irstea, UR RiverLy, Laboratoire d'écotoxicologie, centre de Lyon-Villeurbanne, 5 rue de la Doua CS 20244, F-69625, Villeurbanne, France.

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

Running tittle: Cross-taxon congruence in microbial communities

Abstract

Microbial communities, which drive the major ecosystem functions, are composed by a wide range of interacting species. Understanding how microbial communities are structured and the underlying processes is a crucial task for interpreting ecosystem response to global change but it 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 and lead to analyze a pair of tables. Here, an analytic framework based on co-correspondence analysis (CoCA) is proposed to study the distributions, assemblages and interactions between two microbial communities. We showed 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 eukaryotes assemblages, on one hand, and between microalgae and viruses, on the other hand. We demonstrate also how CoCA can be used, in complement to network analysis, to reorder co-occurrence networks and thus investigate the presence of patterns in ecological networks.

Introduction

Microbial communities are highly diverse (Rappé & Giovannoni, 2003) and drive the 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 a crucial task to predict 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 is still limited (Little et al., 2008; Cordero & Datta, 2016).

The emergence of high-throughput sequencing techniques (next-generation sequencing; NGS) gave access to the diversity of whole microbial communities, including the non-cultivable fraction (Handelsman, 2004; Zimmerman, Izard, Klatt, Zhou, & Aronson, 2014; Zinger, Gobet, & Pommier, 2011). 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 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, correspondence analysis), or after computing a distance matrix (e.g., non-metric 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 analyze a pair of tables (i.e., OTUs composition for the same sampling sites for two different taxonomic groups). To analyze such pair of tables, a common practice consists in merging the two tables into a single one and then applying network analysis (e.g., Kueneman et al., 2016; Banerjee et al., 2016; Ma et al., 2016) and/or multivariate analysis (Cannon et al., 2017; Bergelson, Mittelstrass, & Horton, 2019). 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 normalisz~~z~~ation 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 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., a non-independence of relative abundances induced by the row-sum normalisation) 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 (Ramette, 2007; Buttigieg & Ramette, 2014; Paliy & Shankar, 2016). However, these works 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 normalised (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 microbial communities. This framework is based on correspondence analysis, a method that effectively handles proportional data that contain many zeroes (Gauch et al., 1977; Jackson, 1997; Greenacre, 2009), like NGS data (Paulson, Stine, Bravo, & Pop et al., 2013). We show how this method allows to extract information about the co-structure among two microbial communities to estimate the congruence between them. Finally, we show that the outputs of the method can be used to reorder co-occurrence networks inferred by network analysis to enhance the visualization of microbial association and the understanding of assemblage patterns within networks. Hence, our approach echoes to the current lively debate about the practices to create network visualizations which

are both aesthetically appealing and have high information content (see Pocock et al., 2016 for review). 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.

Materials and methods

Studying cross-taxon congruence by co-correspondence analysis

Co-correspondence analysis (ter Braak & Schaffers, 2004) is part of the class of canonical analyses with the feature to be designed to analyze a pair of tables containing abundance data and to study the co-variations between two types of communities (e.g., plants and pollinators). CoCA is based on correspondence analysis (Benzécri, 1969; Hill, 1973) and preserves its fundamental properties of weighted averaging and the use of the chi-square (χ^2) distance for both tables (Supporting Information). The χ^2 distance is particularly adapted to NGS data as it handles properly zero values (and in particular double absences) and thus is not hampered by zero inflation (Legendre & Legendre, 2012).

Here, co-correspondence analysis 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. Let $X=[x_{ij}]$ and $Y=[y_{ik}]$ be $n \times p$ and $n \times q$ tables containing the relative abundances of each p and q species of two communities measured at the same n samples. In the first case, pCoCA is chosen because we wish to investigate the composition of virus communities under the hypothesis 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 wished to study the relationships between autotrophic and heterotrophic microbial eukaryotes. 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, Chessel, & Thioulouse, 2003). Co-inertia analysis relates two data tables in a symmetric way, by providing a decomposition of the co-inertia criterion on a set of orthogonal axes on which sample scores are projected (Dray, Chessel, & Thioulouse, 2003).

In practice, CoCA identifies associations (or common ecological gradients) between two types of biological assemblages from the same samples, by seeking the factorial axes that maximize the covariance between the weighted average sample scores (projection of rows of one table onto the factorial axes) of one community with those of the other community (Supporting Information, Eq. 12). This requires to determine species scores (projection of columns of one table onto the factorial axes) of one table as weighted averages of sample scores of the other table and sample scores as weighted averages of the species scores of their own table (Supporting Information, Eq. 13–16). So, pCoCA and sCoCA are a weighted version of PLS and COIA, respectively. Given that CoCA is related to the correspondence analysis, it is necessary to circumvent the fact that the sample weights of $\mathbf{Y}$ and $\mathbf{X}$ (R_1 and R_2, which are the row sums of $\mathbf{Y}$ and $\mathbf{X}$, respectively) are imposed and are not similar (Supporting Information, Eq. 6 and Eq. 7). This does not meet a crucial constraint in estimating the co-structure between two tables, namely that the samples must be weighted in the same way for the two tables

(Dray, Chessel & Thioulouse, 2003). Hence, an additional common sample weights matrix (R_0) is defined to replace R_1 and R_2 so that the weighted averaging properties of CA can be retained in CoCA (ter Braak & Schaffers, 2004). In pCoCA, R_0 is equal to the sample weights of the community response table (i.e., $R_0 = R_1$), whereas in sCoCA, $R_0 = (R_1 + R_2)/2$ (Supporting Information, Eq. 10 and Eq. 11). Note that in the specific case of compositional data, the row sums are equal to 1 and $R_0 = R_1 = R_2$.

Ordination of the structure and assemblage of interacting communities

From sCoCA, ordination diagrams can be made in the usual way by jointly plotting the species scores and sample scores (of each community) for the first axes of the analysis (ter Braak & Schaffers, 2004). For an optimal representation of this association in a biplot, the species scores of each axis must be multiplied by the quarter root of the eigenvalue of the axis (ter Braak, 1990). 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 be displayed in ordination diagrams (biplots; ter Braak & Schaffers, 2004). For instance, the joint plot of sample scores of hosts (table $\mathbf{X}$) with species scores of viruses (table $\mathbf{Y}$) displays the fit of the virus OTUs from the host communities. The simultaneous plot of sample scores of microalgae with the loadings (i.e., the coefficient or importance of variables on the first components of the PLS) of predictor species (i.e., hosts) allows representing microalgae communities and their OTUs. Both types of OTUs can also be jointly displayed using scores of the response species (i.e., viruses) with the loadings of predictor species (i.e., microalgae). For all these diagrams, axes are optimized to maximize covariance between assemblages rather than to depict associations within individual species matrices, but interpretation

can also be carried out as in correspondence analysis according to the barycentre principle, where OTUs are placed at the barycentre (weighted average) of the sample points and, by symmetry, samples at the barycentre (weighted average) of the OTUs (Supporting Information, Eq. 13–16).

Real data application

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

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 at the Natural Regional Park of the Chevreuse Valley (South of Paris, France). These systems were characterized by different local environmental conditions. Briefly, 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 0.98 similarity cutoff using CD hit (Fu, Niu, Zhu, Wu, & Li, 2012), and singletons were eliminated, before to assign 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 applied by Simon

et al. (2015) to select the most abundant OTUs. Finally, we subdivided the data set by grouping OTUs in two functional groups (autotrophic and heterotrophic) trained on literature information (Simon et al., 2015; Genitsaris, Monchy, Breton, Lecuyer, & Christaki, 2016 and references within), and we obtained one table for autotrophic microbial eukaryotes ($n=108$, $p=122$) and another table for microbial heterotrophic eukaryotes ($n=108$, $q=104$).

Case study 2: Microalgae-virus system

The data set coming from our research program contains a NGS-based eukaryotic microalgae community table (photosynthetic picoeukaryotes in the class Mamiellophyceae) and a 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 located in 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, i.e. the connection with 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 analyzing 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, & Huse (2009). The genomic sequences of PolB and V9 region were amplified

and sequenced using an Illumina MiSeq platform (GeT-PlaGe, INRA, Castanet-Tolosan, France). Sequences were processed and clustered into OTUs at 0.99 similarity cutoff and singletons were removed using MOTHUR v 1.35.1 (Schloss et al., 2009) for Mamiellophyceae and USEARCH v7 (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 genus *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 obtained one microalgae table ($n=31$, $p=67$) and one virus table ($n=31$, $q=98$).

Statistical analysis

Tables **X** and **Y** of autotrophic and heterotrophic microbial eukaryotes, respectively, were subjected to sCoCA. To test the significance of the global co-variation between the two tables, a Monte-Carlo permutation procedure with 9999 permutations was used. In each permutation, sCoCA (by considering all axes) is reapplied to obtain a value of the co-variance between table $\mathbf{Y}$ and row-permuted $\mathbf{X}$ (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 since we used the symmetric

form of the CoCA. A null distribution was estimated from co-variance calculated for the permuted data. The observed co-variance 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 co-variation between them. For the microalgae-virus data set, both tables were subjected to pCoCA with the SIMPLS algorithm. ter Braak & Schaffers (2004) suggest that the number of axes used to summarize the data can be selected by "leave-one-out" cross-validation procedure to maximize the cross-validatory fit (%) that 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 species scores from the CoCA, and thus from the co-structure between communities. A novel extension of SPIEC-EASI (Tipton et al. 2018) was used to infer the cross-group co-occurrence networks between two data sets. We used the neighborhood (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. Appendix 1 contains a *R* script and example data (from the case study 1 and the case study 2) allowing users to reproduce the analysis and apply them on their own data sets.

Results

Case study 1: Microbial eukaryotes

The common variance between the two microbial groups computed from the sCoCA explained, significantly ( ), 13.21% of the total variation of autotrophic microbial eukaryotes and 15.22% in 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 inter-pond 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, 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 with higher relative abundance of Biocosoecida, and Katablepharida in EVs and LCs and higher relative abundance of Ciliophera in SRs. Species scores of autotrophic microbial eukaryotes indicated that the patterns in heterotrophic communities are associated to a structure of 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, autotrophic communities made up mainly of other specific OTUs of Chlorophyta,

Chrysophyte, and Cryptophyta. It is also worth noting 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, 204 displayed 274 associations (Figure 2). From these associations between OTUs in the network, more positive (98.5%) than negative associations were inferred. All negative associations occurred between OTUs assigned to Ciliphora for heterotrophic microorganisms and Cryptophyta for autotrophic microorganisms. No clear association patterns can be identified in network from raw tables of autotrophic and heterotrophic microbial eukaryotes (Figure 2a). When the co-occurrence network is reordered according to species scores on the first axis of sCoCA, two modules can be distinguished (Figure 2b). The first module (i.e., top right corner) is constituted by OTUs exhibiting the higher relative abundances in brook system (i.e., RSA), while OTUs that compose the second module (i.e., bottom left corner) dominate pond systems (i.e., EV, LC, MG, and SR). Heterotrophic OTUs exhibited major associations with autotrophic OTUs belonging to the same module, with only 1.5% of associations between OTUs from 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 the module 1 (47%) and Ciliophora in the module 2 (59%). In pond systems (i.e., module 2), surprisingly, fungi are involved in very few associations (7%).

Case study 2: Microalgae-virus system

14

The cross-validation procedure identified the best pCoCA model based on the first two significant axes (pCoCA axis 1: $p = 0.001$, pCoCA axis 2: $p = 0.001$), in which Mamiellophyceae community predicted 32.02% of the variation in *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 Mamiellophyceae and *Prasinovirus* community respectively. The biplots indicated that the two communities covaried along a lagoon (samples LAs)/open-sea gradient (samples MAs) on the first axis (from left to right), while a temporal gradient for site LA (intra-site 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 both in lagoon samples, coastal samples and open-sea samples, associated with a similar repartition of MpVs (Figure 3a, b).

Based on the cross-group co-occurrence networks analysis, 67 associations were identified between the major 67 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 species scores on the first axis of pCoCA highlighted a structure in the 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 displaying the higher prevalence in lagoon samples (top right corner,

Figure 4b). Similarly, virus OTUs dominating the open-sea samples were mainly associated with microalgae OTUs from open-sea samples (bottom left corner, Figure 4b). Among the associations contained in the network, 50 were identified between OTUs belonging to an expected host-virus system (i.e. associations *Bathycoccus*/BpV, *Micromonas*/MpV, and *Ostreococcus*/OtV) while the other 17 significant associations were found between OTUs belonging to different host-virus systems. In average 50%, 70.6% and 91.3% 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 reciprocally. On the other hand, dyads were identified, that is specific associations, in *Bathycoccus*/BpV, *Micromonas*/MpV and *Ostreococcus*/OtV systems. Few negative associations inferred from the observation that those OTUs do not co-occur were found in network (Figure 4b). Interestingly, eight negative associations from a total of nine involved virus OTUs and microalgae OTUs belonging to different host-virus systems.

Discussion

Critical review and guidance papers on the analysis of NGS-based community data (Ramette, 2007; Buttigieg & Ramette, 2014; Paliy & Shankar, 2016) do not mention any direct quantitative method for predicting the composition of one community from another. Co-correspondence analysis (ter Braak & Schaffers, 2004) fills this gap. At the level of the case study 1, symmetric form of CoCA indicated that heterotrophic microbial eukaryote assemblages in shallow freshwater ecosystems were strongly associated with patterns of autotrophic microbial eukaryotes presence and abundance with links that can be taxon-specific (Figure 1). Our results shown also that the composition of heterotrophic microbial eukaryote community was dominated by fungi in 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 river than in ponds (Baldy et al., 2002). Higher fungal abundance might potentially be linked to incoming resources from runoff, since in brooks the most important source of imported material is usually deciduous leaves, whose the decomposition processing involved to a large extent fungi (Bärlocher, 1985; Webster & Benfield, 1986). To this, the composition of heterotrophic microbial eukaryote community characterizing brook system is associated a specific composition of microbial autotrophs. Such a result suggests that heterotroph community composition exert a control on the composition of autotroph community, and that microbial autotrophs can be driver of microbial heterotrophs.

Regarding the case study 2, the predictive form of CoCA points out 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 focusing 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), more eutrophic system, compared to *Bathycoccus*, which is found mainly in oligotrophic areas (Vaulot et al., 2012; Wu, Huang, & Zhong, 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 et al. (2010) showing that OtV are more abundant in lagoon than in the open sea.

Cross-taxon congruence description and evaluation (Virtanen, Ilmonen, Paasivirta, & Muotka, 2009; Gioria, Bacaro, & Feehan, 2011) provides a more comprehensive picture of the community similarity than the richness metrics conventionally used (Wolters, Bengtsson, & Zaitsev, 2006; Westgate, Barton, Lane, & Lindenmayer, 2014). 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 et al., 2012). In addition, the study of cross-taxon congruence between Mamiellophyceae and *Prasinovirus* is of definite 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 (Worden, Nolan, &Palenik, 2004; Jardillier, Zubkov, Pearman, & Scanlan, 2010), which accounts for > 50% 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 gradient (Bini, Vieira, Machado, & Machado Velho et al., 2007; Rooney & Bayley, 2012). As an important implication, the level of congruence can inform about the structural pattern among interacting groups (Özkan et al., 2014). That being said, reordering co-occurrence networks, from the species scores on the first axis of CoCA, allows 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). A substantial effort has been made in the development of metrics to estimate and test the level of nestedness (e.g., Rodriguez-Guirones & Santamaria, 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 et al., 2008). It has then been advocated that prior to analysis of 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). These findings taken together with our results validate our proposition to combine CoCA with network analysis to study structural patterns of microbial networks. In addition, in reverse to the expected view of nestedness structure of phage-bacteria network (Flores, Meyer, Valverde, Farr, & Weitz, 2011), the modular structure of Mamiellophyceae/*Prasinovirus* network observed in the field underpinned the modularity patterns previously observed in phage-bacteria network from cross-infection experiments (Flores, Valverde, & Weitz, 2013). Given that the structure of interaction networks is constraint by the coevolutionary processes between species (Peralta, 2016), this would lead to account for phylogenetic signals within co-occurrence network (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. 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 expected host-virus system were positive. These results are consistent with a previous work showing that when parasitism is captured as a significant link in co-occurrence network, it is retrieved as a positive link despite the detrimental effect of parasite on its host (Weiss et al., 2016). This might be explained because the copresence 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, since in our case, OTUs of concerned viruses and microalgae had inverse spatio-temporal 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

493 that the co-occurring Mamiellophyceae are the virus hosts, even if they belong to the
494 expected group of hosts (e.g., *Bathycoccus*, *Micromonas* or *Ostreococcus*). Our approach
495 (CoCA combined with network analysis, or other method to infer associations) could be
496 combined with other approaches, e.g. single cell genomics (Kalisky, Baliney, & Quake,
497 2011; Martinez-Garcia et al., 2012), Epic-PCR (Spencer et al., 2016), to validate the
498 predicted associations in interactions. The justification of the association sign between
499 the two functional groups making up the microbial eukaryote community is also not
500 straightforward, although they can be triggered by ecological interactions or by species
501 abiotic requirements (Derocles et al., 2018).
502 In conclusion, the successful application of co-correspondence analysis over two
503 real data sets of microbial communities exhibiting various types of interactions
504 reinforces that resorting to this method for study the distributions, assemblages and
505 interactions between two microbial communities constitutes a highly valuable approach
506 to understand the cross-taxon congruence between microorganisms. A useful
507 consequence of cross-taxon congruence is that the distribution of well-known taxa may
508 provide insight into the processes structuring the distribution of other taxa (e.g., Bilton,
509 McAbendorth, Bedford, & Ramsay, 2006; Santi et al., 2010). This approach could be used
510 to enhance our understanding of a major problem, the effect of phytoplankton bloom (in
511 particular toxic groups such as cyanobacteria) on the microbial communities and in turn
512 on the ecosystem functions (e.g., Yang et al., 2016; Xue et al., 2018). Our findings also
513 demonstrate that the reordering of co-occurrence networks, according to the
514 congruence information extracted from CoCA, allows to investigate the presence of
515 ecological signals in networks. The advantage of this approach is that the complexity of
516 the network is considerably reduced by the non-random placement of nodes in the
517 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 (Spiegelhalter, Pearson, & Short, 2011; Kjærgaard, 2015; Pocock et al., 2016). Interestingly, the applicability of our approach goes beyond the particular case of data sets with row-sum normalisation (i.e., compositional data). Indeed, CoCA was originally designed to analyze 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 the ecosystem functioning.

Acknowledgments

We are grateful for financial support from the French National Research Agency (DECOVIR ANR-12-BSV7-0009, coordinator Y.D.). We thank Ludwig Jardillier for give us access to genomic data on autotrophic and heterotrophic microbial eukaryotes.

References

Almeida-Neto M, Guimarães P., 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.

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.

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.

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). Netowrk analysis reveals functional redundancy and keystone taxa amongst

Barber M. J. (2007). Modularity and community detection in bipartite networks. *Physical Review*, 76, 066102.

Bärlocher F. (1985). The role of fungi in the nutrition of stream invertebrates. *Botanical Journal of the Linnean Society*, 91, 83–94.

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.

Benzécri F. (1969). Statistical analysis as a tool to make patterns emerge from data. In Watanabe (Ed.), *Methodologies of pattern recognition*. (pp 35–60). Academic Press: New York.

Bergelson J., Mittelstrass J., & Horton M. W. (2019). Characterizing both bacteria and fungi improves understanding of the *Arabidopsis* root microbiome. *Scientific Reports,* 9, 24.

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.

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.

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.

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.

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.

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.

Cordero O. X., & Datta M. (2016). Microbial interactions and community assembly at microscales. *Current Opinion in Microbiology*, 31, 227–234.

Derocles S. A. P., Bohan D. A., Dumbrell A. J., Kitson J. J. N., Massol, F., Pauvert C., … Evans D. M. (2018) Biomonitroing 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.

Dormann C. F., & Strauss R. (2014). A method for detecting modules in quantitative bipartite networks. *Methods in Ecology and Evolution*, 5, 90–98.

Dray S., Chessel D., & Thioulouse J. (2003). Co-inertia analysis and the linking of ecological data tables. *Ecology*, 84, 3078–3089.

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

Edgar R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, 1792–1797.

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

Falkowski P. G., Fenchel T., & Delong E. F. (2008). The microbial engines that drive earth's biogeochemical cycles. *Science*, 320, 1034–1038.

Fang H., Huang C., Zhao H., & Deng M. (2015). CCLasso: correlation inference for compositional data through Lasso. *Bioinformatics,* 31, 3172–3180.

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.

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.

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.

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.

Friedman J., & Alm E. J. (2012). Inferring correlation networks from genomic survey data. *PloS Computational Biology,* 8, e1002687.

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.

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

628 Gauch H. G. Jr., Whittaker R. H., & Wentworth T. R. (1977). A comparative study of
629 reciprocal averaging and other ordination techniques. *Journal of Ecology*, 65, 157–
630 174.
631 Genitsaris S., Monchy S., Breton E., Lecuyer E., & Christaki U. (2016). Small-scale
632 variability of protistan planktonic communities relative to environmental pressures
633 and biotic interactions at two adjacent coastal stations. *Marine Ecology Progress*
634 *Series*, 548, 61–75.
635 Gioria M., Bacaro G., & Feehan J. (2011). Evaluation an interpretating vorss-taxon
636 congruence: Potential pitfalls and solutions. *Acta Oecologia*, 37, 187–194.
637 Graham E. B., Knelman J. E., Schindlbacher A., Siciliano S., Breulmann M., Yannareli A., …
638 Nemergut D. R. (2016). Microbes as engines of ecosystem function: When does
639 community structure enhance predictions of ecosystem processes? *Frontiers in*
640 *Microbiolgy*, 7, 214.
641 Greenacre, M. (2009). Power transformation in correspondence analysis. *Computational*
642 *Statistics & Data Analysis*, 53, 3107–3116.
643 Guillou L., Bachar D., Audic S., Bass D., Berney C., Bittner L., … Christen R. The protest
644 Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-
645 unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 41, D597–D604.
646 Handelsman J. (2004). Metagenomics: Application of genomics to uncultured micro-
647 organisms. *Microbiology and Molecular Biology Reviews*, 68, 669–685.
648 Hill M. O. (1973). Reciprocal averaging: an eigenvector method ordination. *Journal of*
649 *Ecology*, 61, 237–249.
650 Hutchins D. A., & Fu F. (2017). Microorganisms and ocean global change. *Nature*
651 *Microbiology*, 2, 17058.
652 Ings T. C., Montoya J. M., Bascompte J., Blüthgen N., Brown L. Dormann C. F., …
653 Woodward G. (2009). Ecological network – beyond food webs. Journal of Animal
654 Ecology, 78, 253–269.
655 Jackson D. A. (1997). Compositional data in community ecology: the paradigm or peril of
656 proportions? *Ecology*, 78, 929–940.
657 Jackson D. A., & Harvey H. H. (1993). Fish and benthic invertebrates: community
658 concordance and community-environment relationships. *Canadian Journal of*
659 *Fisheries and Aquatic Sciences*, 50, 2641–2651.
660 Jardillier L., Zubkov M. V., Pearman J., & Scanlan D. J. (2010). Significant CO_2 fixation by
661 small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME*
662 *Journal*, 4, 1180-1192.
663 Johnson R. K., & Hering D. (2010). Spatial congruency of benthic diatom, invertebrate,
664 macrophyte, and fish assemblages in European streams. *Ecological Applications*, 20,
665 978–992.
666 Kalisky T., Baliney P., & Quake S. R. (2011). Genomics analysis et the single-cell level.
667 *Annual Review of Genetics,* 45, 431—445.
668 Kjærgaard R. S.. (2015). Data visualization: mapping the topical space. *Nature*, 520, 292–
669 293.
670 Kueneman J. G., Woodhams D. C., Treuren W. V., Archer H. M., Kniht R., & McKenzie V. J.
671 (2016). Inhibitory bacteria reduce fungi on early life stages of endangered Colorado
672 boreal toads (*Anaxyrus boreas*). *ISME Journal*, 10, 934–944.
673 Kurtz Z. D., Müller C. L., Miraldi E. R., Littman D. R., Blaser M. J., & Bonneau R. A. (2015).
674 Sparse and compositionally robust inference of microbial ecological networks. *PloS*
675 *Computational Biology,* 11, e1004226.
676 Legendre L., & Legendre P. *Numerical ecology*. Elsevier, Amsterdam, NL, 2012.

677 Leibold M. A., & Mikkelson G. M. (2002). Coherence, species turnover, and boundary
678 clumping: elements of meta-community structure. *Oikos*, 97, 237–250.
679 Li H. (2015). Microbiome, metagenomics, and high-dimensional compositional data
680 analysis. *Annual Review of Statistics and Its Applications*, 2, 73–94.
681 Little A. E., Robinson C. J., Peterson S. B., Raffa K. F., & Handelsman J. (2008). Rules of
682 engagement: interspecies interactions that regulate microbial communities. *Annual*
683 *Review of Microbiology*, 62, 375–401.
684 Liu H., Roeder K., & Wasserman L. (2010). Stability approach to regularization selection
685 (StARS) for high dimensional graphical models. *Advances in Neural Information*
686 *Processing Systems*, 23, 1432–1440.
687 Ma B., Wang H., Dsouza M., Lou J., He Y., Dai Z., … Gilbert J. A. (2016). Geographic patterns
688 of co-occurrence network topological features fro soil microbiota at continental scale
689 in eastern China. *ISME Journal*, 10, 1981–1901.
690 Maranon E., Holligan P. M., Barciela R., Gonzalez N., Mourino B., Pazo M. J., & Varela M.
691 (2001). Patterns of phytoplankton size structure and productivity in contrasting
692 open-ocean environments. *Marine Ecology Progress Series,* 216, 43-56.
693 Martens H. (2001). Reliable and relevant modelling of real world data: A personal
694 account of the development of PLS Regression. *Chemometrics and Intelligent*
695 *Laboratory Systems,* 58, 85—95.
696 Martinez-Garcia M., Brazel D., Poulton N. J., Swan B. K., Luesma Gomez M., Masland D., …
697 Stepanauskas, R. (2012). Unveiling *in situ* interactions between marine protists and
698 bacteria through single cell sequencing. *ISME Journal,* 6, 703—707.
699 McMurdie P. J., & Holmes S. (2014). Waste not, want not: why rarefying microbiome data
700 is inadmissible. *PloS Computational Biology,* 10, e1003531.
701 Morán X. A., López-Urrutia A., Calvo-Díaz A., & Li W. K. W. (2010). Increasing importance
702 of small phytoplankton in a warmer ocean. *Global Change Biology,* 16, 1137-1144.
703 Ni J., Yan Q., & Yu Y. (2013). How much metagenomic sequencing is enough to achieve a
704 given goal? *Scientific Reports*, 3, 1968.
705 Not F., Latasa M., Marie D., Cariou T. Vaulot D., & Simon N. (2004). A single species,
706 *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the
707 western English Channel. *Applied and Environmental Microbiology*, 70, 4064–4072.
708 Özkan K., Jeppesen E., Davidson T. A., Søndergaard M., Lauridsen T. L., Bjerring R., …
709 Svenning J.-C. (2014). Cross-taxon congruence in lake plankton largely independent
710 of environmental gradients. *Ecology*, 95, 2778–2788.
711 Paliy O., & Shankar V. (2016). Application of multivariate statistical techniques in
712 microbial ecology. *Molecular Ecology*, 25, 1032–1057.
713 Paulson J. N., Stine O. C., Bravo H. C., & Pop M. (2013). Differential abundance analysis for
714 microbial marker-gene surveys. *Nature Methods*, 10, 1200–1020.
715 Peralta G. (2016). Merging evolutionary history into species interaction networks.
716 *Functional Ecology*, 30, 1917–1925.
717 Pocock M. J. O., Evans D. M., Fontaine C., Harvey M., Julliard R., McLaughlin O., … Bohan D.
718 A. (2016). The visualization of ecological networks, and their use as a tool for
719 engagement, advocacy and management. *Advances I Ecological Research*, 54, 41–85.
720 Portillo M. C., Anderson S. P., & Noah F. (2012). Temporal variability in the diversity and
721 composition of stream bacterioplankton communities. *Environmental Microbiology*,
722 14, 2417–2428.
723 R Core Team (2019). *R: A language and environment for statistical computing*. R
724 Foundation for Statistical Computing, Vienna, Australia. Retrieved from
725 https://www.R-project.org/

Ramette A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology,* 62, 142–160.
Rappé M. S., & Giovannoni S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, 57, 369–394.
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.
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.
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.
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.
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.
Simpson G. L. (2009). Cocorresp: Co-correspondence analysis ordination methods. (R package version 0.3-0). (http://cran.r-project.org/package=analogue).
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.
Spiegelhalter D., Pearson M., & Short I. (2011). Visualizing uncertainty about the future. *Science*, 333, 1393–1400.
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.
Teira E., Mourino B., Maranón E., Perez V., Pazo 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.
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.
ter Braak C. J. F. (1990). Interpreting canonical correlation analysis through biplots of structural correlations and weights. *Psychometrika*, 55, 519–531.
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.
Ulrich W., & Gotelli N. J. (2007). Null model analysis of species nestedness patterns. *Ecology*, 88, 1824–1831.

774 Vacher C., Tamaddoni-Nezhad A., Kamenova S., Peyrard N., Moalic Y., Sabbadin R., …
775 Bohan, D. A. (2016). Learning ecological network from next-generation sequencing
776 data. *Advances in Ecological Research,* 54, 1–39.
777 Vaulot D., Lepère C., Toulza E., De la Iglesia R., Poulain J., Gaboyer F., … Piganeau G.
778 (2012). Metagenomes of the Picolaga *Bathycoccus* from the Chile coastal upwelling.
779 *PLoS One*, 7, e39648.
780 Viprey M., Guillou L., Ferréol M., & Vaulot D. (2008). Wide genetic diversity of
781 picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a
782 phylum-biased PCR approach. *Environmental Microbiology*, 10, 1804–1822.
783 Virtanen R., Ilmonen J., Paasivirta L., & Muotka T. (2009). Community concordance
784 between bryophyte and insect assemblages in boreal springs: A broad-scale study in
785 isolated habitats. *Freshwater Biology*, 54, 1651–1662.
786 Webster J. R., & Benfield E.F. (1986). Vascular plant breakdown in freshwater
787 ecosystems. *Annual Review of Ecology and Systematics*, 17, 567–594.
788 Weiss S., Van Treuren W., Lozupone C., Faust K., Friedman J., Deng Y., …Knight R. (2016).
789 Correlation detection strategies in microbial data sets vary widely in sensitivity and
790 precision. *ISME Journal*, 10, 1669–1681.
791 Westgate M. J., Barton P. S., Lane P. W., & Lindenmayer D. B. (2014). Global meta-analysis
792 reveals low consistency of biodiversity congruence relationships. *Nature*
793 *Communications*, 5, 3899.
794 Wisz M. S., Pottier J., Kissling W. D., Pellissier L., Lenoir J., Damgaard C. F., … Svenning J.-
795 C. (2013). The role of biotic interactions in shaping distributions and realised
796 assemblages of species: implications for species distribution modelling. *Biological*
797 *Reviews*, 88, 15–30.
798 Wolters, Bengtsson, & Zaitsev. (2006). Relationship among the species richness of
799 different taxa. *Ecology*, 87, 1886–1895.
800 Worden A. Z., Nolan J. K., & Palenik B. (2004). Assessing the dynamics and ecology of
801 marine picophytoplankton: the importance of the eukaryotic component. *Limnology*
802 *and Oceanography,* 49, 168-179.
803 Wu W., Huang B., & Zhong C. (2013). Photosynthetic picoeukaryote assemblages in the
804 South China Sea from the Pearl River estuary to the SEATS station. *Aquatic Microbial*
805 *Ecology,*71, 271–284.
806 Xue Y., Chen H., Yang J. R., Liu M., Huang B., & Yang J. (2018). Distinct patterns and
807 processes of abundant and rare eukaryotic plankton communities following a
808 reservoir cyanobacterial bloom. *ISME Journal*, 12, 2263–2277.
809 Yang C., Li Y., Zhou Y., Lei X., Zheng W., Tian Y., … Zheng T. (2016). A comprehensive
810 insight into functional profiles of free-living microbial community responses to toxic
811 *Akshiwo sanguinea* bloom. *Scientific Reports*, 6, 34645.
812 Zhu F., Massana R., Not F., Marie D., & Vaulot D. (2005). Mapping of picoeukaryotes in
813 marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology*
814 *Eco*logy, 52, 79–92.
815 Zimmerman N., Izard J., Klatt C., Zhou J., & AronsonE. (2014). The unseen world:
816 environmental microbial sequencing and identification methods for ecologists.
817 *Frontiers in Ecology and the Environment*, 12, 224–231.
818 Zinger L., Gobet A., & Pommier T. (2011). Two decades of describing the unseen majority
819 of aquatic microbial diversity. *Molecular Ecology,* 21, 1878–1896.
820

821 **Data accessibility statement**

Data and R scripts to reproduce the different analyses of the case study 1 and the case study 2 ~~Mamiellophyceae/*Prasinovirus* system~~ are available online in the Appendix 1~~as additional supporting information~~.

Author contributions

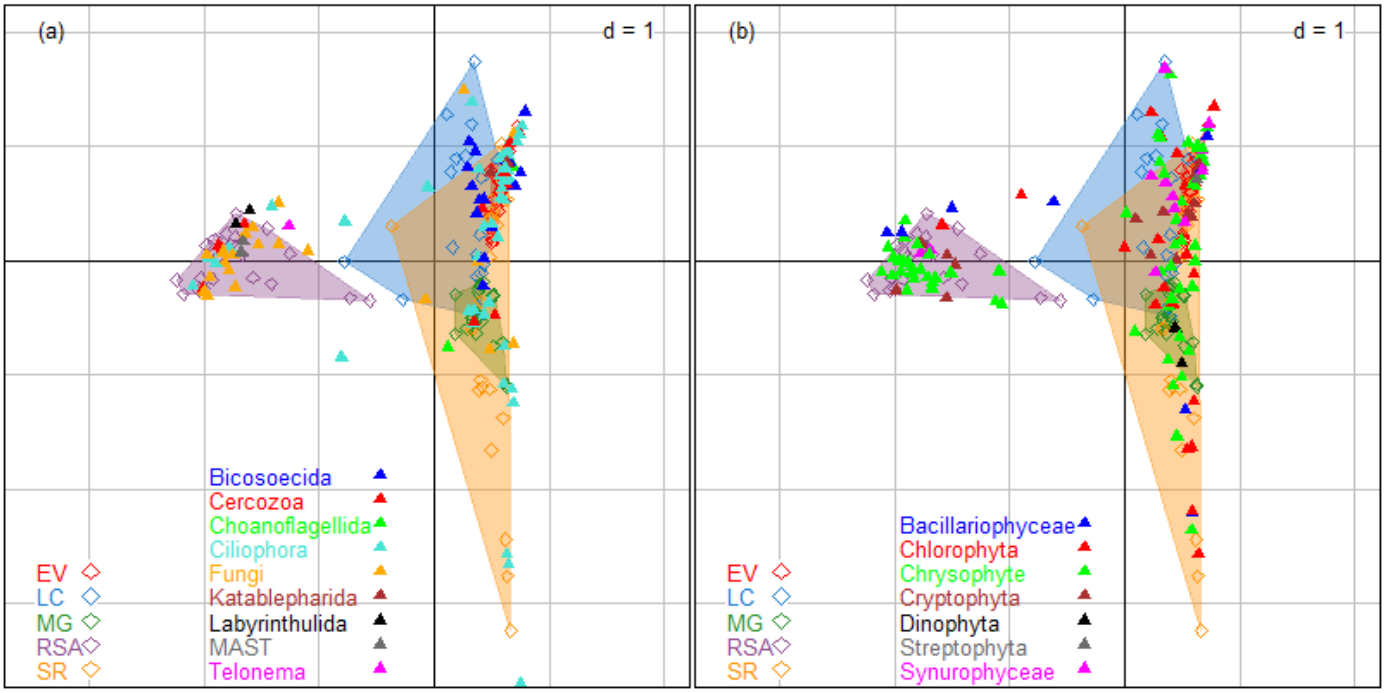
Y.D. and S.D. supervised the project; B.A. performed all the statistical analyses; C.J.F.B. contributed to mathematical development of the method; H.L. contributed to collecting and genotyping the biological materials of microalgae and viruses; B.A. wrote the first draft of the manuscript; B.A., C.J.F.B., Y.D., and S.D. commented and approved the final version of the manuscript.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

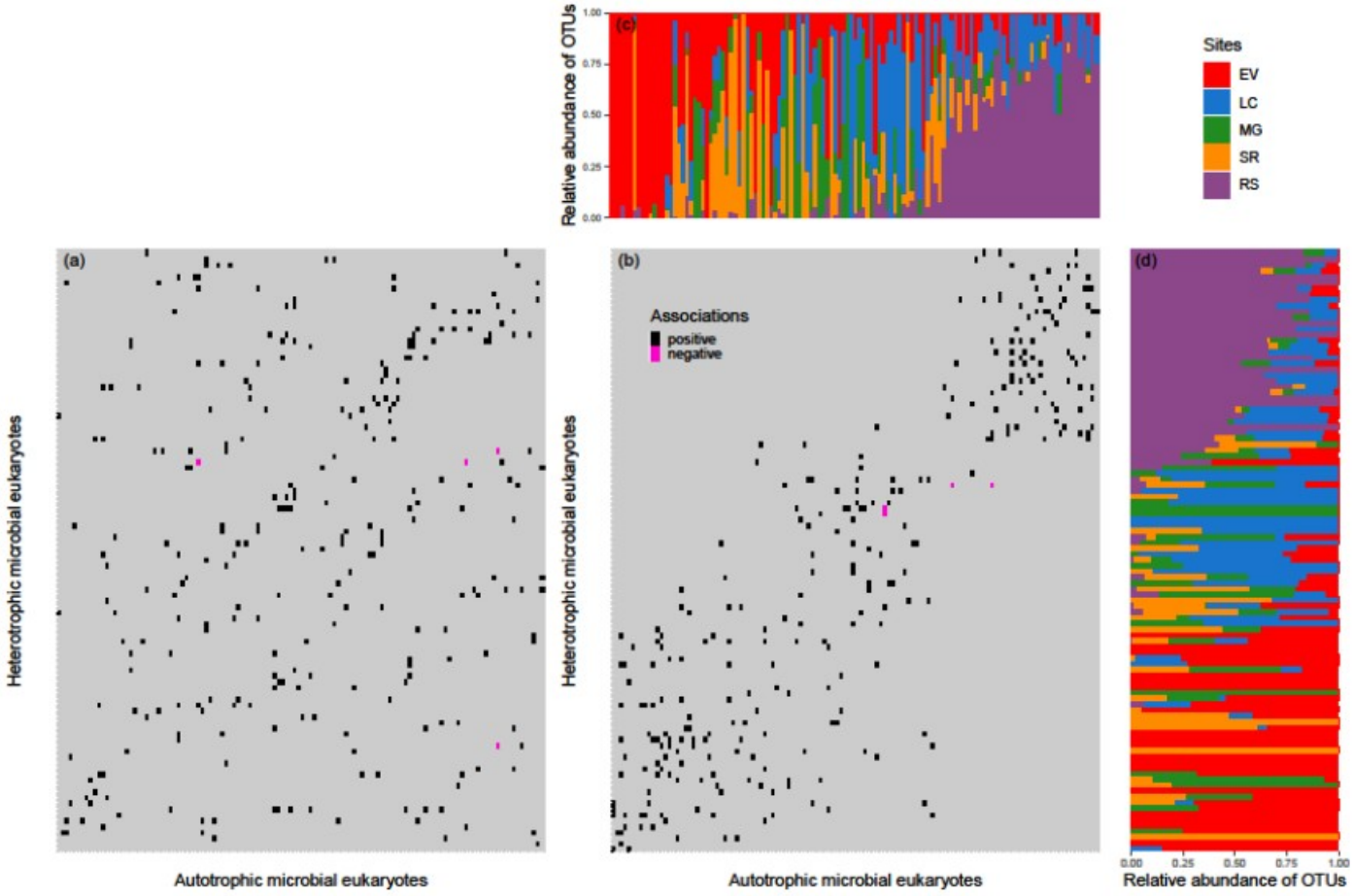
Figures

Figure 1. Ordination biplots of the case study 1, representing ~~of the communities of~~ the positions of sites (open diamond) and species (solid triangle) on the axis 1 × axis 2 factorial plan of the symmetric co-correspondence analysis. (a) Biplot for the heterotrophic microbial eukaryotes and (b) biplot for the autotrophic microbial eukaryotes ~~OTUs obtained from symmetric co-correspondence analysis~~. EV: Etang des Vallées, LC: La Claye, GB: Mare Gabard, SR: Saint Robert, RSA: Ru Sainte Anne. Heterotrophic and autotrophic microbial eukaryotes OTUs are colored according to the phylogenetic group they belong to. "d" indicates the mesh of the grid.



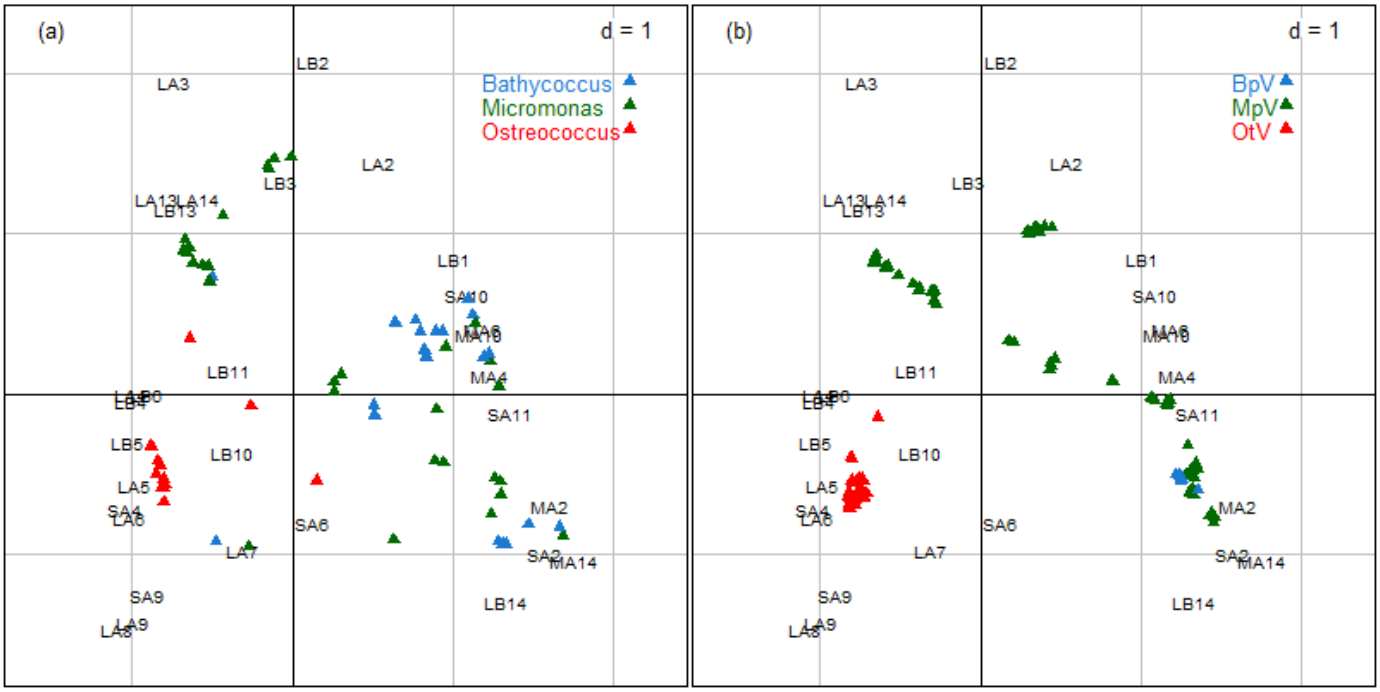
30

Figure 2. Heatmap of the case study 1, representing the association network between the~~of~~ heterotrophic and autotrophic microbial eukaryotes (a) before ~~(a)~~ and (b) after ~~(b)~~ the reordering the position of each species in the network according to their ~~network from the species~~ scores on the first axis of symmetric co-correspondence analysis. Bar plots of the relative abundance of (c) autotrophic microbial eukaryotes and (d) heteretrophic 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, LC: La Claye, GB: Mare Gabard, SR: Saint Robert, RSA: Ru Sainte Anne.



31

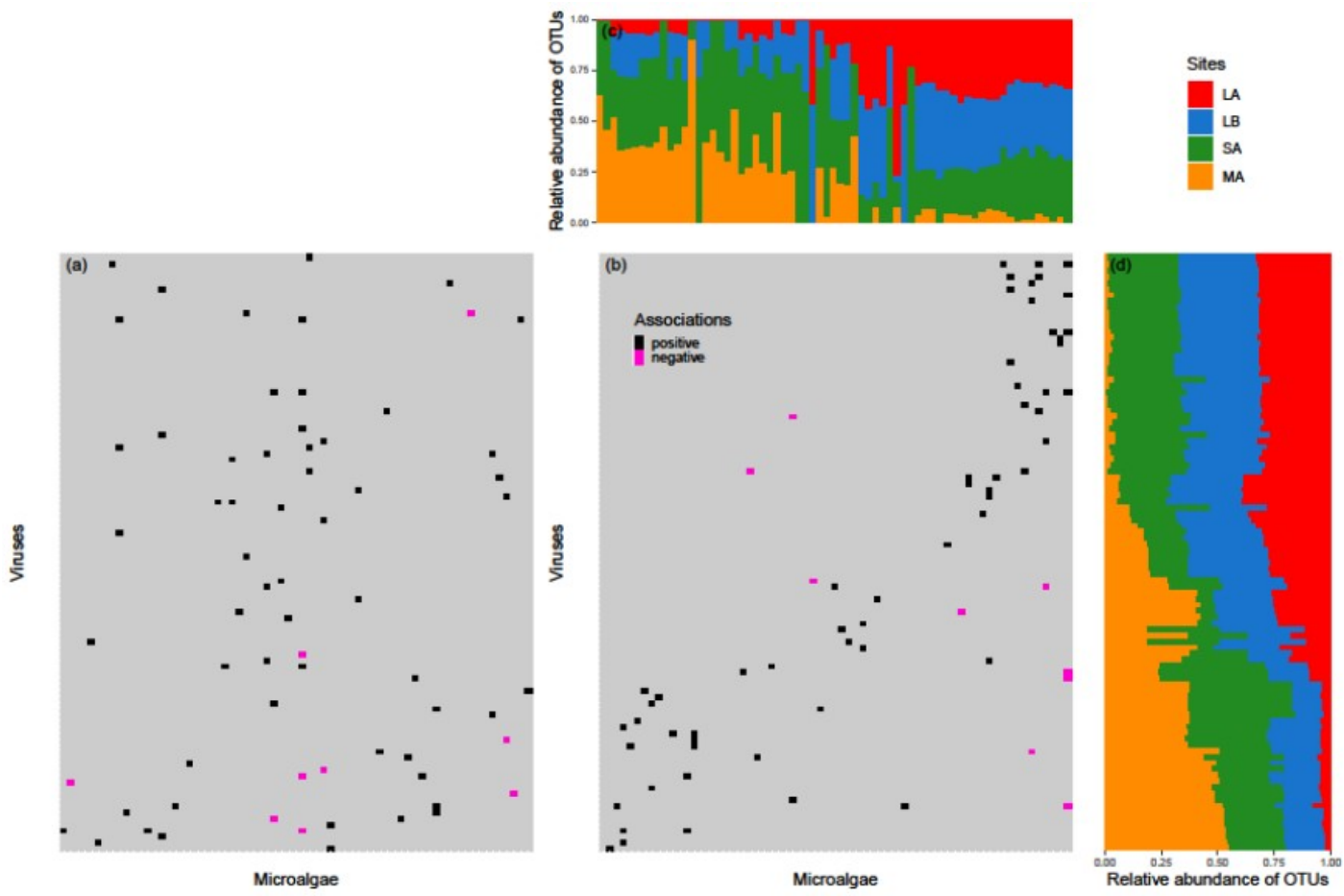
Figure 3. Ordination biplots of, ~~concerning~~ the case study 2, ~~of~~representing the positions of sites (open diamond) and species (solid triangle) on the axis 1 × axis 2 factorial plan of the predictive co-correspondence analysis. (a) Biplot for the Mamiellophyceae and (b) biplot for the *Prasinovirus* ~~OTUs obtained from predictive co-correspondence analysis~~. LB: coastal site in Leucate lagoon, LA: site at the level of the Grau in Leucate lagoon, SA: coastal site, MA: open-sea site. Microalgae and virus OTUs are colored according to the phylogenetic group they belong to. BpV: *Bathycoccus* viruses, MpV: *Micromonas* viruses, OtV: *Ostreococcus* viruses. "d" indicates the mesh of the grid.



~~**Figure 3.** Heatmap of heterotrophic and autotrophic microbial eukaryotes before (a) and after (b) the reordering the network from the species scores on the first axis of symmetric co-correspondence analysis. Bar plots of the relative abundance of (c) autotrophic microbial eukaryotes and (d) heteretrophic 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, LC: La Claye, GB: Mare Gabard, SR: Saint Robert, RSA: Ru Sainte Anne.~~

32

Figure 4. Heatmap of the case study 2, representing the~~of~~ microalgae-virus~~es~~ (Mamiellophycaea/*Prasinovirus*) association network (a) before ~~(a)~~ and (b) after ~~(b)~~ the reordering the position of each species in the network according to their ~~network from the species~~ 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.



33