



HAL
open science

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

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

► **To cite this version:**

Benjamin Alric, Cajo J.F. ter Braak, Yves Desdevises, Hugo Lebretonchel, Stéphane Dray. Investigating microbial associations from sequencing survey data with co-correspondence analysis. *Molecular Ecology Resources*, 2020, 20 (2), pp.468-480. 10.1111/1755-0998.13126 . hal-02400150

HAL Id: hal-02400150

<https://univ-lyon1.hal.science/hal-02400150v1>

Submitted on 31 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

3 Benjamin Alric^{1†}, Cajo J.F. ter Braak², Yves Desdevises³, Hugo Lebretonchel³, Stéphane
4 Dray¹

5
6 ¹*Université de Lyon, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive,*
7 *Université Lyon1, F-69622 Villeurbanne, France*

8 ²*Biometris, Wageningen University & Research, Box 16, 6700 AA Wageningen, The*
9 *Netherlands*

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

12

13

14

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

18

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

21

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

24

25 Running title: Cross-taxon congruence in microbial communities

26 **Abstract**

27 Microbial communities, which drive the major ecosystem functions, are
28 composed by a wide range of interacting species. Understanding how microbial
29 communities are structured and the underlying processes is a crucial task for
30 interpreting ecosystem response to global change but it is challenging as microbial
31 interactions cannot usually be directly observed. Multiple efforts are currently focused
32 to combine next-generation sequencing (NGS) techniques with refined statistical
33 analysis (e.g., network analysis, multivariate analysis) to characterize the structures of
34 microbial communities. However, most of these approaches consider a single table of
35 sequencing data measured for several samples. Technological advances now make it
36 possible to collect NGS data on different taxonomic groups simultaneously for the same
37 samples and lead to analyze a pair of tables. Here, an analytic framework based on co-
38 correspondence analysis (CoCA) is proposed to study the distributions, assemblages and
39 interactions between two microbial communities. We showed the ability of this
40 approach to highlight the relationships between two microbial communities, using two
41 data sets exhibiting various types of interactions. CoCA identified strong association
42 patterns between autotrophic and heterotrophic microbial eukaryotes assemblages, on
43 one hand, and between microalgae and viruses, on the other hand. We demonstrate also
44 how CoCA can be used, in complement to network analysis, to reorder co-occurrence
45 networks and thus investigate the presence of patterns in ecological networks.

46 **Introduction**

47 Microbial communities are highly diverse (Rappé & Giovannoni, 2003) and drive
48 the major ecosystem functions ([e.g., carbon sequestration, climate regulation, gas](#)
49 [regulation, nutrient cycling](#); Ducklow, 2008; Falkowski, Fenchel, & Delong, 2008;
50 Hutchins & Fu, 2017). Understanding how these systems are structured and identifying
51 the underlying processes is a crucial task to predict communities and ecosystem
52 responses to global change (Fuhrman, 2009; Graham et al., 2016). Biotic interactions
53 across taxonomic groups (e.g., predation, parasitism, mutualism or competition) are of
54 broad interest because they are expected to influence the structure and composition of
55 communities (Wisz et al., 2013). Unfortunately, our understanding of the underlying
56 assemblage rules of microbial communities is still limited (Little et al., 2008; Cordero &
57 Datta, 2016).

58 The emergence of high-throughput sequencing techniques (next-generation
59 sequencing; NGS) gave access to the diversity of whole microbial communities, including
60 the non-cultivable fraction (Handelsman, 2004; Zimmerman, Izard, Klatt, Zhou, &
61 Aronson, 2014; Zinger, Gobet, & Pommier, 2011). With the large amount of data
62 generated in a single NGS experiment, powerful statistical methods are needed to assess
63 and explain structural patterns in such complex data sets (Bálint et al., 2016; Paliy &
64 Shankar, 2016). A common approach is to combine NGS techniques with network
65 analysis to represent and characterize interactions between partners in microbial
66 communities (Cardona, Weisenhorn, Henry, & Gilbert, 2016; Vacher et al., 2016).
67 Various computational methods have been developed to infer networks from NGS data
68 sets (e.g., CoNet: Faust et al., 2012; SparCC: Friedman & Alm, 2012; REBACCA: Ban, An, &
69 Jiang, 2015; CCLasso: Fang, Huang, Zhao, & Deng, 2015; SPIEC-EASI: Kurtz et al., 2015).
70 Another popular approach is to use ordination methods to extract information from NGS

71 data and describe the variations in community composition among samples (Paliy &
72 Shankar, 2016). Ordination methods arrange objects in a multidimensional space using
73 directly the original raw data table (e.g., principal component analysis, correspondence
74 analysis), or after computing a distance matrix (e.g., non-metric multidimensional
75 scaling, principal coordinate analysis) (Legendre & Legendre, 2012).

76 These different approaches are based on a single table composed of read counts
77 for each Operational Taxonomic Unit (OTU) measured for several samples.
78 Technological advances now make it possible to acquire NGS data on different
79 taxonomic groups simultaneously for the same samples (Fierer et al., 2007) and lead to
80 analyze a pair of tables (i.e., OTUs composition for the same sampling sites for two
81 different taxonomic groups). To analyze such pair of tables, a common practice consists
82 in merging the two tables into a single one and then applying network analysis (e.g.,
83 Kueneman et al., 2016; Banerjee et al., 2016; Ma et al., 2016) and/or multivariate
84 analysis (Cannon et al., 2017; Bergelson, Mittelstrass, & Horton, 2019). However, this
85 data aggregation is unsuitable especially when NGS data sets, which are a function of
86 sequencing depth (Ni, Yan, & Yu, 2013), are standardized by dividing read counts by the
87 total number of reads in each sample. In this case, the normalisation step and further
88 analysis are very sensitive to the difference in number of OTUs and associated counts in
89 each taxonomic group. Hence, it is important to use techniques allowing for the analysis
90 of a pair of NGS data tables while preserving the original structure of the data. In
91 addition, these approaches must be able to mitigate the statistical bias stemming from
92 high-dimensionality (i.e., a number of samples substantially lower than the number of
93 variables), sparsity (i.e., a high proportion of zero counts), and the compositional nature
94 (i.e., a non-independence of relative abundances induced by the row-sum normalisation)
95 that characterize NGS data (Li, 2015). For network analysis, the SPIEC-EASI method has

96 been adapted to infer associations among microorganisms in a cross-domain analysis
97 (Tipton et al., 2018). For multivariate analyses, several two-table methods exist and
98 have been presented to microbial ecologists in methodological reviews (Ramette, 2007;
99 Buttigieg & Ramette, 2014; Paliy & Shankar, 2016). However, these works focused on
100 asymmetric methods (e.g., canonical correspondence analysis and redundancy analysis)
101 that aim to explain the composition of microbial communities by a limited number of
102 environmental predictors. Unfortunately, these methods are not adapted to link two
103 NGS data tables as they require that there are fewer predictor variables than samples
104 (Dray, Chessel, & Thioulouse, 2003) and thus are not able to deal with the high-
105 dimensionality of NGS data. Moreover, these methods compute linear combinations of
106 the predictor variables, which is not suitable if the table of predictors contains
107 community data that display unimodal structure and/or are sum [normalised](#) (as the NGS
108 data).

109 This study aims to propose an analytical framework based on co-correspondence
110 analysis (CoCA; ter Braak & Schaffers, 2004), a two-table coupling method developed in
111 community ecology, to study the distributions and assemblages between two microbial
112 communities. This framework is based on correspondence analysis, a method that
113 effectively handles proportional data that contain many zeroes (Gauch et al., 1977;
114 Jackson, 1997; [Greenacre, 2009](#)), like NGS data (Paulson, Stine, Bravo, & Pop et al.,
115 2013). We show how this method allows to extract information about the co-structure
116 among two microbial communities to estimate the congruence between them. Finally,
117 we show that the outputs of the method can be used to reorder [co-occurrence](#) networks
118 inferred by network analysis to enhance the visualization of microbial [association](#) and
119 the understanding of assemblage patterns within networks. Hence, our approach echoes
120 to the current lively debate about the practices to create network visualizations which

121 are both aesthetically appealing and have high information content (see Pocock et al.,
122 2016 for review). We illustrate our approach using two real data sets, one on
123 autotrophic and heterotrophic microbial eukaryotes in shallow freshwater systems and
124 another on microalgae and viruses in marine systems.

125

126 **Materials and methods**

127 ***Studying cross-taxon congruence by co-correspondence analysis***

128 Co-correspondence analysis (ter Braak & Schaffers, 2004) is part of the class of
129 canonical analyses with the feature to be designed to analyze a pair of tables containing
130 abundance data and to study the co-variations between two types of communities (e.g.,
131 plants and pollinators). CoCA is based on correspondence analysis (Benzécri, 1969; Hill,
132 1973) and preserves its fundamental properties of weighted averaging and the use of
133 the chi-square (χ^2) distance for both tables (Supporting Information). The χ^2 distance is
134 particularly adapted to NGS data as it handles properly zero values (and in particular
135 double absences) and thus is not hampered by zero inflation (Legendre & Legendre,
136 2012).

137 Here, co-correspondence analysis in its predictive form (pCoCA) is used for the
138 microalgae-virus data set while the symmetric form (sCoCA) is used for the microbial
139 eukaryote data set. Let $X=[x_{ij}]$ and $Y=[y_{ik}]$ be $n \times p$ and $n \times q$ tables containing the
140 relative abundances of each p and q species of two communities measured at the same n
141 samples. In the first case, pCoCA is chosen because we wish to investigate the
142 composition of virus communities under the hypothesis that the occurrence of viruses
143 depends mainly on whether microalgal hosts are present or absent at a particular
144 sample location. In the second case, sCoCA is chosen as we simply wished to study the
145 relationships between autotrophic and heterotrophic microbial eukaryotes. In its

146 | predictive form, CoCA is based on partial least-squares regression analysis (PLS) in a
147 SIMPLS version (ter Braak & Schaffers, 2004) to deal with high-dimensionality and
148 subsequent collinearity in the table of explanatory variables. PLS searches for a linear
149 regression model from a set of orthogonal components (called latent factors) built from
150 collinear explanatory variables with the constraint that these components maximize the
151 covariance with the response variables (Martens, 2001). In its symmetric form, CoCA fits
152 in the framework of co-inertia analysis (COIA, Dolédec & Chessel, 1994) that is not
153 affected by the problem of collinearity (Dray, Chessel, & Thioulouse, 2003). Co-inertia
154 analysis relates two data tables in a symmetric way, by providing a decomposition of the
155 co-inertia criterion on a set of orthogonal axes on which sample scores are projected
156 (Dray, Chessel, & Thioulouse, 2003).

157 | In practice, CoCA identifies associations (or common ecological gradients)
158 | between two types of biological assemblages from the same samples, by seeking the
159 | factorial axes that maximize the covariance between the weighted average sample
160 | scores (projection of rows of one table onto the factorial axes) of one community with
161 | those of the other community (Supporting Information, Eq. 12). This requires to
162 | determine species scores (projection of columns of one table onto the factorial axes) of
163 | one table as weighted averages of sample scores of the other table and sample scores as
164 | weighted averages of the species scores of their own table (Supporting Information, Eq.
165 | 13–16). So, pCoCA and sCoCA are a weighted version of PLS and COIA, respectively.
166 | Given that CoCA is related to the correspondence analysis, it is necessary to circumvent
167 | the fact that the sample weights of Y and X (R_1 and R_2 , which are the row sums of Y and
168 | X, respectively) are imposed and are not similar (Supporting Information, Eq. 6 and Eq.
169 | 7). This does not meet a crucial constraint in estimating the co-structure between two
170 | tables, namely that the samples must be weighted in the same way for the two tables

171 [\(Dray, Chessel & Thioulouse, 2003\)](#). Hence, an additional common sample weights
172 matrix (R_0) [is defined to replace](#) R_1 and R_2 [so that the weighted averaging properties of](#)
173 [CA can be retained in CoCA \(ter Braak & Schaffers, 2004\)](#). In pCoCA, R_0 is equal to the
174 sample weights of the community response table (i.e., $R_0 = R_1$), whereas in sCoCA,
175 $R_0 = (R_1 + R_2) / 2$ ([Supporting Information, Eq. 10 and Eq. 11](#)). Note that in the specific case
176 of compositional data, the row sums are equal to 1 and $R_0 = R_1 = R_2$.

177

178 ***Ordination of the structure and assemblage of interacting communities***

179 From sCoCA, ordination diagrams can be made in the usual way by jointly
180 plotting the species scores and sample scores (of each community) for the first axes of
181 the analysis (ter Braak & Schaffers, 2004). For an optimal representation of this
182 association in a biplot, the species scores of each axis must be multiplied by the quarter
183 root of the eigenvalue of the axis (ter Braak, 1990). From pCoCA, the fit of the response
184 community (viruses) to the predictive community (microalgal hosts) as well as the
185 variations in the composition of communities can be displayed in ordination diagrams
186 (biplots; ter Braak & Schaffers, 2004). For instance, the joint plot of sample scores of
187 [hosts](#) (table X) with species scores of viruses (table Y) displays the fit of the virus OTUs
188 from the [host](#) communities. The simultaneous plot of sample scores of microalgae with
189 the loadings [\(i.e., the coefficient or importance of variables on the first components of](#)
190 [the PLS\)](#) of predictor species (i.e., [hosts](#)) allows representing microalgae communities
191 and their OTUs. Both types of OTUs can also be jointly displayed using [scores](#) of the
192 response species (i.e., viruses) with the loadings of predictor species (i.e., microalgae).
193 For all these diagrams, axes are optimized to maximize covariance between assemblages
194 rather than to depict associations within individual species matrices, but interpretation

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

199

200 | ***Real data application***

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

206 |

207 | ***Case study 1: Microbial eukaryotes***

208 | Surface water was sampled monthly from April 2011 to April 2013 in five small
209 shallow freshwater systems [four ponds: Etang des Vallées (EV), La Claye (LC), Mare
210 Gabard (GB), Saint Robert (SR), and one brook: Ru Sainte Anne (RSA)], located at the
211 Natural Regional Park of the Chevreuse Valley (South of Paris, France). These systems
212 were characterized by different local environmental conditions. Briefly, raw genomic
213 sequences were obtained from 18S rDNA fragments, encompassing the V4
214 hypervariable region, applying 454 pyrosequencing and filtered to remove potential
215 spurious sequences using a local pipeline (Simon et al., 2014). Sequences from all
216 samples were then processed together and clustered into OTUs at 0.98 similarity cutoff
217 using CD hit (Fu, Niu, Zhu, Wu, & Li, 2012), and singletons were eliminated, before to
218 assign OTUs to taxonomic groups based on sequence similarity to the PR2 database
219 (Guillou et al., 2013). From this overall OTU table, we used the method applied by Simon

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

226 |

227 | Case study 2: Microalgae-virus system

228 The data set coming from our research program contains a NGS-based eukaryotic
229 microalgae community table (photosynthetic picoeukaryotes in the class
230 Mamiellophyceae) and a NGS-based virus community table (viruses infecting this class
231 of eukaryotic phytoplankton and belonging to the genera *Prasinovirus* of the family
232 Phycodnaviridae). The data have been acquired across four sites located in northwest
233 Mediterranean Sea (Gulf of Lion) and sampled monthly from March 2013 to April 2014.
234 The Gulf of Lion is characterized by contrasted environments, including eutrophic
235 lagoons connected to the sea, nutrient-rich coastal sites, and oligotrophic open-sea
236 locations. The sample locations included two sites in Leucate lagoon (one coastal site
237 (LB) and another site (LA) at the level of the Grau, i.e. the connection with the sea), a
238 coastal site (SA, marine station included in the French marine monitoring network
239 SOMLIT) and an open-sea site (MA, marine station included in the monitoring network
240 MOOSE). The characterization of *Prasinovirus* was based on analyzing the partial
241 sequence of the DNA polymerase gene (PolB) amplified using two primer sets (Chen &
242 Suttle, 1995; Clerissi et al., 2014). For Mamiellophyceae, the sequence of the V9 region of
243 the 18S rDNA was amplified using primers defined by Amaral-Zettler, McCliment,
244 Ducklow, & Huse (2009). The genomic sequences of PolB and V9 region were amplified

245 and sequenced using an Illumina MiSeq platform (GeT-PlaGe, INRA, Castanet-Tolosan,
246 France). Sequences were processed and clustered into OTUs at 0.99 similarity cutoff and
247 singletons were removed using MOTHUR v 1.35.1 (Schloss et al., 2009) for
248 Mamiellophyceae and USEARCH v7 (Edgar, 2010) combined with MUSCLE software
249 (Edgar, 2004) for *Prasinovirus*. Sequences were then compared against the PR2 database
250 (Guillou et al., 2013) and the NCBI database for Mamiellophyceae and *Prasinovirus*,
251 respectively, in order to assign OTUs to taxonomic groups based on similarity. Finally,
252 we focused specifically on OTUs assigned to the family Mamiellales of the class
253 Mamiellophyceae, and notably the genus *Bathycoccus*, *Micromonas* and *Ostreococcus*
254 which usually dominate this class in the Gulf of Lion and more generally the
255 picoeukaryotic fraction in other ecosystems (Wu, Huang, & Zhong, 2013; Zhu, Massana,
256 Not, Marie, & Vaultot, 2005). Subsequently, the *Prasinovirus* data set was limited to OTUs
257 assigned as *Bathycoccus* viruses (BpVs), *Micromonas* viruses (MpVs) and *Ostreococcus*
258 viruses (OtVs). The dominant microalgae and virus OTUs (i.e., $\geq 0.1\%$ of mean relative
259 abundance for at least two samples) were selected to obtained one microalgae table (
260 $n=31$, $p=67$) and one virus table ($n=31$, $q=98$).

261

262 **Statistical analysis**

263 Tables X and Y of autotrophic and heterotrophic microbial eukaryotes,
264 respectively, were subjected to sCoCA. To test the significance of the global co-variation
265 between the two tables, a Monte-Carlo permutation procedure with 9999 permutations
266 was used. In each permutation, sCoCA (by considering all axes) is reapplied to obtain a
267 value of the co-variance between table Y and row-permuted X (so that samples are
268 randomized while preserving the relative abundance of individuals). Note that the
269 choice of the table to be reordered is not important here since we used the symmetric

270 form of the CoCA. A null distribution was estimated from co-variance calculated for the
271 permuted data. The observed co-variance is then compared to the distribution obtained
272 under the null hypothesis. The positions of the samples on ordination axis of each table
273 are then correlated to show the overall level of co-variation between them. For the
274 microalgae-virus data set, both tables were subjected to pCoCA with the SIMPLS
275 algorithm. ter Braak & Schaffers (2004) suggest that the number of axes used to
276 summarize the data can be selected by “leave-one-out” cross-validation procedure to
277 maximize the cross-validatory fit (%) that measures how well the table **X** (microalgae in
278 our case) predicts the response table **Y** (viruses in our case). Working just with these
279 significant axes provides a measure of association between the tables by removing
280 random noise and keeping only the major dimensions of ecological variability. Finally,
281 we combine CoCA and network analysis so that nodes in [co-occurrence](#) networks are
282 reordered according to the species scores from the CoCA, and thus from the co-structure
283 between communities. A novel extension of SPIEC-EASI (Tipton et al. 2018) was used to
284 infer the cross-group [co-occurrence](#) networks between two data sets. We used the
285 neighborhood (MB) setting and selected the optimal sparsity parameter based on the
286 Stability Approach to Regularization Selection (StARS) (Liu, Roeder, & Wasserman,
287 2010). The StARS variability threshold was set to 0.05 for networks built from the two
288 data sets. All statistical analyses were performed with the R software (R Core Team,
289 2019) and using the *cocorresp* package (Simpson, 2009) for CoCAs and the *SpiecEasi*
290 package (Kurtz et al., 2015) for [co-occurrence](#) networks. [Appendix 1 contains a R script](#)
291 [and example data \(from the case study 1 and the case study 2\)](#) allowing users to
292 reproduce the analysis and apply them on their own data sets.

293

294 **Results**

295 | Case study 1: Microbial eukaryotes

296 | The common variance between the two microbial groups computed from the
297 | sCoCA explained, significantly (), 13.21% of the total variation of autotrophic microbial
298 | eukaryotes and 15.22% in heterotrophic microbial eukaryotes. Of the common variance,
299 | 39.51% was accounted by the first three axes of the sCoCA (sCoCA axis 1: 18.95%,
300 | sCoCA axis 2: 10.55%, sCoCA axis 3: 10.01%). The first three ordination axes of the
301 | autotrophic eukaryotes were highly correlated with the first three ordination axes of the
302 | heterotrophic eukaryotes (correlations being 0.95, 0.92, and 0.92), demonstrating a high
303 | degree of similarity in change between autotrophic and heterotrophic microbial
304 | eukaryote assemblages. Communities of autotrophic and heterotrophic microbial
305 | eukaryotes covaried along a brook/pond gradient on the first axis (from left to right),
306 | and an inter-pond variability on the second axis (Figure 1). Marked differences in the
307 | composition of the two communities are visible in joint plots. In the brook system (i.e.,
308 | RSA), the heterotrophic microbial eukaryote community is mainly composed of fungi,
309 | MAST, Labyrinthulida, and Telonema, whereas in pond systems (i.e., EV, LC, MG, SR)
310 | Ciliophora, Biocosoecida, Katablepharida, and Choanoflagellida dominated the
311 | communities (Figure 1a). Differences between pond systems are explained with higher
312 | relative abundance of Biocosoecida, and Katablepharida in EVs and LCs and higher
313 | relative abundance of Ciliophera in SRs. Species scores of autotrophic microbial
314 | eukaryotes indicated that the patterns in heterotrophic communities are associated to a
315 | structure of the autotrophic community (Figure 1b). In the brook system, the
316 | autotrophic microbial eukaryote community exhibits high relative abundances of
317 | specific OTUs of Bacillariophyceae, Chrysophyte and Cryptophyta. In pond systems,
318 | autotrophic communities made up mainly of other specific OTUs of Chlorophyta,

319 Chrysophyte, and Cryptophyta. It is also worth noting that Dinophyta and Streptophyta
320 were found exclusively in pond systems (in particular in MG and SR, respectively).

321 The community organization of microbial eukaryotes was highlighted from a
322 cross-group co-occurrence network between autotrophic and heterotrophic individuals.

323 Among 226 dominant autotrophic and heterotrophic OTUs, 204 displayed 274
324 associations (Figure 2). From these associations between OTUs in the network, more
325 positive (98.5%) than negative associations were inferred. All negative associations
326 occurred between OTUs assigned to Ciliphora for heterotrophic microorganisms and
327 Cryptophyta for autotrophic microorganisms. No clear association patterns can be
328 identified in network from raw tables of autotrophic and heterotrophic microbial
329 eukaryotes (Figure 2a). When the co-occurrence network is reordered according to
330 species scores on the first axis of sCoCA, two modules can be distinguished (Figure 2b).

331 The first module (i.e., top right corner) is constituted by OTUs exhibiting the higher
332 relative abundances in brook system (i.e., RSA), while OTUs that compose the second
333 module (i.e., bottom left corner) dominate pond systems (i.e., EV, LC, MG, and SR).

334 Heterotrophic OTUs exhibited major associations with autotrophic OTUs belonging to
335 the same module, with only 1.5% of associations between OTUs from distinct module. A

336 striking pattern is that Chrysophyte is the autotrophic group that contributes most to
337 associations in the two modules (module 1: 71%, module 2: 41%), whereas for the
338 heterotrophic group it is fungi in the module 1 (47%) and Ciliophora in the module 2
339 (59%). In pond systems (i.e., module 2), surprisingly, fungi are involved in very few
340 associations (7%).

341 |

342 | **Case study 2: Microalgae-virus system**

343 | The cross-validation procedure identified the best pCoCA model based on the
344 | first two significant axes (pCoCA axis 1: $p=0.001$, pCoCA axis 2: $p=0.001$), in which
345 | Mamiellophyceae community predicted 32.02% of the variation in *Prasinovirus*
346 | community. The first two axes accounted for 37.47% (24.26% and 13.21% for axis 1 and
347 | 2 respectively) and 44.82% (28.01% and 16.81%) of the variation in the structure of
348 | Mamiellophyceae and *Prasinovirus* community respectively. The biplots indicated that
349 | the two communities covaried along a lagoon (samples LAs)/open-sea gradient
350 | (samples MAs) on the first axis (from left to right), while a temporal gradient for site LA
351 | (intra-site variability) could be identified along the second axis (Figure 3). OtVs have a
352 | higher prevalence in the lagoon samples (especially LAs) and coastal samples (SAs) in
353 | which *Ostreococcus* exhibited a high density (Figure 3a, b). Conversely, open-sea
354 | samples (MAs) were dominated by *Bathycoccus* which supported virus assemblages
355 | dominated by BpVs. *Micromonas* showed a wider distribution, with a relative
356 | contribution of its OTUs both in lagoon samples, coastal samples and open-sea samples,
357 | associated with a similar repartition of MpVs (Figure 3a, b).

358 | Based on [the cross-group co-occurrence networks](#) analysis, 67 associations were
359 | identified between the major 67 OTUs assigned to one of the three groups of
360 | Mamiellophyceae (i.e. *Bathycoccus*, *Micromonas*, and *Ostreococcus*) and the major 98
361 | OTUs assigned to *Prasinovirus* (i.e. BpVs, MpVs, and OtVs) (Figure 4a). Reordering the
362 | [co-occurrence](#) network according to the species scores on the first axis of pCoCA
363 | highlighted a structure in the network (Figure 4b). The network topology suggests that
364 | the identity of OTUs contained in co-occurring groups of viruses and microalgae are
365 | related to their respective prevalence along the lagoon/open-sea gradient. Virus OTUs
366 | mostly present in lagoon samples have significant associations primarily with
367 | microalgae OTUs displaying the higher prevalence in lagoon samples (top right corner,

368 Figure 4b). Similarly, virus OTUs dominating the open-sea samples were mainly
369 associated with microalgae OTUs from open-sea samples (bottom left corner, Figure 4b).
370 Among the associations contained in the network, 50 were identified between OTUs
371 belonging to an expected host-virus system (i.e. associations *Bathycoccus*/BpV,
372 *Micromonas*/MpV, and *Ostreococcus*/OtV) while the other 17 significant associations
373 were found between OTUs belonging to different host-virus systems. In average 50%,
374 70.6% and 91.3% of associations found for OTUs of BpVs, MpVs, and OtVs respectively
375 were with OTUs assigned to their respective host group. Within the associations, some
376 single Mamiellophyceae OTUs were associated with many *Prasinovirus* OTUs and
377 reciprocally. On the other hand, dyads were identified, that is specific associations, in
378 *Bathycoccus*/BpV, *Micromonas*/MpV and *Ostreococcus*/OtV systems. Few negative
379 associations inferred from the observation that those OTUs do not co-occur were found
380 in network (Figure 4b). Interestingly, eight negative associations from a total of nine
381 involved virus OTUs and microalgae OTUs belonging to different host-virus systems.

382

383 **Discussion**

384 Critical review and guidance papers on the analysis of NGS-based community
385 data (Ramette, 2007; Buttigieg & Ramette, 2014; Paliy & Shankar, 2016) do not mention
386 any direct quantitative method for predicting the composition of one community from
387 another. Co-correspondence analysis (ter Braak & Schaffers, 2004) fills this gap. At the
388 level of [the case study 1](#), symmetric form of CoCA indicated that heterotrophic microbial
389 eukaryote assemblages in shallow freshwater ecosystems were strongly associated with
390 patterns of autotrophic microbial eukaryotes presence and abundance with links that
391 can be taxon-specific (Figure 1). Our results shown also that the composition of
392 heterotrophic microbial eukaryote community was dominated by fungi in brook system

393 compared to ponds. This is in line with recent observations, based on the estimation of
394 ergosterol level, of a generally higher fungal biomass in river than in ponds (Baldy et al.,
395 2002). Higher fungal abundance might potentially be linked to incoming resources from
396 runoff, since in brooks the most important source of imported material is usually
397 deciduous leaves, whose the decomposition processing involved to a large extent fungi
398 (Bärlocher, 1985; Webster & Benfield, 1986). To this, the composition of heterotrophic
399 microbial eukaryote community characterizing brook system is associated a specific
400 composition of microbial autotrophs. Such a result suggests that heterotroph
401 community composition exert a control on the composition of autotroph community,
402 and that microbial autotrophs can be driver of microbial heterotrophs.

403 Regarding the case study 2, the predictive form of CoCA points out different
404 distribution patterns among the three groups of *Prasinovirus* along the lagoon/open-sea
405 gradient (Figure 3). Note the importance of the dimension reduction step in pCoCA that
406 allows focusing on ecological structures depicted on a limited number of axes and
407 removes random variation from the data. The patterns in *Prasinovirus* assemblages, with
408 a dominance of OtV in lagoon and coastal samples compared to offshore locations, the
409 inverse distribution for BpV, and MpV exhibiting a wider spatial distribution, are in part
410 a consequence of the presence of their respective hosts in lagoon, coastal and open-sea
411 samples. Indeed, *Ostreococcus* is known to be abundant in lagoons (Subirana et al.,
412 2013), more eutrophic system, compared to *Bathycoccus*, which is found mainly in
413 oligotrophic areas (Vaulot et al., 2012; Wu, Huang, & Zhong, 2013) such as offshore sites
414 (i.e., MA). *Micromonas* is ubiquitous and particularly present in nutrient-rich
415 environments (Not et al., 2004; Viprey, Guillou, Ferréol, & Vaulot, 2008). Our findings
416 confirm also the data of Bellec et al. (2010) showing that OtV are more abundant in
417 lagoon than in the open sea.

418 Cross-taxon congruence description and evaluation (Virtanen, Ilmonen,
419 Paasivirta, & Muotka, 2009; Gioria, Bacaro, & Feehan, 2011) provides a more
420 comprehensive picture of [the](#) community similarity than the richness metrics
421 conventionally used (Wolters, Bengtsson, & Zaitsev, 2006; Westgate, Barton, Lane, &
422 Lindenmayer, 2014). Our results reinforce the need to use CoCA to study the cross-taxon
423 congruence in microbial communities from NGS data. This is all the more important
424 because the high co-correspondence between the two functional groups in microbial
425 eukaryote community may be especially informative given the key ecological role of
426 microbial eukaryotes (Caron et al., 2012). In addition, the study of cross-taxon
427 congruence between Mamiellophyceae and *Prasinovirus* is of definite interest in marine
428 ecosystems, warmed by climate change, [where](#) the expected gradual shift towards small
429 primary producers could render the role of small eukaryotes more important than they
430 are today (Morán, López-Urrutia, Calvo-Díaz, & Li, 2010). Microbial eukaryotes are
431 recognized as a significant contributor across various geographical locations of
432 picophytoplankton (Worden, Nolan, & Palenik, 2004; Jardillier, Zubkov, Pearman, &
433 Scanlan, 2010), which accounts for > 50% of phytoplankton biomass and productivity in
434 marine ecosystems (Maranon et al., 2001; Teira et al., 2005).

435 Previous studies have suggested that biotic interactions are the most likely
436 mechanisms underlying cross-taxon congruence at local scales (Jackson & Harvey, 1993;
437 Johnson & Hering, 2010), although concordance is also expected from similar responses
438 to environmental gradient (Bini, Vieira, Machado, & Machado Velho et al., 2007; Rooney
439 & Bayley, 2012). As an important implication, the level of congruence can inform about
440 the structural pattern among interacting groups (Özkan et al., 2014). That being said,
441 reordering [co-occurrence](#) networks, from the species scores on the first axis of CoCA,
442 allows to identify structural patterns in [co-occurrence](#) networks of microbial eukaryote

443 | community (Figure 2b) and in the Mamiellophyceae/*Prasinovirus* system (Figure 4b).
444 For example, the structure of the microbial eukaryote network is characterized by two
445 modules underlying a brook/pond gradient in the composition of heterotrophic and
446 autotrophic microbial eukaryote assemblages. These differences in OTU composition
447 within the two modules suggest that the food-web structure is different between lotic
448 and lentic ecosystems, and reinforce the differences previously observed at the
449 bacterioplankton level (Portillo, Anderson, & Noah, 2012). A substantial effort has been
450 made in the development of metrics to estimate and test the level of nestedness (e.g.,
451 Rodriguez-Guirones & Santamaria, 2006; Ulrich & Gotelli, 2007) and modularity (e.g.,
452 Barber, 2007; Dorman & Strauss, 2014) within interaction networks. The order of
453 individuals of two groups in bipartite matrices affects the magnitude of metrics that
454 represent deviations from an idealized state (e.g., perfect nestedness or modularity)
455 (Almeida et al., 2008). It has then been advocated that prior to analysis of structure of
456 networks, original bipartite matrices should be reordered to maximize the coherence of
457 individual distributions in rows and columns, so that individuals with most similar links
458 are close together (Borgatti & Everett, 1997; Leibold & Mikkelsen, 2002). These findings
459 taken together with our results validate our proposition to combine CoCA with network
460 analysis to study structural patterns of microbial networks. In addition, in reverse to the
461 expected view of nestedness structure of phage-bacteria network (Flores, Meyer,
462 Valverde, Farr, & Weitz, 2011), the modular structure of Mamiellophyceae/*Prasinovirus*
463 network observed in the field underpinned the modularity patterns previously observed
464 in phage-bacteria network from cross-infection experiments (Flores, Valverde, & Weitz,
465 2013). Given that the structure of interaction networks is constraint by the
466 coevolutionary processes between species (Peralta, 2016), this would lead to account
467 for phylogenetic signals within co-occurrence network (Derocles et al., 2018). In this

468 [context, it would be possible to disentangle the confounding effect of phylogeny from](#)
469 [true biotic interactions by developing a partial analysis \(ter Braak, Šmillauer & Dray,](#)
470 [2018\) in the context of CoCA to partial out the phylogenetic effect and focus on patterns](#)
471 [of co-occurrence that are not related to phylogenetic signal.](#)

472 All computational methods used to infer networks from NGS data sets produce
473 species [co-occurrence](#) networks, where a link between two species represents a
474 significant statistical association (positive or negative) between their abundance. This
475 raises a critical issue about the interpretation of inferred associations (Derocles et al.,
476 2018), because [co-occurrence](#) networks differ from interaction networks constructed on
477 observations of both the species and their interactions (Ings et al., 2009). For instance,
478 all inferred associations between Mamiellophyceae and *Prasinovirus* belonging to
479 expected host-virus system were positive. These results are consistent with a previous
480 work showing that when parasitism is captured as a significant link in [co-occurrence](#)
481 network, it is retrieved as a positive link despite the detrimental effect of parasite on its
482 host (Weiss et al., 2016). This might be explained because the copresence of the host
483 species and the parasite species is necessary for the interaction to occur. Another
484 surprising result is that all negative associations (expected one) were between
485 Mamiellophyceae and *Prasinovirus* belonging to different host-virus systems. Such
486 negative associations may account for opposite abiotic requirements, since in our case,
487 OTUs of concerned viruses and microalgae had inverse spatio-temporal dynamics.
488 Positive associations were also found between individuals of different host-virus
489 systems, which could be explained by the increase of a *Prasinovirus* population that
490 removes, by infection, a major competitor of a co-occurring Mamiellophyceae host of
491 another group. It is important to keep in mind that, although these associations between
492 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

518 consequently its readability, as proposed in good practice of data visualization
519 (Spiegelhalter, Pearson, & Short, 2011; Kjærsgaard, 2015; Pocock et al., 2016).
520 Interestingly, the applicability of our approach goes beyond the particular case of data
521 sets with row-sum normalisation (i.e., compositional data). Indeed, CoCA was originally
522 designed to analyze abundance data and is thus able to deal with counts without the
523 need to rarefy data, in accordance with the recent advice against rarefaction (McMurdie
524 & Holmes, 2014). It paves the way for further studies to examine the cross-taxon
525 congruence and structural pattern of co-occurrence networks in microbial communities
526 and in turn their effects on the ecosystem functioning.

527

528 **Acknowledgments**

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

532 References

- 533 Almeida-Neto M, Guimarães P., Guimarães P. R. Jr., Loyola R. D., & Ulrich W. (2008). A
534 consistent metric for nestedness analysis in ecological systems: reconciling concept
535 and measurement. *Oikos*, 117, 1127–1239.
- 536 Amaral-Zettler, L. A., McCliment E. A., Ducklow H. W., & Huse S.M. (2009). A method for
537 studying protistan diversity using massively parallel sequencing of V9 hypervariable
538 regions of small-subunit ribosomal RNA genes. *PLoS One*, 4, e6372.
- 539 Baldy V., Chauvet E., Charcosset J.-Y., & Gessner M. O. (2002). Microbial dynamics
540 associated with leaves decomposing in the mainstem and floodplain pond of a large
541 river. *Aquatic Microbial Ecology*, 28, 25–36.
- 542 Bálint M., Bahram M., Eren A. M., Faust K., Fuhrman J. A., Lindahl B., ... Tedersoo L.
543 (2016). Millions of reads, thousands of taxa: microbial community structure and
544 associations analyzed via marker genes. *FEMS Microbial Reviews* 40, 686–700.
- 545 Ban Y, An L., & Jiang H. (2015). Investigating microbial co-occurrence patterns based on
546 metagenomics compositional data. *Bioinformatics*, 31, 3322–3329.
- 547 Banerjee S., Kirkby C. A., Schmutter D., Bissett A., Kirkegaard J. A., & Richardson A. E.
548 (2016). Network analysis reveals functional redundancy and keystone taxa amongst
549 Barber M. J. (2007). Modularity and community detection in bipartite networks. *Physical
550 Review*, 76, 066102.
- 551 Bärlocher F. (1985). The role of fungi in the nutrition of stream invertebrates. *Botanical
552 Journal of the Linnean Society*, 91, 83–94.
- 553 Bellec L., Grimsley N., Derelle E., Moreau H., & Desdevises Y. (2010). Abundance, spatial
554 distribution and genetic diversity of *Ostreococcus tauri* viruses in two different
555 environments. *Environmental Microbiology Reports*, 2, 313–321.
- 556 Benzécri F. (1969). Statistical analysis as a tool to make patterns emerge from data. In
557 Watanabe (Ed.), *Methodologies of pattern recognition*. (pp 35–60). Academic Press:
558 New York.
- 559 Bergelson J., Mittelstrass J., & Horton M. W. (2019). Characterizing both bacteria and
560 fungi improves understanding of the *Arabidopsis* root microbiome. *Scientific Reports*,
561 9, 24.
- 562 Bilton D. T., McAbendorth L., Bedford A., & Ramsay P. M. (2006). How wide to cast the
563 net? Cross-taxon congruence of species richness, community similarity and indicator
564 taxa ponds. *Freshwater Biology*, 51, 578–590.
- 565 Bini L. M., Vieira L. C. G., Machado J., & Machado Velho L. F., (2007). Concordance of
566 species patterns among micro-crustaceans, rotifers and testate amoebae in a shallow
567 pond. *International Review of Hydrobiology*, 92, 9–22.
- 568 Borgatti S. P., & Everett M. G. (1997). Network analysis of 2-mode data. *Social Networks*,
569 19, 243–269.
- 570 Buttigieg P. L., & Ramette A. (2014). A guide to statistical analysis in microbial ecology: a
571 community-focused, living review of multivariate data analyses. *FEMS Microbiology
572 Ecology*, 90, 543–550.
- 573 Cannon M. V., Craine J., Hester J., Shalkhauser A., Chan E. R., Logue K., ... Serre D. (2017).
574 Dynamic microbial populations along the Cuyahoga River. *PLoS ONE*, 12, e0186290.
- 575 Cardona C., Weisenhorn P., Henry C., & Gilbert J. A. (2016). Network-based metabolic
576 analysis and microbial community modeling. *Current Opinion in Microbiology*, 31,
577 124–131.
- 578 Caron D. A., Countway P. D., Jones A. C., Kim D. Y., & Schnetzer A. (2012). Marine
579 protistan diversity. *Annual Review of Marine Science*, 4, 467–493.

580 Chen F., & Suttle C. A. (1995). Amplification of DNA polymerase gene fragments from
581 viruses infecting microalgae. *Applied and Environmental Microbiology*, 61, 1274–
582 1278.

583 Clerissi C., Grimsley N., Ogata H., Hingamp P., Poulain J., & Desdevises Y. (2014).
584 Unveiling of the diversity of *Prasinoviruses* (Phycodnaviridae) in marine samples by
585 using high-throughput sequencing analyses of PCR-amplified DNA polymerase and
586 major capsid protein genes. *Applied and Environmental Microbiology*, 80, 3150–3160.

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

589 Derocles S. A. P., Bohan D. A., Dumbrell A. J., Kitson J. J. N., Massol, F., Pauvert C., ... Evans
590 D. M. (2018) Biomonitoring for the 21st century: Integrating next-generation
591 sequencing into ecological network analysis. *Advances in Ecological Research*, 58, 3–
592 62.

593 Dolédec S., & Chessel D. (1994). Co-inertia analysis: an alternative method for studying
594 species-environment relationships. *Freshwater Biology*, 31, 277–294.

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

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

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

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

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

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

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

609 Faust K., Sathirapongsasuti J. F., Izard J., Segata N., Gevers D., Raes J., & Huttenhower C.
610 (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS*
611 *Computational Biology*, 8, e1002602.

612 Fierer N., Breitbart M., Nulton J., Salamon P., Lozupone C., Jones R., ... Jackson R. B.
613 (2007) Metagenomic and Small-Subunit rRNA analyses reveal the genetic diversity of
614 bacteria, archaea, fungi, and viruses in soil. *Applied and Environmental Microbiology*,
615 73, 7059–7066.

616 Flores C. O., Meyer J. R., Valverde S., Farr L., & Weitz J. S. (2011). Statistical structure of
617 host-phage interactions. *Proceedings of the National Academy of Sciences of the United*
618 *States of America*, 108, E288–E297.

619 Flores C. O., Valverde S., & Weitz J. S. (2013). Multi-scale structure and geographic
620 drivers of cross-infection within marine bacteria and phages. *ISME Journal*, 7, 520–
621 532.

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

624 Fu L., Niu B., Zhu Z., Wu S., Li W. (2012). CD-HIT: accelerated for clustering the next-
625 generation sequencing data. *Bioinformatics*, 28, 3150–3152.

626 Fuhrman J. A. (2009). Microbial community structure and its functional implications.
627 *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 Microbiology*, 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 CO₂ 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 Kjærsgaard 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., & Mikkelsen 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. Vaultot 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/>

726 Ramette A. (2007). Multivariate analyses in microbial ecology. *FEMS Microbiology*
727 *Ecology*, 62, 142–160.

728 Rappé M. S., & Giovannoni S. J. (2003). The uncultured microbial majority. *Annual Review*
729 *of Microbiology*, 57, 369–394.

730 Rodrigueuz-Gironés M. A., & Santamaría L. (2006). A new algorithm to calculate the
731 nestednes temperature of presence-absence matrices. *Journal of Biogeography*, 33,
732 924–935.

733 Rooney R. C., & Bayley S. E.. (2012). Community congruence of plants, invertebrates and
734 birds in natural and constructed shallow open-water wetlands: Do we need to
735 monitor multiple assemblages? *Ecological Indicators*, 20, 2012, 42–50.

736 Santi E., Maccherini S., Rocchini D., Bonini I., Brunialti G., Favilli L., ... Chiarucci A. (2010).
737 Simple to sample: vascular plants as surrogate group in a nature reserve. *Journal for*
738 *Nature Conservation*, 18, 2–11.

739 Schloss P. D., Westcott S. L., Ryabin T., Hall J. R., Hartmann M., Hollister E. B., ... Weber C.
740 F. (2009). Introducing mothur: Open-source, platform-independent, community-
741 supported software for describing and comparing microbial communities. *Applied*
742 *and Environmental Microbiology*, 75, 7537–7541.

743 Simon M., Jardillier L., Deschamps P., Moriera D., Restoux G., Bertolino P., & López-García
744 P. (2014). Complex communities of small protists and unexpected occurrence of
745 typical marine lineages in shallow freshwater systems. *Environmental Microbiology*,
746 17, 3610–3627.

747 Simon M., López-García P., Deschamps P., Moreira D., Restoux G., Bertolino P., & Jardillier
748 L. (2015). Marked seasonality and high spatial variability of protist communities in
749 shallow freshwater systems. *ISME Journal*, 9, 1941–1953.

750 Simpson G. L. (2009). Cocorresp: Co-correspondence analysis ordination methods. (R
751 package version 0.3-0). (<http://cran.r-project.org/package=analogue>).

752 Spencer S. J., Tamminen M. V., Preheim S. P., Guo M. T., Briggs A. W., Brito I. L., ... Alm E. J.
753 (2016). Massively parallel sequencing of single cells by epicPCR links functional genes
754 with phylogenetic markers. *ISME Journal*, 10, 427–436.

755 Spiegelhalter D., Pearson M., & Short I. (2011). Visualizing uncertainty about the future.
756 *Science*, 333, 1393–1400.

757 Subirana L., Péquin B., Michely S., Escande M.-L., Meilland J., Derelle E., ... Grimsley N. H.
758 (2013). Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus*
759 reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae).
760 *Protist*, 164, 643–659.

761 Teira E., Mourino B., Marañón E., Perez V., Pazo M. J., Serret P., ... Fernández, E. (2005).
762 Variability of chlorophyll and primary production in the Eastern North Atlantic
763 Subtropical Gyre: potential factors affecting phytoplankton activity. *Deep-Sea*
764 *Research Part I*, 52, 569–588.

765 ter Braak C. J. F., & Schaffers A. P. (2004). Co-correspondence analysis: a new ordination
766 method to relate two community compositions. *Ecology*, 85, 834–846.

767 ter Braak C. J. F. (1990). Interpreting canonical correlation analysis through biplots of
768 structural correlations and weights. *Psychometrika*, 55, 519–531.

769 Tipton L., Müller C. L., Kurtz Z. D., Huang L., Kleerup E., Morris A., ... Ghedin E. (2018).
770 Fungi stabilize connectivity in the lung and skin microbial ecosystems. *Microbiome*, 6,
771 12.

772 Ulrich W., & Gotelli N. J. (2007). Null model analysis of species nestedness patterns.
773 *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 *Ecology*, 52, 79–92.

815 Zimmerman N., Izard J., Klatt C., Zhou J., & Aronson E. (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.

821 **Data accessibility statement**

822 | Data and R scripts to reproduce the different analyses of the [case study 1 and the case](#)
823 | [study 2 Mamiellophyceae/Prasinovirus system](#) are available online [in the Appendix 1 as](#)
824 | [additional supporting information](#).
825 |

826 | **Author contributions**

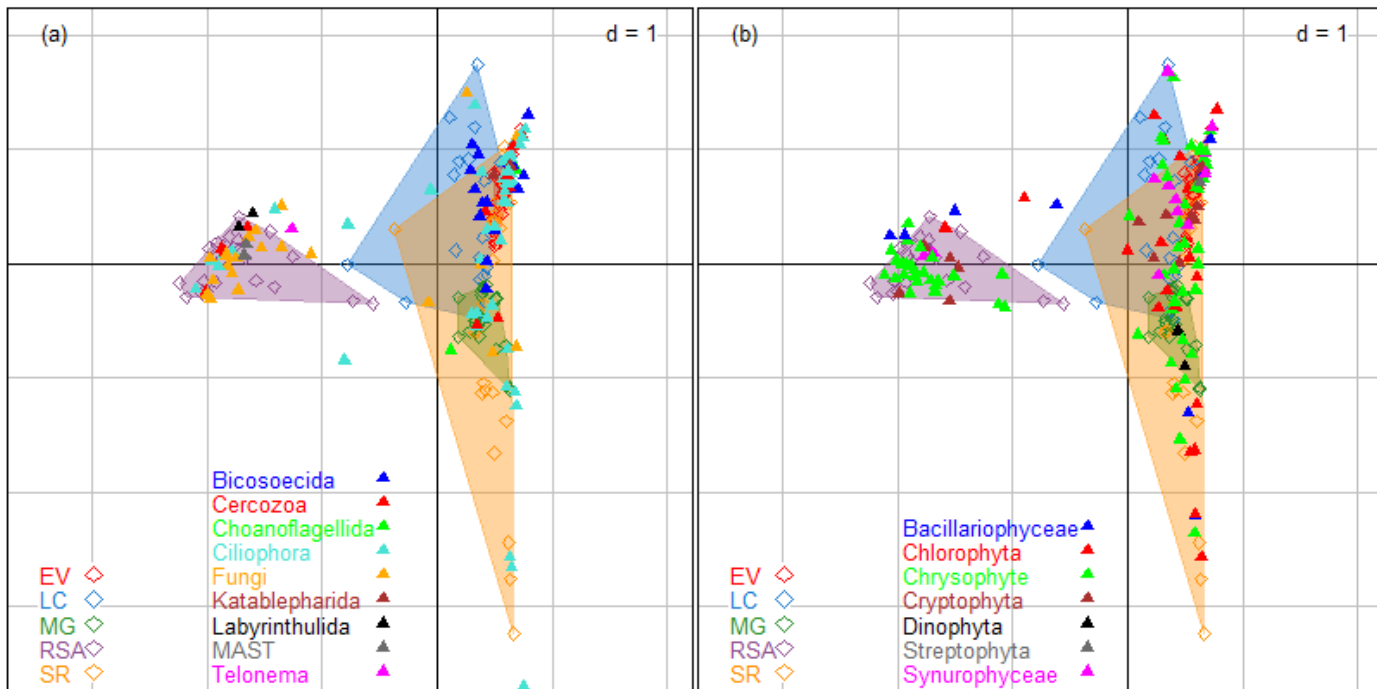
827 | Y.D. and S.D. supervised the project; B.A. performed all the statistical analyses; C.J.F.B.
828 | contributed to mathematical development of the method; H.L. contributed to collecting
829 | and genotyping the biological materials of microalgae and viruses; B.A. wrote the first
830 | draft of the manuscript; B.A., C.J.F.B., Y.D., and S.D. commented and approved the final
831 | version of the manuscript.
832 |

833 | **Supporting information**

834 | [Additional supporting information may be found online in the Supporting Information](#)
835 | [section at the end of the article](#).
836 |

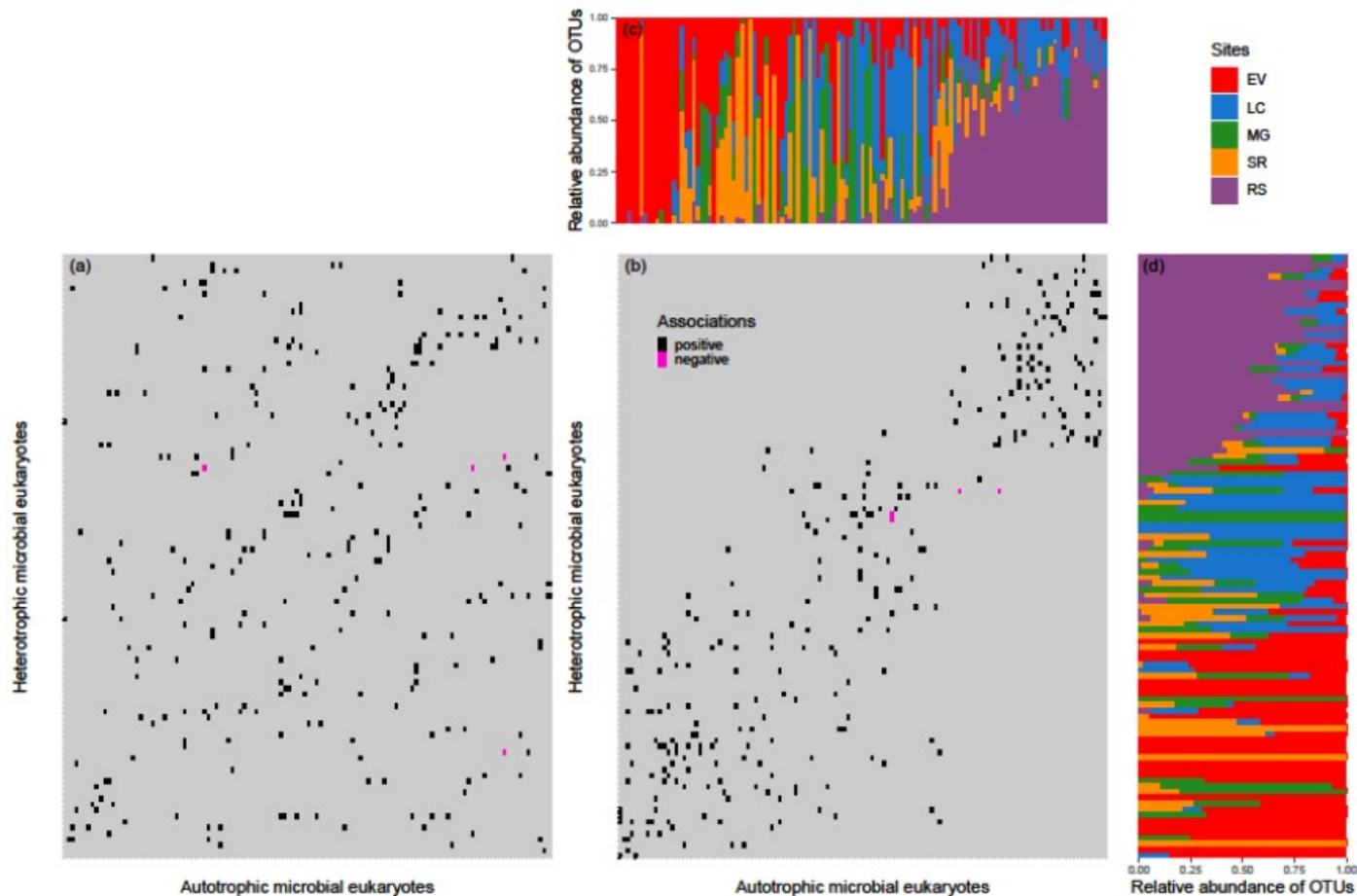
837 **Figures**

838 **Figure 1.** Ordination biplots of the case study 1, representing of the communities of the
 839 positions of sites (open diamond) and species (solid triangle) on the axis 1 × axis 2
 840 factorial plan of the symmetric co-correspondence analysis. (a) Biplot for the
 841 heterotrophic microbial eukaryotes and (b) biplot for the autotrophic microbial
 842 eukaryotes OTUs obtained from symmetric co-correspondence analysis. EV: Etang
 843 des Vallées, LC: La Claye, GB: Mare Gabard, SR: Saint Robert, RSA: Ru Sainte Anne.
 844 Heterotrophic and autotrophic microbial eukaryotes OTUs are colored according to
 845 the phylogenetic group they belong to. "d" indicates the mesh of the grid.
 846



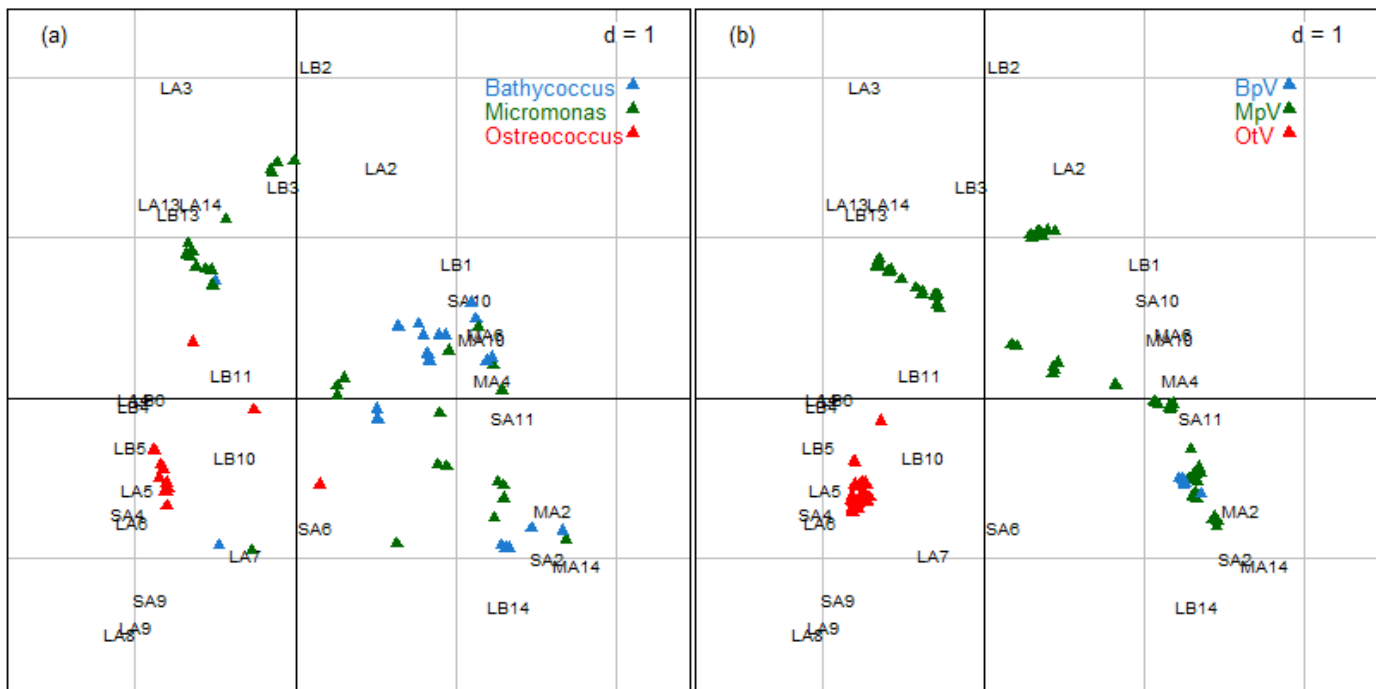
847
 848
 849

850 **Figure 2.** Heatmap of the case study 1, representing the association network between
 851 the heterotrophic and autotrophic microbial eukaryotes (a) before (a) and (b) after
 852 (b) the reordering the position of each species in the network according to their
 853 network from the species scores on the first axis of symmetric co-correspondence
 854 analysis. Bar plots of the relative abundance of (c) autotrophic microbial eukaryotes
 855 and (d) heterotrophic microbial eukaryotes. Each OTU is represented by a vertical
 856 line partitioned into segments corresponding to its relative abundance in one of five
 857 sites. EV: Etang des Vallées, LC: La Claye, GB: Mare Gabard, SR: Saint Robert, RSA: Ru
 858 Sainte Anne.
 859



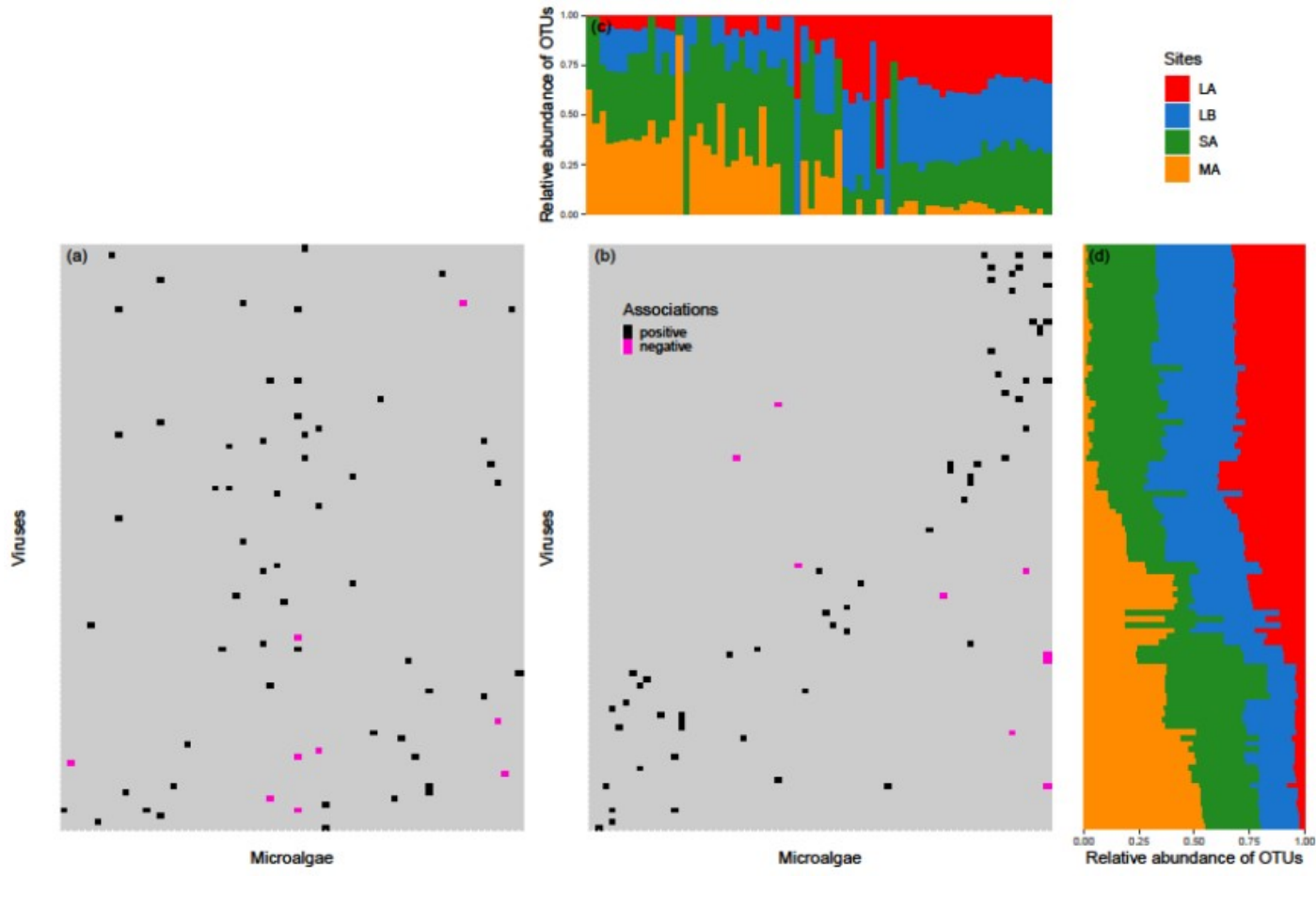
860
 861
 862

863 **Figure 3.** Ordination biplots of, concerning the case study 2, of representing the
 864 positions of sites (open diamond) and species (solid triangle) on the axis 1 × axis 2
 865 factorial plan of the predictive co-correspondence analysis. (a) Biplot for the
 866 Mamiellophyceae and (b) biplot for the Prasinovirus OTUs obtained from predictive
 867 co-correspondence analysis. LB: coastal site in Leucate lagoon, LA: site at the level of
 868 the Grau in Leucate lagoon, SA: coastal site, MA: open-sea site. Microalgae and virus
 869 OTUs are colored according to the phylogenetic group they belong to. BpV:
 870 *Bathycoccus* viruses, MpV: *Micromonas* viruses, OtV: *Ostreococcus* viruses. "d"
 871 indicates the mesh of the grid.
 872



873 **Figure 3.** Heatmap of heterotrophic and autotrophic microbial eukaryotes before (a)-
 874 and after (b) the reordering the network from the species scores on the first axis of
 875 symmetric co-correspondence analysis. Bar plots of the relative abundance of (c)-
 876 autotrophic microbial eukaryotes and (d) heterotrophic microbial eukaryotes. Each
 877 OTU is represented by a vertical line partitioned into segments corresponding to its
 878 relative abundance in one of five sites. EV: Etang des Vallées, LC: La Claye, GB: Mare-
 879 Gabard, SR: Saint Robert, RSA: Ru Sainte Anne.
 880
 881
 882
 883
 884

885 **Figure 4.** Heatmap of the case study 2, representing the microalgae-viruses
 886 (*Mamiellophycaea/Prasinovirus*) association network (a) before (a) and (b) after (b)
 887 the reordering the position of each species in the network according to their network-
 888 from the species scores on the first axis of predictive co-correspondence analysis. Bar
 889 plots of the relative abundance of (c) microalgae and (d) viruses. Each OTU is
 890 represented by a vertical line partitioned into segments corresponding to its relative
 891 abundance in one of four sites. LA: site at the level of the Grau in Leucate lagoon, LB:
 892 coastal site in Leucate lagoon, SA: coastal site, MA: open-sea site.
 893
 894



895
 896