Disentangling environmental effects in microbial association networks

Ina Maria Deutschmann (ina.m.deutschmann@gmail.com)
Institute of Marine Sciences (ICM-CSIC) https://orcid.org/0000-0002-3512-261X

Gipsi Lima-Mendez
Louvain Institute of Biomolecular Science and Technology (IBST), University catholique de Louvain

Anders K. Krabberød
Department of Biosciences/Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo

Jeroen Raes
VIB Center for Microbiology, KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology

Sergio M. Vallina
Spanish Institute of Oceanography (IEO)

Karoline Faust
KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology

Ramiro Logares
Institute of Marine Sciences (ICM-CSIC)

Methodology

Keywords: microbial interactions, association network, effect of indirect dependencies, environmentally-driven edge detection

Posted Date: August 17th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-57387/v1

Version of Record: A version of this preprint was published at Microbiome on November 26th, 2021. See the published version at https://doi.org/10.1186/s40168-021-01141-7.

Disentangling environmental effects in microbial association networks

Ina Maria Deutschmann[1*] (ina.m.deutschmann@gmail.com), Gipsi Lima-Mendez[2] (gipsi.lima-mendez@uclouvain.be), Anders K. Krabberød[3] (a.k.krabberod@ibv.uio.no), Jeroen Raes[4,5] (jeroen.raes@kuleuven.vib.be), Sergio M. Vallina[6] (sergio.vallina@oceanglobe.org), Karoline Faust[5*†] (karoline.faust@kuleuven.be) and Ramiro Logares[1*] (ramiro.logares@icm.csic.es)

[1] Institute of Marine Sciences, CSIC, Passeig Marítim de la Barceloneta, 37, 08003, Barcelona, Spain.
[2] Louvain Institute of Biomolecular Science and Technology (IBST), Université catholique de Louvain, Croix du sud 4-5/L7.07.02, 1348, Louvain-la-Neuve, Belgium.
[3] Department of Biosciences/Section for Genetics and Evolutionary Biology (EVOGENE), University of Oslo, p.b. 1066 Blindern, N-0316, Oslo, Norway.
[4] VIB Center for Microbiology, Herestraat 49-1028, 3000, Leuven, Belgium.
[5] KU Leuven Department of Microbiology, Immunology and Transplantation, Rega Institute, Laboratory of Molecular Bacteriology, Leuven, Belgium, Herestraat 49, 3000, Leuven, Belgium.
[6] Spanish Institute of Oceanography (IEO), Ave Principe de Asturias 70 Bis, 33212, Gijon, Spain.

[*] Corresponding authors

[†] Shared last authors

Manuscript for: BMC Microbiome as Methodology paper

Abstract

Background

Ecolocial interctions among microorganisms are fundamental for ecosystem function, yet they are mostly unknown or poorly understood. High-throughput-omics can indicate microbial interactions by associations across time and space, which can be represented as association networks. Links in these networks could result from either ecological interactions between microorganisms, or from environmental selection, where the association is environmentally-driven. Therefore, before downstream analysis and interpretation, we need to distinguish the nature of the association, particularly if it is due to environmental selection or not.

Results

We present EnDED (Environmentally-Driven Edge Detection), an implementation of four approaches as well as their combination to predict which links between microorganisms in an association network are environmentally-driven. The four approaches are Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality. We tested EnDED on networks from simulated data of 50 microorganisms. The networks contained on average 50 nodes and 1,087 edges, of which 60 were true interactions but 1,026 false associations (i.e. environmentally-driven or due to chance). Applying each method individually, we detected a moderate to high number of environmentally-driven edges—87% Sign Pattern and Overlap, 67% Interaction Information, and 44% Data Processing Inequality. Combining these methods in an intersection approach resulted in retaining more interactions, both true and false (32% of environmentally-driven associations). The addition of noise to the simulated datasets did

not alter qualitatively these results. After validation with the simulated datasets, we applied EnDED on a marine microbial network inferred from 10 years of monthly observations of microbial-plankton abundance. The intersection combination predicted that 14.2% of the associations were environmentally-driven, while individual methods predicted 31.4% (Data Processing Inequality), 38.3% (Interaction Information), and up to 83.4% (Sign Pattern as well as Overlap).

Conclusions

To reach accurate hypotheses about ecological interactions, it is important to determine, quantify, and remove environmentally-driven associations in marine microbial association networks. For that, EnDED offers up to four individual methods as well as their combination. However, especially for the intersection combination, we suggest to use EnDED with other strategies to reduce the number of false associations and consequently the number of potential interaction hypotheses.

Keywords: microbial interactions; association network; effect of indirect dependencies; environmentally-driven edge detection

Background

Association networks to generate microbial interaction hypotheses

There is a myriad of microorganisms on Earth: current estimates indicate ≈1012 microbial species Locey and Lennon (2016), and ≈1030 microbial cells Kallmeyer et al. (2012); Whitman et al. (1998). Microorganisms have crucial roles in the biosphere by

contributing to global biogeochemical cycles Falkowski et al. (2008) and underpinning diverse food webs. The importance of microbes for the functioning of ecosystems cannot be understood without considering their ecological interactions DeLong (2009); Krabberød et al. (2017). These allow transferring carbon and energy to upper trophic levels, and the recycling of nutrients and energy Worden et al. (2015). Furthermore, ecological interactions influence microbial community turnover and composition. These interactions include win-win (e.g. mutual cross-feeding and cooperation), win-loss (e.g. predator-prey and host-parasite), and loss-loss (e.g. resource competition) relationships Faust and Raes (2012). Although microbial communities are highly interconnected Layeghifard et al. (2017), our knowledge about ecological interactions in the microbial world is still limited Bjorbækmo et al. (2019); Krabberød et al. (2017).

Previous studies have shown relationships between a restricted number of microorganisms. However, we need a large number of interactions to understand the functioning of such complex ecosystems. This is challenging, in part, due to the vast number of possible interactions—given n microorganisms, there are $\binom{n}{2} = n(n-1)/2$ potential pairwise interactions. Thus, it is unfeasible to test them experimentally within a reasonable amount of time and cost. The problem of having a large number of potential interactions can be partially circumvented with omics technologies coupled to network analyses.

Omics can identify and quantify a large number of microorganisms from a given sample. Typically, the relative abundance for each identified organism per sample is determined. There are multiple methods to determine associations (normally based on correlations) between microorganisms using their abundances (e.g. eLSA Xia et al. (2011, 2012), CoNet Faust and Raes (2016), SPIEC-EASI Kurtz et al. (2015), or FlashWeave

Tackmann et al. (2019)). These abundance-based associations compose a network, where nodes represent microorganisms and edges represent either co-presence (positive association) or mutual exclusion (negative association) relationships, which constitute microbial interaction hypotheses.

Challenges in using networks as a representation of the microbial ecosystem

Although networks play an essential role in understanding complex systems, microbial ecological networks are not yet as developed in terms of inference and biological interpretation Lv et al. (2019). Network inference from -omics data is difficult Layeghifard et al. (2017); Li et al. (2016) because of both technical and interpretation challenges.

One common technical challenge is the compositional effect—microbial counts become interdependent if normalized by the sample's total number of counts. There are several tools Li et al. (2016) that correct for the compositional effect (e.g. SPIEC-EASI Kurtz et al. (2015)). Other difficulties include data based on a small number of samples relative to the number of microorganisms they contain (i.e. a low sample-to-microorganisms ratio); plus sparse data—too many zeros in the dataset that can wrongly associate microorganisms Aitchison (1981). A zero indicates either the absence of a microorganism (structural zero), or an insufficient detection level or sequencing depth (sampling zero). Thus, we should remove microorganisms that appear in just a few samples.

Interpretation of association networks is challenging because they are not equivalent to ecological networks. Edges in ecological networks represent observed ecological interactions between different microorganisms like parasitism or competition Xiao et al. (2017). Ecological networks are directed graphs, where the directed edges (arcs) point from a start node (source) to an end node (target). In contrast, association networks are undirected.

Although association networks provide ecological insight, they do not necessarily encode causal relationships or observed ecological interactions. Unless we can verify edges, with experiments or additional information, one should be careful when attributing biological meaning to network properties Röttjers and Faust (2018). In sum, association networks inferred from omics data are a great tool to generate microbial interaction hypotheses that must be investigated experimentally. Interpretation and analysis are problematic if inferred association networks are dense, i.e. have too many edges (so-called "hairball" networks). We can obtain lower density networks when lowering the corrected p-value for inferred edges Weiss et al. (2016), or increasing the cut-off for other criteria such as the association strength, prevalence, or abundance filtering Röttjers and Faust (2018). Another strategy is agglomeration using taxonomic or ecological (functional) groupings Lima-Mendez et al. (2015).

The interpretation challenge addressed in this study is the so-called effect of indirect dependencies caused by environmental factors. For most microbial association networks, an edge indicates one of the following three alternatives:

1. ecological interaction between two microorganisms,
2. similar or contrary dependence (i.e. preference) to environmental factor/s or a third microorganisms,
3. association by chance.

The effect of indirect dependencies occurs when two microorganisms are indirectly associated because both are dependent on an abiotic environmental factor (e.g. have the same requirements of nutrients and temperature) or biotic environmental factor (e.g. have the same prey or predator). Here, indirect association describes the computational effect of indirect dependencies, and observing an association when in fact there is none.

Removing indirect dependencies including environmental effects

To distinguish between direct and indirect interactions, several network construction tools use a probabilistic graphical model Kurtz et al. (2015); Yang et al. (2017), e.g. SPIEC-EASI Kurtz et al. (2019, 2015), miic Verny et al. (2017), or FlashWeave Tackmann et al. (2019). FlashWeave can also integrate metadata to remove indirect associations driven by environmental factors. The tool ARACNE Margolin et al. (2006) aims to eliminate indirect associations by using an information theoretic property (the *Data Processing Inequality*, DPI, in Methods). The extension TimeDelay-ARACNE Zoppoli et al. (2010) tries to extract dependencies between different times. Another approach including time-delay is implemented in the tool MIDER Villaverde et al. (2014), which combines mutual information-based distances and entropy reduction to detect indirect interactions (*Mutual Information*, MI, in Methods). PREMER Villaverde et al. (2018), an evolution of MIDER, allows to include previous knowledge, e.g. known non-existent associations.

There are also several approaches to reduce indirect edges that are applied prior to network construction, e.g. a high prevalence filter that preserves microorganisms present in many samples Pascual-García et al. (2014). However, this will keep generalist microorganisms while removing specialist microorganisms. Another approach to reduce environmental effects is to divide datasets displaying a great environmental heterogeneity into sub datasets of similar environmental conditions Röttjers and Faust (2018). For example, a previous work Mandakovic et al. (2018) constructed two microbial co-occurrence networks representing bacterial soil communities from two different sections of a pH, temperature, and humidity gradient. Another work Lima-Mendez et al. (2015) constructed ocean depth-specific networks to account for environmental differences between the surface layer and the

158 deep chlorophyll maximum layer. In addition to dividing samples, an algorithm that aims to
159 correct for habitat filtering effects Brisson et al. (2019), subtracts, for a given habitat, the
160 mean abundance from each microorganisms within each sample. However, this approach is
161 limited to the identified habitat groups that should have a similar sample size.

162 In contrast to the above, there are methods accounting for indirect dependencies after
163 network construction. For instance, global silencing, Barzel and Barabási (2013) and network
164 deconvolution Feizi et al. (2013) aim to recover true direct associations from observed
165 correlations. Both techniques are sensitive to missing variables Alipanahi and Frey (2013).
166 Another post network construction method, called *Sign Pattern*, SP, uses environmental
167 triplets Lima-Mendez et al. (2015). An environmental triplet contains two microorganisms
168 and one environmental factor, which are associated to each other. SP combines the signs of
169 association scores (positive or negative) to determine if a microbial association should be
170 classified as indirect (SP in Methods). Its major drawback is edge removal where
171 microorganisms with similar environmental preference interact. Along SP and network
172 deconvolution, the *Interaction Information*, II, was applied in Lima-Mendez et al. (2015).
173 Within an environmental triplet, the II method aims to distinguish whether an edge is due
174 entirely to shared environmental preferences (II<0) or whether environmental preferences
175 and true interactions are entangled (II>0). However, II cannot determine which of the three
176 associations in a triplet is indirect (II in Methods). Here, we study several indirect edge
177 detection methods: SP, *Overlap*, OL (developed here), II, DPI, as well as their combination.
178

179 **EnDED is an implementation of four methods and their combination**

180 This article presents EnDED, which implements four approaches, and their
181 combination, to disentangle the type of association represented by an edge in order to remove

the environmentally-driven (indirect) associations from the network. The four methods are: Sign Pattern Lima-Mendez et al. (2015), Overlap (developed here), Interaction Information Ghassami and Kiyavash (2017); Lima-Mendez et al. (2015), and Data Processing Inequality Cover and Thomas (2001); Margolin et al. (2006). SP requires an association score that represents co-occurrence when it is positive, and mutual-exclusion when it is negative. OL requires temporal data with a known start and end of the association to determine whether the microbial association occurs in a time window when both microorganism are associated to the same environmental factor. The II method indicates the existence of one indirect dependency between three components that are associated with each other. The DPI method states that the association with the smallest mutual information is the indirect association. Here, we evaluate each method as well as their combination in an intersection approach on how well they detect environmentally-driven associations. By combining methods in an intersection approach, we have retained more true positives than using each method on its own. A union approach was discarded because it would have retained the smallest number of true interactions. We are able to disentangle and filter environmentally-driven edges from microbial association networks (0.95-0.96 in positive predictive value and 0.35-0.83 in accuracy). EnDED contributed to both, generating more reliable hypotheses on microbial interactions, and facilitating network analysis by removing edges from dense "hairball" networks. EnDED is publicly available Deutschmann et al. (2019).

Results

Simulated data

To evaluate EnDED's performance in removing environmentally-driven

205 associations, we simulated 1,000 abundance time-series datasets with 50 microorganisms and
206 known true interactions between them. We obtained another 1,000 datasets by introducing
207 noise to these time-series with Poisson distributions. We constructed the networks (below
208 called simulated networks) with the tool eLSA Xia et al. (2011, 2012) (see methods). The
209 simulated networks contained on average (computed as the median) 50 nodes and 1,087
210 edges (1,063 for data with noise; hereafter dwn), of which 60 (59 dwn) were true interactions
211 (edges present in the inferred and true network) and 1,026 (1,005 dwn) false associations
212 (edges present in the inferred but absent in the true network). A simple approach to
213 discriminate true interactions (desired) from false associations (undesired) would be to use a
214 threshold for the association strength, which could be suitable if the values for true
215 interactions and false associations are i) following different distributions, and ii) the
216 distributions are mainly non-overlapping. We tested the former requirement with a two-
217 sample Kolmogorov-Smirnov test with the R R Core Team (2019) function ks.test. Using a
218 95% (99%, 99.9%) confidence level, the distributions were significantly different for 358
219 (192, 66) simulated datasets and 355 (173, 68) simulated datasets with noise, which is slightly
220 more than one third of them. This indicates that an association strength cut-off is unsuitable
221 to separate true interactions from false associations. More sophisticated approaches than a
222 simple threshold include the methods implemented in EnDED: SP, OL, II, DPI, as well as
223 their combination.

224 Combining the methods in an intersection approach (hereafter referred to as
225 intersection combination), we classified on average 348 (228 dwn), that is 32% (22% dwn)
226 of the associations, to be environmentally-driven. The number of correctly detected false
227 associations was on average 332 (219 dwn), i.e. 96% of the removed edges. The resulting
228 networks contained on average 737 (828 dwn) edges. When each method was individually

applied more edges were removed: 87% (86% dwn) for SP and OL, 67% (60% dwn) for II, and 44% (32% dwn) for DPI. The fraction of correctly removed edges was on average 95% for each individual method. Individual methods removed more edges from the network than the intersection combination, where all methods must agree. However, a method's performance is not solely determined by the number of removed edges.

To evaluate the removal of environmentally-driven edges, we scored the different approaches based on the true positive rate, TPR, true negative rate, TNR, false positive rate, FPR, positive predicted value, PPV, and accuracy, ACC, (evaluation measurements see Methods). In order to determine these measurements, we first determined true and false positives, as well as true and false negatives. A true positive is a false association in the network that is correctly removed by a method, and a false negative is a false association that incorrectly is not removed. A false positive is a true interaction in the network that is incorrectly removed by a method, and a true negative is a true interaction that correctly is not removed by a method. The ideal method maximizes true positives and true negatives and minimizes false positives and false negatives. The intersection combination under-performed compared to each individual method when considering the TPR, FPR and ACC as shown in Figure 1. However, applying each method individually has the drawback of removing more true interactions. On average there are 60 (59 dwn) true interactions in the simulated networks. The individual methods remove 86% (85% dwn) (SP), 85% (84% dwn) (OL), 60% (51% dwn) (II), and 38% (28% dwn) (DPI). Therefore, although the intersection combination removes fewer edges, it outperforms the others according to the TNR because it eliminates fewer of the true interactions, 25% (16% dwn). We summarized the performance of EnDED in Additional file Table S1. In Figure 1, we plotted the TPR against the FPR for each simulated network and environmentally-driven edge detection methods as well as their

intersection combination. According to the PPV, intersection combination performs best and SP and OL perform worst. SP and OL perform best according to TPR, FPR, and ACC. II performs better than DPI according to TPR, FPR, and ACC. Regarding PPV, the former two methods perform similarly on average. All methods have high PPV values with half of all measured PPV above ≈0.95.

Real data

After testing EnDED's performance on simulated networks, we applied it to a real microbial association network, which was constructed from 10 years of monthly samples from January 2004 to December 2013 at the Blanes Bay Microbial Observatory (BBMO) Gasol et al. (2016). These samples included bacteria and eukaryotes of two size-fractions: picoplankton (0.2-3 μm) and nanoplankton (3-20 μm). We estimated community composition via metabarcoding of the 16S and 18S rRNA gene, and inferred an association network, hereafter referred to as BBMO network (See Methods). The BBMO network contained 844 nodes and 33, 832 edges before applying EnDED. The network contained more positive than negative microbial associations (Figure 2).

By applying EnDED, we found that 28,210 of the network edges (≈83.4% of all edges) were in at least one and in maximum nine environmental triplets (see Additional file Table S2). The set of environmental factors included abiotic factors, e.g. temperature, nutrients as well as cell counts (cells/ml) of heterotrophic prokaryotes, *Synechococcus*, *Cryptomonas*, *Micromonas*, photosynthetic and heterotrophic nanoflagellates. Overall, we detected 66,964 environmental triplets within the BBMO network. Of the 16 considered environmental factors, PO_4^{3-} and salinity were not associated to any microorganism in the network, and turbidity, NH_4^+ and *Cryptomonas* were not found within a triplet. The influence

277 of the remaining 11 environmental factors affecting microbial associations is displayed in
278 Figure 2. Temperature and day length (hours of light) are the top two environmental factors
279 affecting microbial associations followed by *Micromonas*, photosynthetic and heterotrophic
280 nanoflagellates.

281 The intersection combination of the four methods removed 4,806 (≈14.2%)
282 associations from the BBMO network(see Figure 2). When analysing these indirect edges,
283 we discovered that over 51.6% are between bacteria, 38.5% between bacteria and eukaryotes,
284 and 10% between eukaryotes. Figure 2shows the number of edges and fraction of
285 environmentally-driven edges between the two domains. Considering size fractions, these
286 environmentally-driven edges correspond 19.5% to picoplankton, 44.2% to nanoplankton,
287 and 36.3% to edges between size fractions. We have summarized the number of associations
288 in Additional file Table S3. Compared to the intersection combination approach, each method
289 individually removed more edges: 83.4% (SP and OL removing all microbial edges present
290 in a triplet), 38.3% (II), and 31.4% (DPI); that is, between a factor of x2 and x6 larger
291 removal.

292 We also determined for each association the Jaccard index, also known as Jaccard
293 similarity coefficient (See Methods), which indicates how often two microorganisms appear
294 together in the dataset. Although we are aware of time-delayed interactions and eLSA Xia et
295 al. (2011, 2012) could account for them, we did not take them into consideration for our
296 BBMO dataset, as we consider our sampling interval is too large (1 month) for inferring
297 associations that may have a solid ecological basis. Thus, in our study, we focused on
298 contemporary interactions between co-occurring microbes. We found that only 29.8% of
299 indirect associations have a Jaccard index above 0.5, i.e. microorganisms appeared together
300 over 50% of the time, compared to 63.3% of the associations that appeared in at least one

triplet but were not removed from the BBMO network (see Table 1). We assume that two microbes that appear together < 50% of the time are less likely to have true contemporary ecological interactions and the corresponding association is more likely to be false. The fact that over 70% of environmentally-driven associations have a Jaccard index equal or below 0.5 strengthens the decision of their removal. When considering the sign of an association, we found that only 3% of negative associations obtain a Jaccard index over 0.5, compared to 68% of the positive associations.

In the BBMO network, the intersection combination approach removed roughly the same number of negative and positive edges, 2,263 and 2,543, respectively (see Figure 2). The pre-EnDED network contained 81.9% positive and only 18.1% negative edges, so the method removed 41.5% of the negative and only 8.2% of the positive edges. If we randomly removed 4,806 edges, we would expect 18.1% to be negative (i.e. 870) and 81.9% of them to be positive (i.e. 3,936). If we restrict these calculations to the 28,210 microbial associations that were found in at least one environmental triplet, with 22,492 of them being positive and 5,718 being negative, we would expect to remove 20.3% (i.e. 976) of negative and 79.7% (i.e. 3, 830) of positive edges. The probability of randomly removing an equal number of positive and negative associations is nearly zero, since it follows a multivariate hypergeometric distribution:

$$P(k_{neg}, k_{pos}) = \frac{\binom{N_{neg}}{k_{neg}} \cdot \binom{N_{pos}}{k_{pos}}}{\binom{N}{n}}, \quad \text{Eq. (1)}$$

where N_{pos} and N_{neg} are the number of positive and negative associations in the network, respectively, k_{pos} is the number of removed positive and k_{neg} the removed negative associations from the network, N is the number of associations in the network, and n is the

number of removed associations from the network. The intersection combination removing an equal number of positive and negative edges, indicates that the removal was not random and that the removal is biased towards negative associations.

In order to evaluate the performance of EnDED on the BBMO network, we considered interactions described in literature and collected in the Protist Interaction Database (PIDA) Bjorbækmo et al. (2019). In order to use these known interactions, we taxonomically classified the microbes in the BBMO network (see methods). Studies typically compare the associations of a network to those reported in the literature at the genus level Lima-Mendez et al. (2015). The ambiguity in taxonomic classification and the large number of edges challenge this comparison. Thus, we implemented a function to compare strings and match the taxonomic classification of a microorganism in the BBMO network to those in the scientific literature (PIDA). We found that only 31 (< 0.09%) associations were supported by interactions described in the literature (see Table 2). That is, 99.91% of associations in the BBMO network (before applying EnDED) could not be used to evaluate EnDED's performance. These 31 associations describe nine unique interactions between 9 microorganisms, and 18 edges were in an environmental triplet to which each method as well as their combination were applied (see summary in Table 2). Ideally none of these described associations should be removed by EnDED. Yet, the intersection combination removed 8 associations: one between a diatom (*Thalassiosira*) and a dinoflagellate (*Heterocapsa*), and seven associations between a diatom (Thalassiosira) and an unknown *Flavobacteriia*. In contrast and even worse, SP and OL removed all 18 edges, II 11 and DPI 14 edges. The additionally removed edges by individual methods are associations between a diatom (*Thalassiosira*) and an unknown *Flavobacteriia*. Considering only the genus level, there were 179 unique microbes (level genus) in the BBMO network, and 700 in PIDA, combined

346 there were 843 microorganisms, and 36 microorganisms are in both. Thus, 20.1% of
347 microorganisms found in the BBMO network were also in PIDA, and 5.1% of
348 microorganisms found in PIDA were also found in the BBMO network. Regarding
349 interactions, and not considering prokaryote-prokaryote associations, there were 1,266
350 unique interactions in PIDA, but 3,422 in the BBMO network (considering only genus). Only
351 5 unique interactions are in both, i.e. 0.39% of PIDA interactions were found in the BBMO
352 network, and 0.15% of BBMO associations were found in PIDA.
353

354 Discussion

355 Using EnDED to disentangle environmental effects in microbial association networks

356 EnDED

357 EnDED makes several indirect-edge removal techniques accessible to microbial
358 ecologists and does not require previous programming experience. These techniques can be
359 used individually or combined. In addition, this work systematically evaluates the different
360 techniques and their combination to remove indirect edges from microbial association
361 networks. Here, we tested only the union and intersection combination of all four methods,
362 but other combination strategies are possible to obtain with EnDED. EnDED requires the
363 data of the environmental factors in order to predict if an association is environmentally-
364 driven, but we understand that it may be impossible to consider all environmental factors Lv
365 et al. (2019). Despite this limitation, EnDED can perform well if the major environmental
366 factors, such as, e.g. temperature and nutrient concentrations for marine microbes, are
367 provided. Moreover, knowledge of microbial interactions in nature is rather limited and
368 therefore determining the performance of EnDED for real networks is challenging and carries

some degree of uncertainty. Thus, the analysis of EnDED's results without previous validation should be interpreted with care. Here, we first applied EnDED on simulated networks in order to measure the performance of individual methods as well as their intersection combination. Then, we applied EnDED to an association network constructed from observational data. We used monthly data from the BBMO marine time-series Gasol et al. (2016), which provided 10 years of microbial abundance data and 16 environmental factors.

For the simulated networks (and simulated networks with noise), we found that each method individually removed on average a moderate to high number of edges: 44% (32% dwn) DPI, 67% (60% dwn) II, and 87% (86% dwn) SP and OL. The intersection combination of the four methods removed on average 32% (22% dwn) of the edges, while also keeping more true interactions. To understand the impact of the environment, Röttjers and Faust simulated an increasing environmental influence and observed a decrease in retrieving true interactions from inferred associations, i.e. a decrease in precision Röttjers and Faust (2018). They also compared several microbial correlation network construction methods for cross-sectional data, including CoNet Faust et al. (2012), SparCC Friedman and Alm (2012), SPIEC-EASI Kurtz et al. (2015), and Spearman correlations—all exhibited a reduced precision on data with simulated environmental effects. For our simulation networks, we observed a slight increase in precision when removing environmentally-driven associations. In summary, intersection combination removed the smallest edge number, outperforming individual methods in terms of number of remaining true interactions, TNR, and PPV. Regarding TPR and ACC, all methods performed better on simulated networks without noise. Fewer edges have been removed from these networks along with fewer true interactions. All approaches performed similarly according to the PPV, which could be a result of the removal

of fewer edges and also fewer true interactions.

In our BBMO dataset, the intersection combination removed 14.2% of the edges—41.5% of the negative and only 8.2% of the positive edges. We argue that several negative associations are probably due to different environmental preference (different niches) of microbes. The Jaccard index representing a level of microbial co-occurrence, scored equal or below 50% for 97% of the negative associations. These may partially represent microbes adapted to different seasons. Previous work on the eukaryotic pico- and nano-plankton at the BBMO, using the same basal 10-year dataset used here, indicated a strong seasonality at the community level Giner et al. (2019).

Comparisons of indirect edge detection on other datasets

In our BBMO network we found 28,210 (83.4%) microbial edges that were within at least one environmental triplet. Thus, the fraction was 2.6 times higher than what was found for an association network called "global plankton interactome" containing 29,912 (32.3%) edges, associated to microbes as well as small metazoans, that were within at least one environmental triplet Lima-Mendez et al. (2015). In the previous study 29,900 (≈100% of triplets and 32% of all edges) were attributed to environmental factors by SP. II indicated 11,043 environmentally-driven edges (≈37% of triplets and 12% of all edges) with p-value below 0.05 in a permutation test with 500 iterations. Network deconvolution suggested 22,439 environmentally-driven edges (≈75% of triplets and 24% of all edges). These three methods agreed for 8,209 edges (≈27% of triplets and 8.9% of all edges). In comparison, we found more environmentally-driven associations for the BBMO network (14.2% of all edges) based on a decade of temporal data from one location and one depth including two size

fractions than what was detected for the global ocean interactome covering two depths, 68 stations around the world and various size fractions Lima-Mendez et al. (2015). These differences suggest that microbial temporal turnover may induce more indirect edges than spatial turnover. Thus, the effects of indirect dependencies may depend on dataset type (e.g. temporal vs. spatial), and this should be further investigated.

Using II for the BBMO network, we removed 38.3% of the edges (45.9% when considering only triplets), which would indicate a moderate number of associations explained by an environmental factor. The DPI also identified a moderate number (31.4%, 37.6% when considering only triplets) of environmentally-driven associations in the BBMO network, whereas SP or OL identified a ubiquitous number of environmentally-driven edges (83.4%, 100% when considering only triplets). This indicates that SP and OL are strict and should be used in combination with other methods in an intersection approach. In another study, the tool FlashWeave Tackmann et al. (2019) predicted direct microbial interactions in the human microbiome using the Human Microbiome Project (HMP) dataset, including heterogeneous microbial abundance data of 68, 818 samples The Human Microbiome Project Consortium: Huttenhower et al. (2012). The inferred networks (with and without metadata) were sparser than our networks. The network with metadata contained 10.7% fewer associations compared to the network without metadata, which suggests a minor number of environmentally-driven edges in the tested dataset, slighlty less than in our results from BBMO. Considering the previously mentioned comparison between the number of indirect edges detected in BBMO (higher) vs. the ocean interactome (lower), it remains to be tested whether FlashWeave would detect more indirect edges in temporal than in spatial or in environmental vs. host-associated datasets.

441 **Factors causing indirect microbial associations**

442 From the simulated networks, we found that using the intersection combination
443 instead of each method individually, we maintained more true interactions at the cost of more
444 false associations in the network—more when considering simulated networks including
445 noise. Comparing our simulated network against the BBMO network, the intersection
446 combination classified a higher number of edges as environmentally-driven in the simulated
447 networks 32% (22% dwn) than in the BBMO network (14.2%). For the simulated data, we
448 previously knew the environmental factor influencing pairwise microbial associations. For
449 the BBMO data, we used 16 available environmental factors, but not all factors that could
450 affect microbial dynamics. Even though the most important factors influencing microbial
451 seasonal dynamics at BBMO were considered Giner et al. (2019), there are several other
452 environmental factors that were not measured and that could generate indirect edges. The
453 indirect edges associated to these factors were not detected in our analyses. Similarly, indirect
454 edges associated to biotic interactions (e.g. two bacteria sharing a positive edge as they are
455 symbionts in the same protists) were not considered. Future sampling for microbial
456 interaction research should expand metadata collection in order to detect more abiotic and
457 biotic factors that could generate indirect edges. While we identified temperature and day
458 length (hours of light) to be the top two environmental factors affecting microbial
459 associations in the BBMO network, followed by *Micromonas*, photosynthetic and
460 heterotrophic nanoflagellates, the frequent environmental factors in the global plankton
461 interactome Lima-Mendez et al. (2015) were phosphate concentration, and temperature,
462 followed by nitrite concentration, and mixed-layer depth. Although we considered PO_4^{3-} it
463 was not associated to any microorganism in the network along with salinity, which could be
464 explained by the fact that these environmental factors were more homogenous in our BBMO

dataset. For instance, the standard deviation in BBMO dataset was < 1 for PO_4^{3-} and salinity in contrast to Tara samples Lima-Mendez et al. (2015), where it was about 20-30 when considering all samples. During the Malaspina 2010 Circumnavigation Expedition, the concentrations of trace metals were determined for 110 surface water samples Pinedo-González et al. (2015). The previous study indicates relationships between primary productivity and trace nutrients, more specifically for the Indian Ocean Cd, the Atlantic Ocean Co, Fe, Cd, Cu, V and Mo, and the Pacific Ocean Fe, Cd, and V. Thus, trace metals may play an important role in regulating oceanic primary productivity.

Limitations of EnDED

It may be promising to update EnDED—currently using environmental triplets—to allow any closed triplet, (i.e. three microorganisms being all connected), as done with gene triplets Margolin et al. (2006). A recent update of the network construction tool eLSA Xia et al. (2011, 2012) permits to examine how a factor, such as a microorganism or environmental variable, mediates the association of two other factors Ai et al. (2019), which allows the study of interactions between three factors. Furthermore, triplets limit the study to first-order indirect dependencies, neglecting higher-order indirect dependencies. Such limitation was solved for the DPI method by examining associations in quadruplets, quintuplets, and sextuplets Jang et al. (2013). Implementing higher-order DPI and adjusting the other three methods to account for higher-order indirect dependencies may be promising but one need to be aware that incorporating higher-order dependencies will also increase the risk of over-fitting. Further, all relevant environmental factors could be incorporated into the calculation of II, which would combine several environmental triplets. However, we reason that such adjustments would require a larger sample size. Both II and DPI methods calculate MI that

measures the dependence between two random variables. EnDED is limited by including one function to estimate the MI. A comparison of four different estimates of MI revealed that obtaining the true value of MI is not straightforward, and minor variations of assumptions yield different estimates Fernandes and Gloor (2010). Lastly, the conditional mutual information, CMI, which quantifies nonlinear direct relationships among variables, can be underestimated if variables have tight associations in a network Zhao et al. (2016). The so-called part mutual information, PMI, measurement can help overcome CMI's underestimations. Although using PMI instead of CMI looks promising, calculating PMI is computationally more demanding Zhao et al. (2016).

Future Perspective

In this study, we have shown that EnDED with an intersection combination approach provides less dense networks, but still with many potential interactions. Specific associations may be validated with experiments or microscopy Krabberød et al. (2017); Lima-Mendez et al. (2015). However, we suggest to first further reduce the set of potential interaction hypotheses. To improve the selection of interaction hypotheses, we propose to score associations based on re-occurrence: in time, as done with microbial abundance seasonality Giner et al. (2019), or space, where an association appears in different networks based on different datasets, or different regions of the world. In a previous study using 313 samples, including seven size-fractions, four domains (Bacteria, Archaea, Eukarya, and viruses), and two depths from 68 stations across eight oceanic provinces, 14% of the 81,590 predicted biotic interactions were identified as local Lima-Mendez et al. (2015). Thus, re-occurrent associations suggest a higher likelihood that the association represents a true ecological interaction, reducing the number of interaction hypotheses to the strongest ones. Another

strategy to shortlist interaction hypotheses is to incorporate additional data into the network and use a multi-layer network approach. Such data could be environmental preferences such as temperature or salinity optima, size of cells, presence of chloroplasts, or more sophisticated data obtained from High-Throughput Cultivation Faust (2019), microbial community transcriptomes that reveal microbes and metabolic pathways McCarren et al. (2010), or interactions inferred from Single-Cell genome data Krabberød et al. (2017); Yoon et al. (2011).

Conclusion

In this paper, we present EnDED, an analysis tool to reduce the number of environmentally induced indirect edges in inferred microbial networks. We applied EnDED to networks based on time-series of simulated data and observed marine microbial abundance data. Our simulated networks indicate that false associations, driven by environmental variables instead of true interactions, are ubiquitous in inferred association networks. However, EnDED's intersection combination classified a minority of associations as environmentally-driven in a real (BBMO) network. Depending on the single method used, we classified a moderate to high number of associations as environmentally-driven in the same network. Nevertheless, associations driven by environmental factors must be determined and quantified to generate more accurate insights regarding true microbial interactions. EnDED provides a step forward in this direction.

Methods

Simulated dataset: time series based on an adjusted generalized Lotka-Volterra model

We simulated a time series using an adjusted version of the standard *generalized Lotka-Volterra model*, gLV. The standard gLV generates simulated data for evaluation Bashan et al. (2016); Berry and Widder (2014). The gLV can describe the dynamics of microbial communities, by including a first order approach of the microbial interactions. The model's simplicity arises from the assumption of linear interactions, which facilitates implementation and allows fast numerical simulations. The gLV has, however, several limitations Gonze et al. (2018). For example, gLV neglects higher-order interactions and the additivity of interaction strengths is a weakness because they may be combined in different ways. Also, interactions are often assumed to be constant parameters, but a reducing level of a nutrient may weaken cross-feeding relationships. Moreover, gLV omits the influence of environmental factors, which, for example, can induce oscillations in natural communities Beninca` et al. (2011). Using a model that accounts for nutrients Kettle et al. (2018) is more realistic but also more complex. More elaborate mechanistic models of microbial dynamics than gLV solve explicitly the global cycling of nutrients and are coupled to the oceanic circulation (see Vallina et al. (2019) for a review), but the added complexity can hamper understanding about the ecological interactions among microorganisms when compared to a simpler gLV approach. Thus, we chose to use a simpler extension of the gLV to account for the influence of environmental factors Dam et al. (2016); Stein et al. (2013). In order to allow the growth rates to vary when the environmental variables change, environmental variables can be incorporated directly into the gLV Dam et al. (2016); Röttjers and Faust (2018). We simulated a time series using the Klemm-Eguíluz algorithm Klemm and Eguíluz (2002), and an adjusted gLV. We adjusted the model by defining microorganisms growth rates as a function dependent on one seasonal abiotic environmental factor, and added an abiotic environmental factor in the interaction matrix. We then used the time series generated by the

gLV to obtain temporal microbial abundance data. With this simulated data, we inferred a network that contained environmentally-driven associations, needed to evaluate the performance of EnDED. We repeated this procedure 1,000 times to obtain a large set of simulated networks, and then used the determined abundance tables and Poisson distribution to obtain another 1,000 simulated networks including noise. The addition of noise was done by randomly drawing an abundance from the Poisson distribution with λ equaling the original abundance of a specific microorganisms to a specific time.

Adjusting the gLV

To evaluate EnDED, we simulated a time series of microbial abundances with a gLV including true pairwise interactions between 50 microorganisms and adjusted it by incorporating two environmental factors:

$$\frac{dy(t)}{dt} = y(t)[b + Ay(t)]\,, \qquad \text{Eq. (2)}$$

where t is time, $dy(t)/dt$ is the rate of change of microbial abundances as a column vector, $y(t)$ is the vector of microbial abundance at time t, b is the growth rate vector determined through microorganisms specific growth rate functions that depend on an environmental factor (see equation (4)), and A is the interaction matrix.

Interaction matrix

In the interaction matrix A, each coefficient a_{ji} provides the linear effect that a change in the abundance of microorganism i has on the growth of microorganism j Novak et al. (2016). We simulated the interaction coefficients a_{ji} with the Klemm-Eguíluz algorithm Klemm and Eguíluz (2002), which generates a modular and scale-free matrix. We also set

582 the interaction probability to 0.01, the percentage of positive coefficients to 30%, and
583 diagonal coefficients to zero. Negative diagonal coefficients a_{ji} (i.e. the interaction of a
584 microorganism with itself) can represent intra-specific competition and provides the carrying
585 capacity for each microorganism, preventing its explosive growth Haydon (1994). We set the
586 diagonal coefficients $a_{ii} = -0.5$ to avoid excessive microbial abundances in the simulations.
587
588 Two abiotic environmental factors
589 We adjusted the gLV by including two environmental factors. For simplicity, we
590 assume no feedback between the microorganisms and the environmental factors. That is, the
591 environmental factors affect the growth of the microorganisms but not vice-versa. The first
592 environmental factor affects the specific growth rate of each microorganism by interacting
593 with two of their traits: optimal environmental value for growth and tolerance range of
594 environmental values. We simulated the environmental factor using a periodic sinusoidal
595 function (see equation (3)), rounded to 3 digits:

$$\epsilon(t) \triangleq round(\sin(\omega \cdot t)\,, \text{digits} = 3)\,, \qquad \text{Eq. (3)}$$

596 where t is the time axis (months), $\omega = (-2\pi/T)$ is the signal frequency (radians) and $T =$
597 12 is the signal periodicity (months); resulting in a signal phase shift of $T/4$ (months). While
598 the first environmental factor is considered to be "external" to the microbial community, the
599 second environmental factor is considered to be "internal", and therefore it is included in the
600 interaction matrix. The interaction coefficients between the microorganisms and the second
601 environmental factor were generated by splitting the microorganisms into two groups: the
602 second abiotic environmental factor influenced positively one half and negatively the other
603 half of the microorganisms. We obtained the interaction coefficients from two uniform

distributions defined to range between [-0.8, -0.2] and [0.2, 0.8] respectively. As the microorganisms did not influence the abiotic factor, the corresponding interaction coefficients were set to zero.

Species growth rate

The external seasonal abiotic environmental variable affects the growth rate, g, of each microorganism. This dependency is given by:

$$g(t) \triangleq g_{max}^2 \exp\left(-\frac{1}{2}\frac{\left(\epsilon_{opt} - \epsilon(t)\right)^2}{\sigma^2}\right), \quad \text{Eq. (4)}$$

where $E(t)$ is the environmental parameter that affects the microorganisms growth rate $g(t)$ at time t, g_{max} is the microorganism' specific maximum growth rate that determines the amplitude of the growth-rate curve, ϵ_{opt} is the microorganism' specific optimal environmental value that determines the peak of the growth-rate curve, and σ is the microorganism' specific ecological tolerance (niche width) determining the environmental range in which the microorganism grows, which determines the length (niche spread) of the growth-rate curve. We obtained the two constant parameters g_{max}, and σ for each microorganism from a uniform distribution ranging between 0.3 and 1 to assure positive values. The values ϵ_{opt} were drawn from a uniform distribution ranging between the minimal and maximal value of the seasonal environmental factor. We defined the internal abiotic environmental factor, which is included in the interaction matrix, through the same function with $g_{max} = 0.8$, $\epsilon_{opt} = 0.5$, and $\sigma = 0.5$. Since the growth rates depend on the environmental factor, they vary seasonally. Different microorganisms will grow better or worse at different times of the year following their environmental niches. This will lead to

an asynchrony of their growth rate responses to the environment that will translate into an asynchrony of their abundances in time.

Initial abundances

To obtain the microbial abundances in time with the adjusted gLV, we simulated the initial microbial abundances with a stick-breaking process such that abundances add up to 1, using the function bstick Jackson (1993); Legendre and Legendre (2012), and the package vegan Oksanen et al. (2019). We generated uneven initial microbial abundances without introducing zeros and set the initial value for the internal abiotic environmental factor included in the interaction matrix to 0.001.

Species abundances in time

Once we have set the initial conditions, we simulated microbial abundances over time by solving the equations given in the adjusted gLV (see equation (2)). Start time was 0, end time 49.5, and sample resolution 0.5 resulting in 100 samples. We used the solver function lsoda Soetaert et al. (2010). The simulated abundances in time were used to construct an association network, which is referred to as the simulated network.

Real dataset: Blanes Bay Microbial Observatory time series

Microbial abundances

Surface water (< 1m depth) was sampled monthly from January 2004 to December 2013 at the BBMO in the North-Western Mediterranean Sea Gasol et al. (2016). About 6L of seawater were filtered and separated into two size fractions, picoplankton (0.2-3 µm) and nanoplankton (3-20 µm), as described in Giner et al. (2019). Community DNA was extracted,

649 and the 18S ribosomal RNA-gene (V4 region) was amplified Giner et al. (2019). The 16S
650 ribosomal RNA-gene (V4 region) was also amplified from the same DNA extracts using the
651 primers Bakt 341F Herlemann et al. (2011) and 806R Apprill et al. (2015). DADA2 v1.10.1
652 Callahan et al. (2016) was used for read quality control, trimming and inference of
653 Operational Taxonomic Units (OTUs) as Amplicon Sequence Variants (ASVs). In both
654 microbial eukaryotes and prokaryotes, the maximum number of expected errors (MaxEE)
655 was set to 2 and 4 for the forward and reverse reads, respectively. OTU tables were generated
656 for both the 16S and 18S rRNA genes. Before network construction all samples were
657 individually subsampled using the function rrarefy, in R package vegan Oksanen et al.
658 (2019), to the size of the sample with the lowest sequencing depth (4,907 reads). Due to
659 suboptimal sequencing of the amplicons from some months, we did not use nanoplankton
660 samples from the period May 2011 to July 2012 (27 samples) as well as 4 additional samples.
661 OTU abundances for the missing samples were estimated using seasonally aware missing
662 value imputation by weighted moving average for time series as implemented in the R
663 package imputeTS Moritz and Gatscha (2017).
664
665 Taxonomic classification
666 The taxonomic classification of each OTU was inferred with the naïve Bayesian
667 classifier method Wang et al. (2007) together with the SILVA version 132 Quast et al. (2012)
668 database as implemented in DADA2 Callahan et al. (2016). In addition eukaryotic
669 microorganisms were BLASTed Altschul et al. (1990) against the Protist Ribosomal
670 Reference database [PR2, version 4.10.0; Guillou et al. (2012)]. If the taxonomic assignment
671 for eukaryotes disagreed between SILVA and PR2, we used the PR2 classification. We
672 removed microorganisms identified as either Metazoa, or Streptophyta, plastids and

673 mitochondria. In addition, we removed Archeas since primers were not optimal for
674 recovering this domain.
675
676 Environmental factors
677 We measured environmental factors that may affect the ecosystem's dynamics. We
678 considered a total of 16 contextual abiotic and biotic variables: temperature (C_4°), turbidity
679 (Secchi depth m), salinity, total cholorophyll (µg/l), inorganic nutrients— PO_4^{3-} (µM), NH_4^{+}
680 (µM), NO_2^{-} (µM), NO_3^{-} (µM), and SiO_2 (µM)—heterotrophic prokaryotes (cells/ml),
681 *Synechococcus* (cells/ml), *Cryptomonas* (cells/ml), *Micromonas* (cells/ml), photosynthetic
682 nanoflagellates (cells/ml), heterotrophic nanoflagellates (cells/ml), day length (hours of light)
683 Giner et al. (2019). Water temperature and salinity were sampled in situ with a CTD
684 (Conductivity, Temperature, and Depth) measuring device. Inorganic nutrients were
685 measured with an Alliance Evolution II autoanalyzer Grasshoff et al. (2009). See Gasol et al.
686 (2016) for specific details on how other variables were measured.
687
688 **Network construction**
689 For both the simulated and BBMO dataset, we constructed association networks from
690 microbial abundance tables and environmental parameters using eLSA Xia et al. (2011,
691 2012). We included default normalization and a z-score transformation using median and
692 median absolute deviation. We estimated the p-value with a mixed model that performs a
693 random permutation test of a co-occurrence if the theoretical p-values for the comparison are
694 below 0.05; the number of iterations was 2,000, and we considered no delay. For the BBMO
695 dataset, the Bonferroni false discovery rate, q, was calculated for all edges based on the p-
696 values using the R function p.adjust R Core Team (2019). For the network inference, we used

the significance threshold for the p and q value of 0.001 as suggested in other works Weiss et al. (2016).

Intersection combination of EnDED—Environmentally-Driven Edge Detection methods

For the intersection combination approach implemented in EnDED, to determine if a microbial association is environmentally-driven or not, all four individual methods (i.e. Sp, OL, II, and DPI, described below) must converge to the same solution. We applied the methods to find environmentally-driven associations of microorganisms that were within an environmental triplet, as already used in Lima-Mendez et al. (2015). An environmental triplet is a special case of a closed triplet where one of the nodes corresponds to an environmental component and the other two nodes correspond to microorganisms. We define the closed triplet, where there is an edge between each pair of three nodes, as $T = \{v, w, f\}$ where v and w are two microorganisms, and f is an environmental component (see Figure 3). For an environmental triplet, if all methods classify the microbial edge as environmentally-driven, the edge is removed from the network. If a microbial association is within several environmental triplets, at least one of them must indicate the association as environmentally-driven in order to remove the edge from the network. In sum, the intersection combination retains an association in the network if no triplet classifies the association as environmentally-driven.

Sign Pattern

The SP method Lima-Mendez et al. (2015) filters environmentally-driven edges from a network in which a positive association score indicates co-occurrence, and a negative

721 association score indicates mutual exclusion. Let s_{vw} be the sign of the association score of
722 the association between v and w (i.e. $s_{vw} = +$ or $s_{vw} = -$). A closed triplet T has eight SP
723 combinations that group into two sets (see Figure 3). If the product of the three association
724 scores is positive, then the SP suggests that the edge between the two microorganisms is
725 environmentally-driven. Otherwise, if the product of the three association scores is negative,
726 SP does not suggest that the association is environmentally-driven.
727
728 Overlap
729 We have developed the OL method to support the SP for temporal data: a microbial
730 edge should be disregarded as environmentally-driven when the associations are misaligned
731 in time. Thus, OL requires the time when the association begins as well as how long the
732 associations lasts, i.e. duration or length of association in time, both determined by the
733 network construction tool eLSA Xia et al. (2011, 2012). Given an association between v
734 and w, let b_{vw}^{v} be the beginning of the association for v, b_{vw}^{w} the beginning of the
735 association for w, and d_{vw} be the duration of the association between v and w. Although
736 not used in the BBMO network, OL can consider time-delays by assuming that the
737 beginning of the association is the minimum of the two beginnings, $b_{vw} = \min(b_{vw}^{v}, b_{vw}^{w})$,
738 and the end of the association is the maximum, $e_{vw} = \max(b_{vw}^{v} + d_{vw}, b_{vw}^{w} + d_{vw})$. We
739 indicate two microorganisms with v and w, and the factor by f. The OL method calculates
740 the overlap O of the microbial association with the two microorganism-environment
741 associations through equation (5). As depicted in Figure 3, if O>60%, the microbial
742 association is considered environmentally-driven.

$$O = 100\,\frac{min\left(e_{vw}, e_{vf}, e_{wf}\right) - \max\left(b_{vw}, b_{vf}, b_{wf}\right)}{e_{vw} - b_{vw}} \quad \text{Eq. (5)}$$

743 Mutual Information and Conditional Mutual Information

744 The method II employs two measurements: MI and CMI. The former is also used by
745 DPI. Thus, before describing the methods, we first describe the two measurements. MI is a
746 measure of the degree of statistical dependency between two variables Margolin et al. (2006).
747 We first consider $\boldsymbol{v} = v_1, \ldots, v_n$, $\boldsymbol{w} = w_1, \ldots, w_n$, and $\boldsymbol{f} = f_1, \ldots, f_n$ as discrete random
748 variables. The marginal probability of each discrete state (value) of the variable is denoted
749 by $p(v_i) = P(\boldsymbol{v} = v_i)$, the joint probability by $p(v_i, w_j)$, and $p(v_i, w_j, f_k)$, and the
750 conditional probability by $p(v_i|f_k)$, and $p(v_i, w_j|f_k)$. To obtain MI, we calculate the entropy
751 of $\boldsymbol{v}$ as

$$S(\boldsymbol{v}) = -\sum_{i=1}^{n} p(v_i)\log\left(p(v_i)\right), \quad \text{Eq. (6)}$$

752 and the joint entropy of $\boldsymbol{v}$ and $\boldsymbol{w}$ as

$$S(\boldsymbol{v}, \boldsymbol{w}) = -\sum_{i=1,j=1}^{n} p\left(v_i, w_j\right)\log\left(p\left(v_i, w_j\right)\right), \quad \text{Eq. (7)}$$

753 using the natural logarithm. The MI of $\boldsymbol{v}$ and $\boldsymbol{w}$ is defined through the sum of their entropies
754 subtracted by their joint entropy:

$$\mathrm{MI}(\boldsymbol{v}; \boldsymbol{w}) = S(\boldsymbol{v}) + S(\boldsymbol{w}) - S(\boldsymbol{v}, \boldsymbol{w}) \quad \text{Eq. (8)}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} p\left(v_i, w_j\right)\log\left(\frac{p(v_i, w_i)}{p(v_i)p(w_j)}\right), \quad \text{Eq. (9)}$$

755 with marginal probabilities $p(v_i) = \sum_{j=1}^{n} p(v_i, w_j)$, and $p(w_j) = \sum_{i=1}^{n} p(v_i, w_j)$.

756 The measurement CMI is the expected value of the MI of two random variables given

757 a third random variable. It is defined as

$$\mathrm{CMI}(\boldsymbol{v};\boldsymbol{w}|\boldsymbol{f}) = S(\boldsymbol{v},\boldsymbol{f}) + S(\boldsymbol{w},\boldsymbol{f}) - S(\boldsymbol{v},\boldsymbol{w},\boldsymbol{f}) - S(\boldsymbol{f}) \quad \text{Eq. (10)}$$

$$= \sum_{k=1}^{n} p(f_k) \sum_{i=1}^{n}\sum_{j=1}^{n} p(v_i, w_j|f_k) \log\left(\frac{p(v_i, w_i|f_k)}{p(v_i|f_k)p(w_j|f_k)}\right) \quad \text{Eq. (11)}$$

$$= \sum_{k=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} p(v_i, w_j, f_k) \log\left(\frac{p(f_k)p(v_i, w_i, f_k)}{p(v_i, f_k)p(w_j, f_k)}\right).$$

758

759 Interaction Information

760 The II is calculated with microbial abundance and environmental data. In this study,
761 as in Lima-Mendez et al. (2015), II is computed as the difference of the CMI and MI:

$$\mathrm{II} = \mathrm{CMI} - \mathrm{MI}\,. \quad \text{Eq. (12)}$$

762 In other works Ghassami and Kiyavash (2017), the II is defined with a different sign
763 convention: II = MI-CMI. In our study, if II is positive, the method suggests that the microbial
764 association is not environmentally-driven. If II is negative, there is an environmentally-
765 driven association within the closed triplet. However, this method cannot detect which of the
766 three associations is indirect. In other works Lima-Mendez et al. (2015), the microbial
767 association is assumed to be environmentally-driven if II is negative, but here we suggest to
768 combine it with DPI (see below).

769

770 Significance of Interaction Information

771 We determined the significance of II following a strategy from Ref. North et al.
772 (2002); Veech (2012). We used a null model without prior assumptions on the distribution

773 and computed the p-value by randomizing the environmental vector f. Since the MI is
774 independent of the environmental factor and therefore remains constant, the significance of
775 the II is the same as the CMI. Thus, we determined the significance of CMI with 1,000
776 permutations: we randomized the environmental vector f and recalculated the CMI 1,000×,
777 obtaining a CMI_i with $i \in \{1,\ldots,1{,}000\}$. Afterwards, we quantified with c how many
778 random CMI_i were at least as small as the original CMI_i: $c = |i\colon \mathrm{CMI}_i \leq \mathrm{CMI}_{original}, i \in$
779 $\{1,\ldots,1{,}000\}|$. We calculate the p-value as

$$p = \frac{c+1}{1{,}000+1}\,. \qquad \text{Eq. (13)}$$

780

781 Data Processing Inequality

782 As mentioned above, the II method can detect if an indirect association exists within
783 a triplet but cannot determine which of the three associations is indirect. Thus, we added DPI
784 to EnDED. DPI states that if two components v and w interact only through a third
785 component f (i.e. in a network v and w are connected through a path containing f and there
786 is no alternative path between v and w), then the MI of v and w, MI(**v**; **w**) is smaller than
787 MI(**v**; **f**) and MI(**w**; **f**) Cover and Thomas (2001):

$$\mathrm{MI}(\boldsymbol{v};\boldsymbol{w}) \leq \min\{\mathrm{MI}(\boldsymbol{v};\boldsymbol{f}),\mathrm{MI}(\boldsymbol{w};\boldsymbol{f})\}\,. \qquad \text{Eq. (14)}$$

788 While DPI has been used in previous works on gene triplets Margolin et al. (2006), we used
789 the DPI method for environmental triplets. We compared the MI between the two
790 microorganisms with the MI between a microorganism and the environmental factor. If the
791 MI between the microorganisms is the smallest, then the method suggests that the edge is
792 environmentally-driven. This method complements the II method.

793

Equal Width Discretization

To compute the MI, CMI, and subsequently II, we discretized the abundance data and environmental parameters. EnDED uses the equal width discretization algorithm, which creates equal sized ranges (also called bins or buckets) for an abundance vector $\boldsymbol{v} = (v_1, \ldots, v_n)$ between the lowest value (v_{min}) and highest value (v_{max}). It is a procedure implemented in other works Meyer et al. (2008). Given vector $\boldsymbol{v}$ of length n (that is sample size) and number of bins $|B| = \lfloor \sqrt{n} \rfloor$, the discretized value v_d of $\boldsymbol{v}$ is:

$$v_d = \left\lceil \frac{(v - v_{min}) \cdot |B|}{v_{max}} \right\rceil . \qquad \text{Eq. (15)}$$

This equation assumes positive values. However, if $\boldsymbol{v}$ contains negative values, $v_{min} < 0$, we adjust equation (15) by substituting v_{max} for $v'_{max} = v_{max} - v_{min}$. This method does not fill in missing values, and it is limited in the presence of outliers as most values would go within the same bin. We can solve this problem with a different discretization method (where bins have the same number of elements) but we have not implemented it in the current version of EnDED.

Applying EnDED to networks constructed from simulated and real data

We applied EnDED to association networks constructed from time series of simulated abundances and observed microbial abundances. The simulated networks were based on a gLV, while the real network was based on data from the BBMO. For the methods II and DPI we also included the corresponding abundance tables, and environmental factors. EnDED was run with the OL threshold of 60%. We set the significance threshold for the II score to 0.05 and used 1,000 iterations.

Evaluation of EnDED's performance

Simulated network

We evaluated EnDED with the simulated interaction matrices, which revealed the number of true positives (TP), true negatives (TN), false negatives (FN), and false positives (FP) before and after removing associations that were classified as environmentally-driven. We have assumed that all associations not present in the interaction matrices, are environmentally-driven. We consider P as the number of all false associations, both true positive and false negative detected environmentally-driven edges: $P = TP + FN$, and N as the number of all true interactions, i.e. all true negative and false positive detected environmentally-driven edges: $N = TN + FP$. Then, we can calculate the true positive rate (sensitivity) by dividing the number of true positives by the number of all real positives: $TPR = (TP)/(P)$. Equivalently, we can also calculate the true negative rate (specificity) by dividing the number of true negatives by the number of all real negatives, $TNR = (TN)/(N)$. The false positive rate (fall out) is the complementary to TNR, i.e. $FPR = 1 - TNR$. The positive predictive value can be calculated by dividing the number of true positives by the sum of all predicted positives, $PPV = (TP)/(TP + FP)$. Note PPV is also called precision. In the Discussion section we used precision to refer to the networks ability in retrieving true interactions. Here we use PPV in EnDEDs ability to remove false associations. The accuracy is calculated by dividing the sum of true positives and true negatives by the sum of all real positives and real negatives, $ACC = (TP + TN)/(P + N)$.

Real Dataset

Literature based database

Real network evaluation is limited since the true interactions and the microorganisms that do not interact with each other are poorly known. We assessed true interactions known in the literature based on the genus, which are compiled within the Protist Interaction Database, PIDA Bjorbækmo et al. (2019). On October 15th 2019, PIDA contained 2,448 interactions. Although our dataset contains protists as well as bacteria, we were unable to evaluate interactions between bacteria.

Jaccard index

In ecology, the Jaccard index (Jaccard similarity coefficient) is normally used for communities. Here, for each pair of microorganisms in the BBMO network, we computed the Jaccard index as the number of samples in which both microorganisms occur, divided by the number of samples in which at least one of the two microorganisms is present.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

EnDED is available here: https://github.com/InaMariaDeutschmann/EnDED. It also contains the file FromDataSimulationToEvaluatingEnDED.RMD, which contains R code to generate simulated abundance tables, commands to run eLSA network construction and EnDED, as well as the command to run a C++ program (included as well) and R code used for evaluation. The folder BBMO data contains the BBMO abundance table, the taxonomic classification table, and the BBMO network including results of EnDED.

Competing interests

The authors declare that they have no competing interests.

Funding

This project and IMD received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 675752 (Singek: http://www.singek.eu). RL was supported by a Ramón y Cajal fellowship (RYC-

2013-12554, MINECO, Spain). This work was also supported by the projects INTERACTOMICS (CTM2015-69936-P, MINECO, Spain) and MicroEcoSystems (240904, RCN, Norway) to RL.

Author's contributions

IMD, GLM, JR, KF and RL designed and conceived the project. IMD performed data analysis, data simulation, and implementation of EnDED. IMD received substantial feedback on established indirect detection methods from GLM and KF, on data simulation from SMV and KF, on network construction from AKK, and on evaluation of EnDED from AKK (literature based database for real dataset), GLM and KF (measurements for simulation dataset). RL processed the amplicon data from BBMO generating OTU tables. AKK ran the eLSA network construction tool for the BBMO dataset and IMD ran the tool for the simulation datasets. RL provided funding for the project. The original draft was written by IMD. IMD, GLM, AKK, SMV, KF and RL contributed substantially to manuscript revisions. All authors approved the final version of the manuscript.

Acknowledgements

We thank all members of the Blanes Bay Microbial Observatory sampling team and the multiple projects funding this collaborative effort over the years. Thanks to collaborators at www.thepapermill.eu for help with critical reading in the early stages of the manuscript. Part of the analyses have been performed at the Marbits bioinformatics core at ICM-CSIC (https://marbits.icm.csic.es).

References

Ai, D., Li, X., Pan, H., Chen, J., Cram, J.A., Xia, L.C.: Explore mediated co-varying dynamics in microbial community using integrated local similarity and liquid association analysis. BMC Genomics 20(2), 185 (2019). doi:10.1186/s12864-019-5469-8

Aitchison, J.: A new approach to null correlations of proportions. Journal of the International Association for Mathematical Geology 13(2), 175–189 (1981). doi:10.1007/BF01031393

Alipanahi, B., Frey, B.J.: Network cleanup. Nature Biotechnology 31(8), 714–715 (2013). doi:10.1038/nbt.2657

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2

Apprill, A., McNally, S., Parsons, R., Weber, L.: Minor revision to v4 region ssu rrna 806r gene primer greatly increases detection of sar11 bacterioplankton. Aquatic Microbial Ecology 75(2), 129–137 (2015). doi:10.3354/ame01753

Barzel, B., Barabási, A.-L.: Network link prediction by global silencing of indirect correlations. Nature Biotechnology 31(8), 720–725 (2013). doi:10.1038/nbt.2601

Bashan, A., Gibson, T.E., Friedman, J., Carey, V.J., Weiss, S.T., Hohmann, E.L., Liu, Y.-Y.: Universality of human microbial dynamics. Nature 534(7606), 259–262 (2016). doi:10.1038/nature18301

Benincà, E., Dakos, V., Van Nes, E.H., Huisman, J., Scheffer, M.: Resonance of plankton communities with temperature fluctuations. The American Naturalist 178(4), 85–95 (2011). doi:10.1086/661902. PMID: 21956036

Berry, D., Widder, S.: Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Frontiers in Microbiology 5, 219 (2014). doi:10.3389/fmicb.2014.00219

Bjorbækmo, M.F.M., Evenstad, A., Røsæg, L.L., Krabberød, A.K., Logares, R.: The planktonic protist interactome: where do we stand after a century of research? The ISME Journal (2019). doi:10.1038/s41396-019-0542-5

Brisson, V., Schmidt, J., Northen, T.R., Vogel, J.P., Gaudin, A.: A new method to correct for habitat filtering in microbial correlation networks. Frontiers in Microbiology 10, 585 (2019). doi:10.3389/fmicb.2019.00585

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature Methods 13(7), 581–583 (2016). doi:10.1038/nmeth.3869

Cover, T.M., Thomas, J.A.: Inequalities in information theory. In: Schilling, D.L., Cover, T.M., Thomas, J.A. (eds.) Elements of Information Theory, pp. 482–509. John Wiley & Sons, Ltd, (2001). Chap. 16. doi:10.1002/0471200611.ch16. https://onlinelibrary.wiley.com/doi/abs/10.1002/0471200611.ch16

Dam, P., Fonseca, L.L., Konstantinidis, K.T., Voit, E.O.: Dynamic models of the complex microbial metapopulation of lake mendota. npj Systems Biology and Applications 2(1), 16007 (2016). doi:10.1038/npjsba.2016.7

DeLong, E.F.: The microbial ocean from genomes to biomes. Nature 459(7244), 200–206 (2009). doi:10.1038/nature08059

Deutschmann, I.M., Lima-Mendez, G., Krabberød, A.K., Raes, J., Faust, K., Logares, R.: Environmentally-Driven Edge Detection Program. (2019). doi: 10.5281/zenodo.3271729. EnDED v1.0.1, 15. June 2019. https://github.com/InaMariaDeutschmann/EnDED

Falkowski, P.G., Fenchel, T., Delong, E.F.: The microbial engines that drive earth's biogeochemical cycles. Science 320(5879), 1034–1039 (2008). doi:10.1126/science.1153213

Faust, K.: Towards a better understanding of microbial community dynamics through high-throughput cultivation and data integration. mSystems 4(3) (2019). doi:10.1128/mSystems.00101-19

965
966 Faust, K., Raes, J.: Microbial interactions: from networks to models. Nature Reviews
967 Microbiology 10(8), 538–550 (2012). doi:10.1038/nrmicro2832
968
969 Faust, K., Raes, J.: Conet app: inference of biological association networks using cytoscape
970 [version 2; peer review: 2 approved]. F1000Research 5(1519) (2016).
971 doi:10.12688/f1000research.9050.2
972
973 Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., Huttenhower, C.:
974 Microbial co-occurrence relationships in the human microbiome. PLoS Computational
975 Biology 8(7), 1–17 (2012). doi:10.1371/journal.pcbi.1002606
976
977 Feizi, S., Marbach, D., Médard, M., Kellis, M.: Network deconvolution as a general method
978 to distinguish direct dependencies in networks. Nature Biotechnology 31(8), 726–733
979 (2013). doi:10.1038/nbt.2635
980
981 Fernandes, A.D., Gloor, G.B.: Mutual information is critically dependent on prior
982 assumptions: would the correct estimate of mutual information please identify itself?
983 Bioinformatics 26(9), 1135–1139 (2010). doi:10.1093/bioinformatics/btq111
984
985 Friedman, J., Alm, E.J.: Inferring correlation networks from genomic survey data. PLOS
986 Computational Biology 8(9), 1–11 (2012). doi:10.1371/journal.pcbi.1002687
987
988 Gasol, J.M., Cardelús, C., G Morán, X.A., Balagué, V., Forn, I., Marrasé, C., Massana, R.,
989 Pedrós-Alió, C., Montserrat Sala, M., Simó, R., Vaqué, D., Estrada, M.: Seasonal patterns in
990 phytoplankton photosynthetic parameters and primary production at a coastal nw
991 mediterranean site. Scientia Marina 80(S1), 63–77 (2016). doi:10.3989/scimar.04480.06E
992
993 Ghassami, A., Kiyavash, N.: Interaction information for causal inference: The case of
994 directed triangle. In: 2017 IEEE International Symposium on Information Theory (ISIT), pp.
995 1326–1330 (2017)
996
997 Giner, C.R., Balagué, V., Krabberød, A.K., Ferrera, I., Reñé, A., Garcés, E., Gasol, J.M.,
998 Logares, R., Massana, R.: Quantifying long-term recurrence in planktonic microbial
999 eukaryotes. Molecular Ecology 28(5), 923–935 (2019). doi:10.1111/mec.14929
1000
1001 Gonze, D., Coyte, K.Z., Lahti, L., Faust, K.: Microbial communities as dynamical systems.
1002 Current Opinion in Microbiology 44, 41–49 (2018). doi:10.1016/j.mib.2018.07.004.
1003 Microbiota
1004
1005 Grasshoff, K., Kremling, K., Ehrhardt, M.: Methods of Seawater Analysis. John Wiley &
1006 Sons, (2009)
1007
1008 Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G.,
1009 de Vargas, C., Decelle, J., del Campo, J., Dolan, J.R., Dunthorn, M., Edvardsen, B.,
1010 Holzmann, M., Kooistra, W.H.C.F., Lara, E., Le Bescot, N., Logares, R., Mahé, F., Massana,
1011 R., Montresor, M., Morard, R., Not, F., Pawlowski, J., Probert, I., Sauvadet, A.-L., Siano, R.,

1012 Stoeck, T., Vaulot, D., Zimmermann, P., Christen, R.: The protist ribosomal reference
1013 database (pr2): a catalog of unicellular eukaryote small sub-unit rrna sequences with curated
1014 taxonomy. Nucleic Acids Research 41(D1), 597–604 (2012). doi:10.1093/nar/gks1160
1015
1016 Haydon, D.: Pivotal assumptions determining the relationship between stability and
1017 complexity: An analytical synthesis of the stability-complexity debate. The American
1018 Naturalist 144(1), 14–29 (1994). doi:10.1086/285658
1019
1020 Herlemann, D.P., Labrenz, M., Jürgens, K., Bertilsson, S., Waniek, J.J., Andersson, A.F.:
1021 Transitions in bacterial communities along the 2000km salinity gradient of the baltic sea. The
1022 ISME Journal 5(10), 1571–1579 (2011). doi:10.1038/ismej.2011.41
1023
1024 Jackson, D.A.: Stopping rules in principal components analysis: A comparison of heuristical
1025 and statistical approaches. Ecology 74(8), 2204–2214 (1993). doi:10.2307/1939574
1026
1027 Jang, I.S., Margolin, A., Califano, A.: haracne: improving the accuracy of regulatory model
1028 reverse engineering via higher-order data processing inequality tests. Interface Focus 3(4),
1029 20130011 (2013). doi:10.1098/rsfs.2013.0011
1030
1031 Kallmeyer, J., Pockalny, R., Adhikari, R.R., Smith, D.C., D'Hondt, S.: Global distribution
1032 of microbial abundance and biomass in subseafloor sediment. Proceedings of the National
1033 Academy of Sciences 109(40), 16213–16216 (2012). doi:10.1073/pnas.1203849109
1034
1035 Kettle, H., Holtrop, G., Louis, P., Flint, H.J.: micropop: Modelling microbial populations and
1036 communities in r. Methods in Ecology and Evolution 9(2), 399–409 (2018).
1037 doi:10.1111/2041-210X.12873
1038
1039 Klemm, K., Eguíluz, V.M.: Growing scale-free networks with small-world behavior.
1040 Physical Review E 65(5), 057102 (2002). doi:10.1103/PhysRevE.65.057102
1041
1042 Krabberød, A.K., Bjorbækmo, M.F.M., Shalchian-Tabrizi, K., Logares, R.: Exploring the
1043 oceanic microeukaryotic interactome with metaomics approaches. Aquatic Microbial
1044 Ecology 79(1), 1–12 (2017). doi:10.3354/ame01811
1045
1046 Kurtz, Z.D., Bonneau, R., Müller, C.L.: Disentangling microbial associations from hidden
1047 environmental and technical factors via latent graphical models. bioRxiv (2019).
1048 doi:10.1101/2019.12.21.885889
1049
1050 Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse
1051 and compositionally robust inference of microbial ecological networks. PLOS
1052 Computational Biology 11(5), 1–25 (2015). doi:10.1371/journal.pcbi.1004226
1053
1054 Layeghifard, M., Hwang, D.M., Guttman, D.S.: Disentangling interactions in the
1055 microbiome: A network perspective. Trends in Microbiology 25(3), 217–228 (2017).
1056 doi:10.1016/j.tim.2016.11.008
1057
1058 Legendre, P., Legendre, L.F.: Numerical Ecology vol. 24. Elsevier, (2012)

Li, C., Lim, K.M.K., Chng, K.R., Nagarajan, N.: Predicting microbial interactions through computational approaches. Methods 102, 12–19 (2016). doi:10.1016/j.ymeth.2016.02.019. Pan-omics analysis of biological data

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A.M., Coppola, L., Cornejo-Castillo, F.M., d'Ovidio, F., De Meester, L., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans coordinators, Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S.G., Sunagawa, S., Bork, P., Sullivan, M.B., Karsenti, E., Bowler, C., de Vargas, C., Raes, J.: Determinants of community structure in the global plankton interactome. Science 348(6237), 1262073 (2015). doi:10.1126/science.1262073

Locey, K.J., Lennon, J.T.: Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences 113(21), 5970–5975 (2016). doi:10.1073/pnas.1521291113

Lv, X., Zhao, K., Xue, R., Liu, Y., Xu, J., Ma, B.: Strengthening insights in microbial ecological networks from theory to applications. mSystems 4(3), 00124–19 (2019). doi:10.1128/mSystems.00124-19

Mandakovic, D., Rojas, C., Maldonado, J., Latorre, M., Travisany, D., Delage, E., Bihouée, A., Jean, G., Díaz, F.P., Fernández-Gómez, B., Cabrera, P., Gaete, A., Latorre, C., Gutiérrez, R.A., Maass, A., Cambiazo, V., Navarrete, S.A., Eveillard, D., González, M.: Structure and co-occurrence patterns in microbial communities under acute environmental stress reveal ecological factors fostering resilience. Scientific Reports 8(1), 5875 (2018). doi:10.1038/s41598-018-23931-0

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., Califano, A.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(1), 7 (2006). doi:10.1186/1471-2105-7-S1-S7

McCarren, J., Becker, J.W., Repeta, D.J., Shi, Y., Young, C.R., Malmstrom, R.R., Chisholm, S.W., DeLong, E.F.: Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. Proceedings of the National Academy of Sciences 107(38), 16420–16427 (2010). doi:10.1073/pnas.1010732107

Meyer, P.E., Lafitte, F., Bontempi, G.: minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 9(1), 461 (2008). doi:10.1186/1471-2105-9-461

Moritz, S., Gatscha, S.: imputeTS: Time Series Missing Value Imputation. (2017). R package version 2.8. https://github.com/SteffenMoritz/imputeTS

North, B.V., Curtis, D., Sham, P.C.: A note on the calculation of empirical p values from monte carlo procedures. The American Journal of Human Genetics 71(2), 439–441 (2002). doi:10.1086/341527

Novak, M., Yeakel, J.D., Noble, A.E., Doak, D.F., Emmerson, M., Estes, J.A., Jacob, U., Tinker, M.T., Wootton, J.T.: Characterizing species interactions to understand press perturbations: What is the community matrix? Annual Review of Ecology, Evolution, and Systematics 47(1), 409–432 (2016). doi:10.1146/annurev-ecolsys-032416-010215

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H.: Vegan: Community Ecology Package. (2019). R package version 2.5.-6 (network simulation), and 2.4-2 (BBMO data).
https://CRAN.R-project.org/package=vegan

Pascual-García, A., Tamames, J., Bastolla, U.: Bacteria dialog with santa rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? BMC Microbiology 14, 284 (2014). doi:10.1186/s12866-014-0284-5

Pinedo-González, P., West, A.J., Tovar-Sánchez, A., Duarte, C.M., Marañón, E., Cermeño, P., González, N., Sobrino, C., Huete-Ortega, M., Fernández, A., López-Sandoval, D.C., Vidal, M., Blasco, D., Estrada, M., Sañudo-Wilhelmy, S.A.: Surface distribution of dissolved trace metals in the oligotrophic ocean and their influence on phytoplankton biomass and productivity. Global Biogeochemical Cycles 29(10), 1763–1781 (2015). doi:10.1002/2015GB005149

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O.: The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic Acids Research 41(D1), 590–596 (2012). doi:10.1093/nar/gks1219

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019). R Foundation for Statistical Computing. https://www.R-project.org/

Röttjers, L., Faust, K.: From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiology Reviews 42(6), 761–780 (2018). doi:10.1093/femsre/fuy030

Soetaert, K., Petzoldt, T., Setzer, R.W.: Solving differential equations in R: Package deSolve. Journal of Statistical Software 33(9), 1–25 (2010). doi:10.18637/jss.v033.i09

Stein, R.R., Bucci, V., Toussaint, N.C., Buffie, C.G., Rätsch, G., Pamer, E.G., Sander, C., Xavier, J.B.: Ecological modeling from time-series inference: Insight into dynamics and stability of intestinal microbiota. PLOS Computational Biology 9(12), 1–11 (2013). doi:10.1371/journal.pcbi.1003388

Tackmann, J., Rodrigues, J.a.F.M., von Mering, C.: Rapid inference of direct interactions in

1153 large-scale ecological networks from heterogeneous microbial sequencing data. Cell Systems
1154 9(3), 286–2968 (2019). doi:10.1016/j.cels.2019.08.002
1155
1156 The Human Microbiome Project Consortium: Huttenhower, C., Gevers, D., Knight, R.,
1157 Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G.,
1158 Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V.,
1159 Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M.,
1160 Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-
1161 Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi,
1162 H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri,
1163 V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A., Buhay,
1164 C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.-M.A.,
1165 Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S.,
1166 Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C.,
1167 Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael
1168 Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust,
1169 K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di
1170 Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia,
1171 N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A.,
1172 Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn,
1173 T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M.,
1174 Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley,
1175 S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar,
1176 C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M.,
1177 Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C.A., Dwayne
1178 Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M.,
1179 Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes,
1180 P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C.,
1181 O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J.,
1182 Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J.,
1183 Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller,
1184 B., Rogers, Y.-H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J.,
1185 Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata,
1186 N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U.,
1187 Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G.,
1188 Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J.,
1189 Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren,
1190 W., Watson, M.A., Wellington, C., Wetterstrand, K.A., White, J.R., Wilczek-Boney, K., Wu,
1191 Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang,
1192 L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K.,
1193 Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K., White, O.:
1194 Structure, function and diversity of the healthy human microbiome. Nature 486(7402), 207–
1195 214 (2012). doi:10.1038/nature11234
1196
1197 Vallina, S.M., Martinez-Garcia, R., Smith, S.L., Bonachela, J.A.: Models in microbial
1198 ecology. In: Schmidt, T.M. (ed.) Encyclopedia of Microbiology (Fourth Edition), Fourth
1199 edition edn., pp. 211–246. Academic Press, Oxford (2019). doi:10.1016/B978-0-12-809633-

8.20789-9. http://www.sciencedirect.com/science/article/pii/B9780128096338207899

Veech, J.A.: Significance testing in ecological null models. Theoretical Ecology 5(4), 611–616 (2012). doi:10.1007/s12080-012-0159-z

Verny, L., Sella, N., Affeldt, S., Singh, P.P., Isambert, H.: Learning causal networks with latent variables from multivariate information in genomic data. PLOS Computational Biology 13(10), 1–25 (2017). doi:10.1371/journal.pcbi.1005662

Villaverde, A.F., Becker, K., Banga, J.R.: Premer: A tool to infer biological networks. IEEE/ACM Trans. Comput.Biol. Bioinformatics 15(4), 1193–1202 (2018). doi:10.1109/TCBB.2017.2758786

Villaverde, A.F., Ross, J., Morán, F., Banga, J.R.: Mider: Network inference with mutual information distance and entropy reduction. PLOS ONE 9(5), 1–15 (2014). doi:10.1371/journal.pone.0096732

Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R.: Naïve bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73(16), 5261–5267 (2007). doi:10.1128/AEM.00062-07

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., Xia, L.C., Xu, Z.Z., Ursell, L., Alm, E.J., Birmingham, A., Cram, J.A., Fuhrman, J.A., Raes, J., Sun, F., Zhou, J., Knight, R.: Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal 10(7), 1669–1681 (2016). doi:10.1038/ismej.2015.235

Whitman, W.B., Coleman, D.C., Wiebe, W.J.: Prokaryotes: The unseen majority. Proceedings of the National Academy of Sciences 95(12), 6578–6583 (1998). doi:10.1073/pnas.95.12.6578

Worden, A.Z., Follows, M.J., Giovannoni, S.J., Wilken, S., Zimmerman, A.E., Keeling, P.J.: Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. Science 347(6223) (2015). doi:10.1126/science.1257594

Xia, L.C., Steele, J.A., Cram, J.A., Cardon, Z.G., Simmons, S.L., Vallino, J.J., Fuhrman, J.A., Sun, F.: Extended local similarity analysis (elsa) of microbial community and other time series data with replicates. BMC Systems Biology 5(2), 15 (2011). doi:10.1186/1752-0509-5-S2-S15

Xia, L.C., Ai, D., Cram, J., Fuhrman, J.A., Sun, F.: Efficient statistical significance approximation for local similarity analysis of high-throughput time series data. Bioinformatics 29(2), 230–237 (2012). doi:10.1093/bioinformatics/bts668

Xiao, Y., Angulo, M.T., Friedman, J., Waldor, M.K., Weiss, S.T., Liu, Y.-Y.: Mapping the ecological networks of microbial communities. Nature Communications 8(1), 2042 (2017). doi:10.1038/s41467-017-02090-2

Yang, Y., Chen, N., Chen, T.: Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical bayesian statistical model. Cell Systems 4(1), 129–1375 (2017). doi:10.1016/j.cels.2016.12.012

Yoon, H.S., Price, D.C., Stepanauskas, R., Rajah, V.D., Sieracki, M.E., Wilson, W.H., Yang, E.C., Duffy, S., Bhattacharya, D.: Single-cell genomics reveals organismal interactions in uncultivated marine protists. Science 332(6030), 714–717 (2011)

Zhao, J., Zhou, Y., Zhang, X., Chen, L.: Part mutual information for quantifying direct associations in networks. Proceedings of the National Academy of Sciences 113(18), 5130–5135 (2016). doi:10.1073/pnas.1522586113

Zoppoli, P., Morganella, S., Ceccarelli, M.: Timedelay-aracne: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics 11(1), 154 (2010). doi:10.1186/1471-2105-11-154

Figures

Figure 1: **Evaluation of EnDED: intersection combination and individual methods on simulated networks** Using 1,000 simulated networks, and 1,000 simulated networks incorporating noise, we evaluate EnDED's performance. Plot A) displays the evaluation measurements true positive rate (TRP), true negative rate (TNR), accuracy (ACC), and positive predictive value (PPV) for each individual method, i.e. Sign Pattern (SP), Overlap (OL), Interaction Information (II), and Data Processing Inequality (DPI), as well as the intersection combination (COMBI). SP and OL perform best according to TRP and ACC, while the intersection combination performs best according to TNR. All methods performed well according to PPV. The intersection combination, DPI and II performed better on noisy data according to TNR because less edges were removed along with less true interactions. Plot B) displays the ROC curve for each environmentally-driven edge detection method as well as their intersection combination.

Figure 2: **Quantification of environmentally-driven associations in the BBMO network**

For A) to D), the upper left figure shows the number (or fraction) of microbial associations divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations (BE), and Eukaryote-Eukaryote associations (E). The upper right figure shows the number (or fraction) of associations divided by size-fractions: association within the nano size fraction (n), within the pico size fraction (p), and between these two size fractions (np). The figure in the middle shows all triplets connected to an environmental parameter: Temperature (Temp), Day length (Day), *Micromonas* (Mic), photosynthetic Nanoflagellates (PNF), heterotrophic Nanoflagellates (HNF), Chlorophyll (Chl), *Synechococcus* (Syn), inorganic nutrients NO_3^- (NO3), SiO_2 (Si), and NO_2^- (NO2), as well as heterotrophic Prokaryotes (H.Pro). The figure at the bottom shows the number of edges divided in how many triplets they have been found ranging from no triplets (0) to nine triplets. Figure A) and B) display the number of microbial associations of the BBMO network before applying EnDED. Positive associations are indicated with black, negative associations with red. Figure C) and D) indicate in blue the fraction of environmentally-driven edges among the positive (C) and negative (D) microbial associations. E) The upper left network shows in black the positive and in red the negative associations. The upper right network shows in green the number of triplets a microbial edge is in ranging from one (light green) to nine (dark green), and no triplet (black). The lower network shows in blue the environmentally-driven associations that were detected by the intersection combination of the four methods Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality.

Figure 3: **EnDED Methods Overview** EnDED is an implementation of four methods aiming to determine whether an edge between two microorganisms is indirect through the action of an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction

Information, and Data Processing Inequality (see Methods). Each method can be used individually or in combination. Here, we show the intersection combination approach, i.e. only if all methods classify an edge as indirect, it is removed from the network. Otherwise, the edge is classified as not indirect and kept in the network.

Tables

Table 1 Jaccard index of edges The BBMO network before applying EnDED contained 33, 832 edges of which 4,806 (14.2%) are environmentally-driven (indirect). Considering the Jaccard index for these indirect edges, 1,433 (29.8% of indirect edges) score above 50%, and 3, 373 (70.2%) score below or equal to 50%. In contrast, 60.6% of edges not considered as indirect have a Jaccard index above 50%, and 39.4% of all not indirect edges have a Jaccard index equal or below 50%.

	All	Jaccard index>50	Jaccard index≤50
BBMO network	33, 832 (100%)	19, 015 (56.2%)	14, 817 (43.8%)
positive edges	27, 700 (81.9%)	18, 832 (68.0%)	8, 868 (32.0%)
negative edges	6, 132 (18.1%)	183 (3.0%)	5, 949 (97.0%)
indirect (intersection combination)	4, 806 (14.2%)	1,433 (29.8%)	3, 373 (70.2%)
not indirect (all)	29, 026 (85.8%)	17, 582 (60.6%)	11,444 (39.4%)
not indirect (min 1 triplet)	23, 404 (69.2%)	14, 822 (63.3%)	8, 582 (36.7%)
not indirect (no triplet)	5, 622 (16.7%)	2, 760 (49.1%)	2, 862 (50.9%)
Sign Pattern	28, 210 (83.4%)	16, 255 (57.6%)	11,955 (42.4%)
Overlap	28, 210 (83.4%)	16, 255 (57.6%)	11,955 (42.4%)
Interaction Information	12, 960 (38.3%)	7, 808 (60.2%)	5, 152 (39.8%)
Data Processing Inequality	10, 610 (31.4%)	3, 328 (31.4%)	7, 282 (68.6%)

Table 2 Interactions found in the BBMO network These that appear in the literature, including whether or not the associations were removed or kept by EnDED. For example, the interaction between OTUs classified as *Dia. Thalassiosira* and OTUs classified as F. unknown *Flavobacteriia* has been found 18 times in the network: 7 were removed and 11

1320 were kept.

Microorganisms	EnDED	ID in PIDA
Dia. Thalassiosira - *Dino. Heterocapsa*	1 removed	1665
Dia. Thalassiosira - F. unknown *Flavobacteriia*	7 removed	2199
	11 kept	
Dino. Heterocapsa - *Dino. Prorocentrum*	1 kept	1501, 1511
Dino. Gymnodinium - *C. Strombidium*	1 kept	1209
Dino. Gyrodinium - *Dino. Heterocapsa*	1 kept	1313, 1314, 1780, 1783
Dino. Prorocentrum - *Dino. Gymnodinium*	2 kept	1499
Dino. Prorocentrum - *Dino. Prorocentrum*	4 kept	1509, 1510
Dino. Prorocentrum - *Dino. Scrippsiella*	2 kept	1513
F. unknown *Flavobacteriia* - *Dia. Pseudo-nitzschia*	1 kept	2196

1321 Abbreviations indicate Dia - Diatomea; Dino - Dinoflagellata; C - Ciliophora; F - *Flavobacteriia*; ID in PIDA
1322 refers to the number PIDA gave an interaction described in literature.

Figures

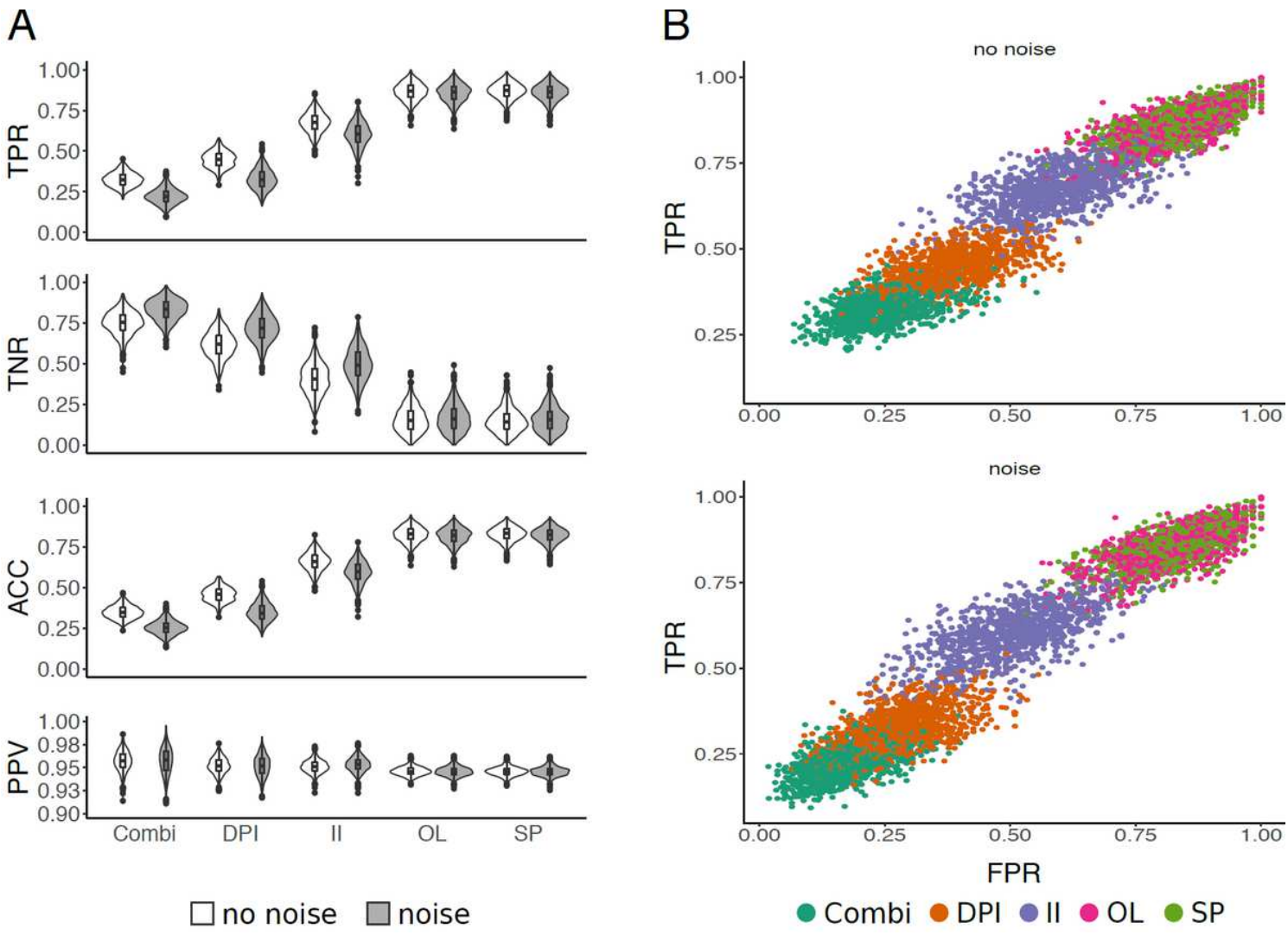

Figure 1

Evaluation of EnDED: intersection combination and individual methods on simulated networks Using 1,000 simulated networks, and 1,000 simulated networks incorporating noise, we evaluate EnDED's performance. Plot A) displays the evaluation measurements true positive rate (TRP), true negative rate (TNR), accuracy (ACC), and positive predictive value (PPV) for each individual method, i.e. Sign Pattern (SP), Overlap (OL), Interaction Information (II), and Data Processing Inequality (DPI), as well as the intersection combination (COMBI). SP and OL perform best according to TRP and ACC, while the intersection combination performs best according to TNR. All methods performed well according to PPV. The intersection combination, DPI and II performed better on noisy data according to TNR because less edges were removed along with less true interactions. Plot B) displays the ROC curve for each environmentally-driven edge detection method as well as their intersection combination.

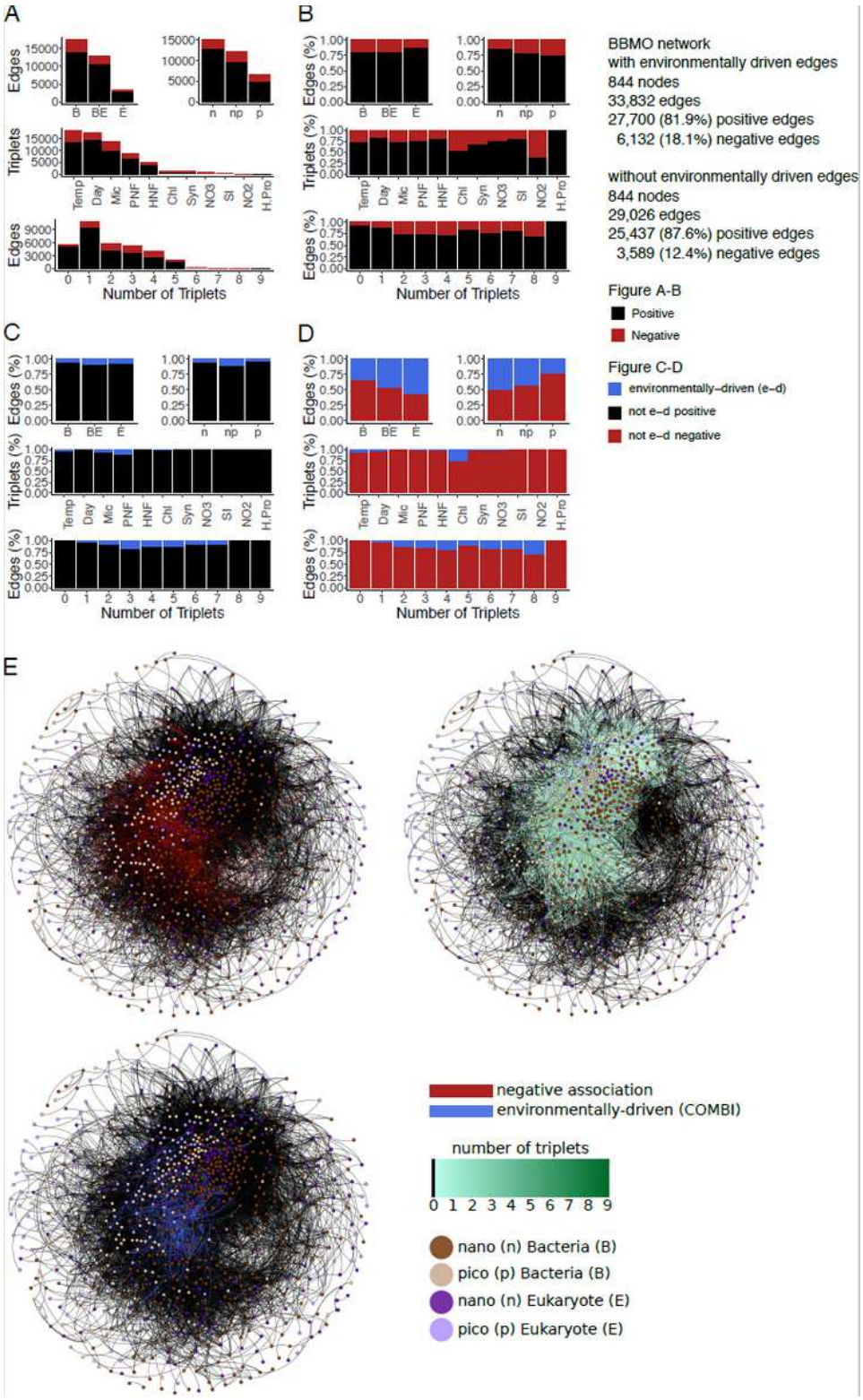

Figure 2

Quantification of environmentally-driven associations in the BBMO network For A) to D), the upper left figure shows the number (or fraction) of microbial associations divided by domain: Bacteria-Bacteria associations (B), Bacteria-Eukaryote associations (BE), and Eukaryote-Eukaryote associations (E). The upper right figure shows the number (or fraction) of associations divided by size-fractions: association within the nano size fraction (n), within the pico size fraction (p), and between these two size fractions

(np). The figure in the middle shows all triplets connected to an environmental parameter: Temperature (Temp), Day length (Day), Micromonas (Mic), photosynthetic Nanoflagellates (PNF), heterotrophic Nanoflagellates (HNF), Chlorophyll (Chl), Synechococcus (Syn), inorganic nutrients NO3- (NO3), SiO2 (Si), and NO2- (NO2), as well as heterotrophic Prokaryotes (H.Pro). The figure at the bottom shows the number of edges divided in how many triplets they have been found ranging from no triplets (0) to nine triplets. Figure A) and B) display the number of microbial associations of the BBMO network before applying EnDED. Positive associations are indicated with black, negative associations with red. Figure C) and D) indicate in blue the fraction of environmentally-driven edges among the positive (C) and negative (D) microbial associations. E) The upper left network shows in black the positive and in red the negative associations. The upper right network shows in green the number of triplets a microbial edge is in ranging from one (light green) to nine (dark green), and no triplet (black). The lower network shows in blue the environmentally-driven associations that were detected by the intersection combination of the four methods Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality.

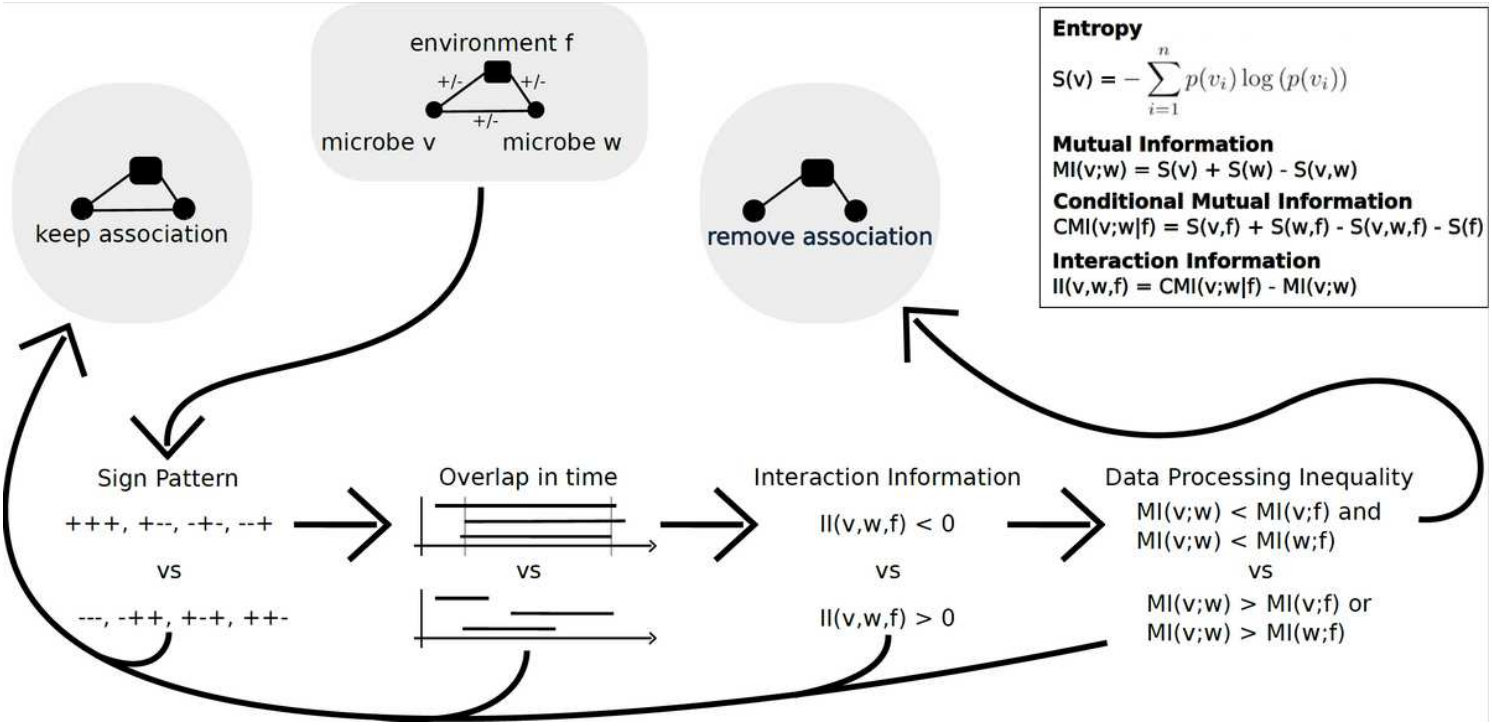

Figure 3

EnDED Methods Overview EnDED is an implementation of four methods aiming to determine whether an edge between two microorganisms is indirect through the action of an environmental factor. The four methods are: Sign Pattern, Overlap, Interaction Information, and Data Processing Inequality (see Methods). Each method can be used individually or in combination. Here, we show the intersection combination approach, i.e. only if all methods classify an edge as indirect, it is removed from the network. Otherwise, the edge is classified as not indirect and kept in the network.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AdditionalFilesTableS1S2S3.docx