Relating network analyses to phylogenetic relatedness to infer protistan co-occurrences and co-exclusions in marine and terrestrial environments

Guillaume Lentendu[1,*] and Micah Dunthorn[2,3]

[1]Laboratory of Soil Biodiversity, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland

[2]Department of Eukaryotic Microbiology, University of Duisburg-Essen, D-45141 Essen, Germany

[3]Centre for Water and Environmental Research (ZWU), University of Duisburg-Essen, D-45141 Essen, Germany

[*]corresponding author: phone +41 32 718 22 61; email guillaume.lentendu@unine.ch

Abstract

We used two large-scale metabarcoding datasets to evaluate phylogenetic signals at global marine and regional terrestrial scales using co-occurrence and co-exclusion networks. Phylogenetic relatedness was estimated using either global pairwise sequence distance or phylogenetic distance and the significance of observed patterns relating networks and phylogenies were evaluated against two null models. In all datasets, we found that phylogenetically close OTUs significantly co-occurred more often, and OTUs with intermediate phylogenetic relatedness co-occurred less often, than expected by chance. Phylogenetically close OTUs co-excluded less often than expected by chance in the marine datasets only. Simultaneous excess of co-occurrences and co-exclusions were observed in the inversion zone between close and intermediate phylogenetic distance classes in marine surface. Similar patterns were observed by using either pairwise sequence or phylogenetic distances, and by using both null models. These results suggest that environmental filtering and dispersal limitation are the preponderant forces driving co-occurrence of protists in both environments, while signal of competitive exclusion was only detected in the marine surface environment. The discrepancy in the co-exclusion pattern is potentially linked to the individual

environments: water bodies are more homogeneous while tropical forest soils contain a myriad of nutrient rich micro-environment reducing the strength of mutual exclusion.

Introduction

There is a long history of research trying to elucidate why species are present in a specific environment and why multiple species are found together (Darwin, 1859; Gause, 1934; Humboldt & Bonpland, 1805). Species sharing the same ecological niche tend to co-occur due to environmental filtering and dispersal limitation. In turn, closely-related species are more likely to co-occur due to their shared evolutionary history (e.g., common ancestor, shared traits) and their potential limited dispersal- and establishment-abilities. These processes can be balanced by density-dependent negative biotic interactions, like competitive exclusion when functionally similar species are after the same resource and co-exclude themselves. Environmental filtering and dispersal limitation have been identified as the main drivers shaping the assembly of most protists in the environment (Boenigk et al., 2018; de Vargas et al., 2015; del Campo et al., 2015; Lentendu et al., 2018; Mahé et al., 2017; Singer et al., 2018; Wetzel et al., 2012), while competition have been only formally tested in laboratory conditions (Saleem, Fetzer, Dormann, Harms, & Chatzinotas, 2012; Violle, Nemergut, Pu, & Jiang, 2011). These mechanisms have been largely evaluated for macro-organisms in different environments (Cavender-Bares, Kozak, Fine, & Kembel, 2009; Kraft et al., 2015), but have not yet been broadly evaluated for microbes in natural environments, for which community ecological analyses have rarely integrated phylogenetic information.

In environmental microbial ecology, environmental filtering is often considered as the prevalent limiting parameter of species occurrence (Khomich, Kauserud, Logares, Rasconi, & Andersen, 2017; Lauber, Strickland, Bradford, & Fierer, 2008; Lentendu et al., 2018; Philippot et al., 2010; Singer et al., 2018; Tedersoo et al., 2016; Weißbecker et al., 2018; Zinger et al., 2011) and is directly linked to the ecological niche of microbes (i.e., the set of abiotic parameter ranges in which a species can leave in). Ecological niche of microbes is hardly measurable without

cultivation (Lennon, Aanderud, Lehmkuhl, & Schoolmaster, 2012; J. B. H. Martiny, Jones, Lennon, & Martiny, 2015), so that for large scale studies mostly based on non-cultivable microbes, function and functional similarity are either deducted from taxonomic or phylogenetic similarity of recovered sequences. Environmental filtering is inferred from the non-random co-occurrence of members of a taxa or a clade or from clade or taxa occurring in a restricted set of habitats. Thus, environmental filtering, when analyses in a phylogenetic context, often assumes phylogenetic niche conservatism, that is the long-term retention of ecological traits among closely related species (Wiens et al., 2010). Phylogenetic niche conservatism was shown in bacteria, mainly for complex functional traits which are conserved inside single clades (A. C. Martiny et al., 2013). Under phylogenetic niche conservatism, evolutionary close species are more likely to share the same ecological niche and thus tend to be filtered into the same habitats. With this assumption, environmental filtering can be tested using measures of phylogenetic divergence (e.g. MPD, MNTD, but see Tucker et al., 2017), with phylogenetic over-clustering (i.e. low phylogenetic divergence) being interpreted as sign for environmental filtering. This sample-wide approach has been used to support environmental filtering of trees, bacteria and protists along habitat and nutrient gradients (Horner-Devine & Bohannan, 2006; Kembel & Hubbell, 2006; Singer et al., 2018). However, it appears that most studies concluding on environmental filtering do not account for biotic interactions which could produce similar results (Kraft et al., 2015).

Competition is long known experimentally and it was hypothesized to drive co-exclusion in an initial experimental study involving protists (Gause, 1934). Competitive exclusion was first viewed as an evolutionary pressure which trigger trait divergence of related species, allowing them to escape competition and to persist in the same habitat, as originally observed for Darwin's finches (Darwin, 1859). This assumption was further formalized with the phylogenetic limiting similarity hypothesis, in which phylogenetic related species do compete stronger due to niche overlap thus limiting the number of related species which can coexist (Macarthur & Levins, 1967). By assuming

phylogenetic niche conservatism, it is expected that competitive exclusion will only affect closely related species, so that phylogenetic over-dispersion (i.e. high phylogenetic divergence) of natural communities is interpreted as a sign of competitive exclusion. This approach have allowed to identify one tree family presenting signs of competitive exclusion in a tropical forest (Manel et al., 2014). However, competition do not necessarily lead to exclusion when for example competition is symmetric or when other biotic interactions (e.g., mutualism or herbivory) reduce or neutralize the competition (Lamb & Cahill Jr., 2008; Müller, Hauzy, & Hulot, 2012; Olff & Ritchie, 1998). Further experimental evidences have shown that for protists species, competition will more quickly lead to exclusion when species are phylogenetically related, with a direct relation to phylogenetically conserved traits (e.g. mouth size Violle et al., 2011). The "paradox of the plankton" was also considered to be an opposite example of competitive exclusion, with the co-existence of high number of species using the same resources (Hutchinson, 1961). It was however shown that this pattern is explained by the competition itself which only leads to short term exclusion in a system never reaching an equilibrium (Huisman & Weissing, 1999). In plant ecology, studies measuring competition strength have shown that depending on clades or depending on soil conditions, there will be more or less competition between related species, so that no generalization of the 'competition-relatedness' hypothesis is possible (Burns & Strauss, 2011; Cahill, Kembel, Lamb, & Keddy, 2008). The exclusion of closely related species due to competition can thus be viewed as a special case of the coexistence theory (Mayfield & Levine, 2010). But so far, no large-scale study has tested for phylogenetic overdispersion and exclusion patterns in protists.

In today's very large environmental sequencing datasets, microbial taxa are characterize using operational taxonomic units (OTU) which are used as proxy to molecular species (Blaxter et al., 2005). At the same time, co-occurrence and co-exclusion networks analyses have become standard in environmental microbial ecology, with a predominance of studies interested in co-occurrence patterns among and between taxonomic groups with a presumed function (Chow, Kim,

Sachdeva, Caron, & Fuhrman, 2014; Lima-Mendez et al., 2015; Milici et al., 2016; Steele et al., 2011). To contrast with phylogenetic divergence analyses conducted at the sample level, co-occurrence and co-exclusion network analyses allow to extract statistically significant pair of co-occurring/co-excluding OTUs at the whole study level. By comparing observed co-occurrences to random co-occurrences among the regional pool of OTUs, signal for potential biotic interactions like parasitism, predation or viral infection have been disclosed (Lentendu et al., 2014; Lima-Mendez et al., 2015; Steele et al., 2011). By taking advantage of the modularity structure of the co-occurrence networks, microbial occurrences have also been linked to habitat preference, which can be interpreted as the signal for environmental filtering (de Menezes et al., 2014; Lentendu et al., 2014; Milici et al., 2016; Morriën et al., 2017). However, studies have yet to integrate the phylogenetic relatedness as an explaining parameter for network structure.

Here we describe a new analytical approach that aims to evaluate community assembly processes by decomposing the co-occurrence and co-exclusion networks among phylogenetic relatedness classes. By looking at excess or deficit of co-occurrence or co-exclusion in class of organism with increasing phylogenetic relatedness, we can test the possible assembly mechanisms in natural protistan communities. Under the assumption of phylogenetic niche conservatism, we tested the following hypotheses: i) if environmental filtering dominate, phylogenetically related OTUs will co-occur more and co-exclude less often than expected by chance and conversely for pairs of OTUs with intermediate phylogenetic relatedness; ii) if competitive exclusion dominate, phylogenetically related OTUs will co-occur less and co-exclude more often than expected by chance and conversely for pairs of OTUs with intermediate phylogenetic relatedness. To evaluate these hypotheses, we use two of the largest environmental sequencing protist datasets to date: the global marine subsurface dataset of de Vargas et al. (de Vargas et al., 2015), and the Neotropical rainforest soil dataset of Mahé et al. (2017). While both studies were primarily concerned by describing the occurrence of different taxa in different water bodies or forest soils, the current study

try to evaluate how phylogenetic relatedness could explain the distributions of protists at global and regional scales.

Material and Methods

The complete bash and R (R Core Team, 2017) scripts to reproduce the analyses are provided in HTML format (**File S1**). The full network calculation procedure is also available as a stand-alone software with multiple matrix normalization, randomization and thresholding options (https://github.com/lentendu/NetworkNullHPC).

Datasets

Two large-scale environmental sequencing projects that focused on protistan diversity were used here (available upon request). Protistan OTUs from the world's open oceans and seas came from de Vargas et al. (2015). This marine dataset is composed of 355 samples collected at the surface and deep chlorophyll maximum (DCM), which produced 366,800,845 protist reads of the V9 hyper-variable region of the SSU-rRNA locus that clustered into 302,663 OTUs. To allow for comparison, the version of this marine dataset used here was re-analyzed by Mahé et al. (2017). All filter-size classes libraries of either the surface or DCM at a single station were pooled together, thus the number of samples used here reduced to 47 for surface and 32 for DCM waters. Protistan OTUs from three lowland Neotropical rainforests came from Mahé et al. (2017). This terrestrial dataset is composed of 144 samples collected at the soil surface, which produced 46,652,206 protist reads of the V4 hyper-variable region of the SSU-rRNA locus that clustered into 26,860 OTUs. For sampling and sequencing information see the original publications (de Vargas et al., 2015; Mahé et al., 2017); for bioinformatic pipeline of reads cleaning, clustering with Swarm v2 (Mahé, Rognes, Quince, de Vargas, & Dunthorn, 2015), and taxonomic assignments using the Protist Ribosomal Reference database (Guillou et al., 2013) to protists see Mahé et al. (2017). It is important to note that this reference database does not reflect the exact current international agreement on the

taxonomy of protists (S. M. Adl et al., 2019) and each taxonomic path is reduced to eight taxonomic levels.

Co-occurrence and co-exclusion networks

To infer protistan co-occurrences and co-exclusions from the marine and terrestrial datasets, networks were constructed using OTUs following Connor et al. (2017). This method infer positive correlations (co-occurrences), which was expanded here to also infer negative correlations (co-exclusions). Resulting networks were composed of nodes (OTUs) that were connected by edges to one or more other nodes; these edges were either instances of co-occurrences or co-exclusions. First, to reduce computational load, OTUs occurring in less than 30% of marine and 10% of terrestrial samples were removed as well as samples with less than 20% of median read counts per sample in the terrestrial dataset. Low occurrence OTUs would never show any significant co-occurrence or co-exclusion using this method (Connor et al., 2017). The OTUs which passed the occurrence filter are later referred as the candidate OTUs. Second, read counts per sample were normalized using the log-ratio count method: reads were log transformed in order to reduce abundance bias due to PCR; counts were then normalized per sample to a median sequencing depth by multiplying read counts by the ratio of a minimum expected sequencing depth (half the median of original sample's read count) by the sample's total sum of read counts and rounding to integer. This normalization is preferable to rarefaction and/or relative abundance normalization, because it avoids random subsampling and variance inflation while taking into account the compositionality of the data (Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017; McMurdie & Holmes, 2014). Third, random noise was added to the normalized matrices in order to break ties when calculating Spearman's rank correlation coefficient (*rho*). Fourth, this random noise addition was repeated 1000 times (i.e., Monte Carlo sampling) to obtain a normal distribution of Spearman's *rho*. Fifth, the thresholds to detect a biological significant positive (co-occurrence) or negative (co-exclusion) correlation were determined with randomly shuffled and noise-added OTU matrices. This threshold

was set at the Spearman's *rho* for which the largest connected component of a network, build with edges equal and above this threshold for co-occurrence, or equal and below this threshold for co-exclusion, contains less than 1% of the total OTU number in at least 90 % of the 1000 random OTU matrices. The random shuffling was based on OTU abundance swaps constrained to each sample and was prefer to the original full count shuffling without fixed row and column sums because it preserved the slight positive shift in Spearman's rho as observed in natural communities (**Figure S1**). Sixth, observed edges with a Spearman's *rho* above or below the selected threshold in at least 90 % of the Monte Carlo sampling and with corrected Spearman's *rho* p.values (Benjamini & Hochberg, 1995) ≤ 0.01 in at least 90 % of the Monte Carlo sampling were considered as biological co-occurrence or co-exclusion, respectively. This procedure sets Spearman's *rho* co-occurrence thresholds at 0.58 for marine surface, 0.68 for marine DCM, and 0.45 for terrestrial. Spearman's *rho* co-exclusion thresholds were set at -0.52 for marine surface and -0.64 for marine DCM, and -0.24 for terrestrial (**Table 1**).

Pairwise sequence and phylogenetic distances

To infer the phylogenetic relatedness between the OTUs (nodes) in the constructed co-occurrence or co-exclusion networks, the OTU representatives (the most abundant strictly-identical amplicon) were used. These phylogenetic relatedness values between the OTUs were then overlaid along the edges in the networks. Two methods were used to infer the phylogenetic relatedness. First, pairwise sequence distances were calculated using a Needleman-Wunsh approximation as implemented in SUMATRA v1.0.34 (Mercier, Boyer, Bonin, & Coissac, 2013). This global pairwise sequence comparison did not account for any model of evolution. Second, phylogenetic distances were calculated by aligning the sequences using the FFT-NS-i strategy in MAFFT v7.407 (Katoh & Standley, 2013) and by finding the best maximum-likelihood tree using the GTRCAT model in RAxML 8.2.12 (Stamatakis, 2014) with 256 random starting trees. The phylogenetic distance between each tree tip was then calculated with the "cophenetic" function in R (R Core Team, 2017).

Null models

To infer if the associations between the networks (both co-occurrences or co-exclusions) and the phylogenetic relatedness differed significantly from randomness, two null models were constructed. Null model 1 followed Hardy (model 1s, 2008) by generating random phylogenetic relatednesses values between nodes. These random values were made by a custom script that randomly shuffled the tip of the phylogenetic tree limited to the OTUs presented in the co-occurrence or co-exclusion networks. The same random re-ordering of OTUs was applied to both pairwise sequence and phylogenetic distance matrices (i.e. re-ordering row and column names) and the distance value for each co-occurring or co-excluding OTU pair was extracted. Null model 1 aimed to test whether co-occurring or co-excluding OTUs are more or less phylogenetically related than expected by chance. Null model 2 followed Chung and Lu (2002) by generating random edges between nodes. In these random networks, the total amount of edges remained the same as in the observed network, but the number of edges from an individual node was drawn from a probability distribution in which edge probability depends on the cumulative observed degree of the two nodes involved. This null Chung-Lu model produced networks with characteristics (e.g. modularity, diameter, clustering coefficient) more similar to natural networks compared to the most widely used null Erdős-Rényi model (Connor et al., 2017), and thus minimizes the number of parameters modified compared to the observed network. The random networks were made using the “sample_fitness” function in the R igraph package (Csardi & Nepusz, 2006). Null model 2 aimed to test whether phylogenetically related OTU co-occurred or co-excluded more or less than expected by chance.

Statistical analyses

Null model constructions were repeated 1,000 times in order to test for statistic difference with the observed data. Phylogenetic relatedness was aggregated step-wisely, using a step of 0.01 for pairwise sequence distances and a step of 0.1 for phylogenetic distances. For each distance class, the number of co-occurring or co-excluding OTUs was accounted in the observed and random

networks and a non-parametric p-values was calculated as the amount of time the observed number of co-occurrence or co-exclusion was higher or lower than in the null models. Differences between the observed networks and the null models were considered significant if the p-values were ≤ 0.05. Results were summarized for each distance class into standardized effect size (SES), calculated following Gotelli & McCabe (2002). By convention, a SES is considered as strong if it is ≥ 2.

Results

Networks coverage

In order to test for a phylogenetic signal between co-occurring and co-excluding OTUs with different phylogenetic relatedness, co-occurrence and co-exclusion networks were related to pairwise sequence and phylogenetic distances: edges of connected OTUs in the networks were labeled with the phylogenetic relatedness distances and the number of edges in each distance class were compared to two null models. The marine protist networks consisted of 32 to 53 % of candidate OTUs, while terrestrial protist networks included only 6 to 12 % of candidate OTUs (**Table 1**). The network OTUs occurred in at least 32 % of marine surface, 37 % of marine DCM or 17 % of terrestrial samples. The terrestrial co-exclusion network included the lowest amount of candidate OTUs (6 %) and candidate edges (0.02 %) compared to all the other networks.

The occurrence patterns of network OTUs were slightly skewed toward OTUs occurring in the highest number of samples and thus in the highest number of geographical units, compared to candidate OTUs (**Figure S2**). Marine protist networks included mainly OTUs occurring in 6 to 8 sea and oceans, and most candidate OTUs occurring in only 4 to 5 of this geographical units were not included in the networks. Terrestrial protists networks included mostly OTUs occurring in 2 to 3 forests while candidate OTUs occurring in a single forest were largely absent from the networks.

The taxonomic coverage of network OTUs remain unchanged in marine datasets compared to candidate OTUs (**Figure S3**). OTUs of the two clades with the lowest abundance in the terrestrial

dataset, Dinophyta and Haptophyta, were not included in the networks as well as Chlorophyta OTUs in the co-occurrence network and MAST (Marine Stramenopiles, polyphiletic basal clade; Massana, Campo, Sieracki, Audic, & Logares, 2014) OTUs in the co-exclusion network.

Phylogenetic signal in co-occurrences networks

Using null model 1 in which phylogenetic relatedness values were randomized along the edges of the networks, co-occurring OTUs from the marine datasets had positive SES that were significant and strong for low pairwise sequence distances <0.27 and phylogenetic distances <1.7, and OTUs from the terrestrial dataset had positive SES that were significant and strong for pairwise sequence distances <0.25 and phylogenetic distances <0.9 (**Figure 2**). Conversely, OTUs from the marine datasets had negative SES that were significant for intermediate and large pairwise sequence distances (0.27 to 0.5) and phylogenetic distances (2.1 to 4.3 and 6.3 to 9.5 for marine surface, 1.9 to 6.3 and 7.7 to 9.2 for marine DCM), and OTUs from the terrestrial dataset had negative SES for intermediate values that were significant in only four pairwise sequence distance classes (0.28 to 0.35) and seven phylogenetic distance classes (1.1 to 2.3). Interestingly, co-occurrence in Neotropical soils showed significant positive SES for OTUs pairs with large dissimilarities at one pairwise sequence and four phylogenetic distance classes.

Similar co-occurrence results to null model 1 were observed when using null model 2, in which the edges were randomized in the networks (**Figure S4**). Co-occurring OTUs from the marine datasets had positive SES that were significant and strong for pairwise sequence distances <0.23 and phylogenetic distances <1, and OTUs from the terrestrial dataset had positive SES that were significant and strong for pairwise sequence distances <0.04 and phylogenetic distances <0.9. And conversely, OTUs from the marine and terrestrial datasets had negative SES that were significant for intermediate pairwise sequence distances (>0.23) and phylogenetic distances (>1.1 to 6.7).

These results using the two null models mean that pairs of OTUs that are closely related phylogenetically co-occurred more often than expected by chance in the marine and terrestrial protistan communities, phylogenetically distant OTUs predominantly co-occured less often than expected by chance, and some phylogenetically far OTUs co-occurred more often than expected by chance. Additionally, for co-occurrences, using either pairwise sequence distances or phylogenetic distances in these comparisons results in similar SES values.

Phylogenetic signal in co-exclusion networks

Using null model 1, co-excluding OTUs from the marine datasets had negative SES that were significant and strong for low pairwise sequence distances <0.23 and phylogenetic distances <1.1 in surface and <1.4 in DCM waters (**Figure 2**). Conversely, OTUs from the marine datasets had positive SES that were significant for intermediate pairwise sequence distances (surface: 0.24 to 0.33; DCM: 0.25 to 0.42) and phylogenetic distances (surface: 1.3 to 3; DCM: 3 to 3.4), while at higher distance classes a mix of significant positive and negative SES were retrieved. In the terrestrial dataset, however, no significant SES were observed except for the pairwise distance class between 0.29 an 0.3 and three phylogenetic distance classes between 1.7 and 2.3 with significant positive SES and pairwise distances between 0.23 and 0.24 and phylogenetic distances between 1.2 and 1.3 with a significant negative SES each.

Similar co-exclusion results to null model 1 were also observed when using null model 2 respectively (**Figure S4**). Co-excluding OTUs from the marine datasets had negative SES that were significant and strong for pairwise sequence distances <0.23 in surface and <0.16 in DCM waters, and phylogenetic distances <1.2 in surface <0.8 in DCM waters. No significant SES were observed in the terrestrial dataset except for the pairwise distance class between 0.29 an 0.3, three phylogenetic distance classes between 1.7 and 2.3 with significant positive SES and phylogenetic distances between 0 and 0.1 with a significant negative SES.

These results using the two null models mean that pairs of OTUs that are closely related phylogenetically co-excluded less often than expected by chance, and phylogenetically distant OTUs co-excluded more often than expected by chance, in the marine protistan communities. In the terrestrial protistan communities, though, there was an independence between phylogenetic relatedness and co-exclusion. Additionally, for co-exclusions, as in the co-occurrences, using either pairwise sequence distances or phylogenetic distances in these comparisons results in similar SES values.

Synchrony and convergence in co-occurrence and co-exclusion patterns

In all datasets and for most distance classes, positive SES in co-occurrence networks were reflected by negative SES in co-exclusion networks and conversely. However, the negative SES in co-exclusion networks for phylogenetically close OTUs were comparatively much lower or non-significant than the positive SES in the co-occurrence networks. These patterns are confirmed by the edge sampling along distance classes (**Figure S5**), with co-occurrence networks sampling most of candidate edges in low pairwise sequence and phylogenetic distances values, while co-exclusion networks lack of edges in those low distance values. It implied higher sampling of edges between OTUs from same genera in the marine datasets or from same species in the terrestrial dataset for co-occurrence networks (**Figure S6**).

For some distance classes there was, at the same time, significant positive or negative SES in both co-occurrence and co-exclusion networks (**Figure 2** and **S4**, shaded areas). This was particularly obvious for the marine surface dataset with null model 1 for which a SES inversion zone with positive SES in both co-occurrence and co-exclusion networks was observed over large ranges of pairwise sequence (0.24-0.27) and phylogenetic (1.3-1.7) distance classes (**Figure 2**). In the inversion zone, more than 80% of the co-occurrences and co-exclusions in the marine surface dataset were between taxa of different kingdoms and the distribution of edges among shared taxonomic levels did not differed significantly from the candidate edges in these same ranges

(**Figure 3** and **S7**). A closer look at the taxonomic groups connected by co-occurrences and co-exclusions in the inversion zone revealed important shifts in proportion of edges compared to all candidate edges (**Figure S8**). Ciliophora were under-represented in both of these co-occurrence and co-exclusion sub-networks compared to all candidate edges as well as Apicomplexa, Bacillariophyta (diatoms), Dinophyta and Radiolaria in the co-occurrence sub-netwroks, while there were increase for almost all other pairs of clades in both sub-networks, in particular for Haptophyta in the co-occurrence sub-network and for Bacillariophyta vs. Dinophyta and Haptophyta in the co-exclusion sub-network (**Figure 4**). Interestingly, there were simultaneous excess of intra-clade co-occurrences and co-exclusions for Haptophyta, MAST and Telonemia and simultaneous lack of intra-class co-occurrences and co-exclusions for Ciliophora, Dinophyta and Radiolaria. The amount of changes was particularly important when comparing to the same sub-networks in the 0.24-027 pairwise sequence distance range of the marine DCM dataset (**Figure S9**). Edges involved less pairs of clades, and the lowest range of fold changes showed a much less divergent sampling of all potential edges than in the marine surface dataset, so that no inversion zone was visible for the marine DCM dataset.

Discussion

We assessed the non-random phylogenetic relatedness of co-occurring and co-excluding OTUs in two of the largest environmental sequencing datasets of marine and terrestrial protists. By decomposing assembly patterns in phylogenetic relatedness classes and by comparing observed results to two null models, we could show that phylogenetic close OTUs co-occurred more often than expected by chance and that co-occurring OTUs are phylogenetically closer than expected by chance in both environments. The opposite trend was observed for OTUs with intermediate phylogenetic distances, which co-occurred less often than expected by chance. These co-occurrence results tend to support the preponderant effect of environmental filtering under the assumption of phylogenetic niche conservatism. These results could also be explained by the dispersal limitation

of recently diverging taxa, which was demonstrated for the dominant protistan taxa in terrestrial dataset used here (Lentendu et al., 2018) or in marine ciliates (Azovsky, Chertoprud, Garlitska, Mazei, & Tikhonenkov, 2020).

Phylogenetic close OTUs were found to co-exclude less often than expected by chance while OTUs with intermediate phylogenetic distances co-excluded more often than expected by chance in the marine environments, in opposition to the co-occurrence patterns. There was, however, no clear limit between close and intermediate phylogenetic distances so that some distances classes displayed significant excess of both co-occurrences and co-exclusions in this transition zone in the marine surface dataset. In the terrestrial environment, however, co-exclusion was almost independent from phylogenetic relatedness. Under the assumption of phylogenetic niche conservatism, these co-exclusion patterns would also reflect the effect of environmental filtering in both marine surface and DCM waters, while neither environmental filtering nor competitive exclusion appeared to impact the distribution of protists in Neotropical soils. One explanation to this discrepancy would be the relatively higher level of homogenization and increased dispersal potential in the marine waters, which allows protists to more easily reach a suitable habitat, while the larger amount of soil protist microhabitats (M. S. Adl & Gupta, 2006) and the high local diversity in the Neotropics (Mahé et al., 2017) should blur the impact of environmental filtering and limit potential competitors to come into contact. Simultaneous excess of co-occurrences and co-exclusions in Haptophyta and Telonemia in the SES inversion zone could reflect simultaneous effect of environmental filtering and competitive exclusion. While the “paradox of the plankton” and its resolution based on the theory of chaos support the co-occurrence of functionally similar plankton (Huisman & Weissing, 1999; Hutchinson, 1961), here we show that indeed phylogenetic related plankton co-occur but could simultaneously co-exclude themselves more than expected by chance at the marine surface. Other large-scale processes affect the assembly patterns of marine protist like the mean annual temperature responsible of the latitudinal diversity gradient or the

372 sunlight exposure and currents responsible of the depth stratification in the water column (Giner et
373 al., 2020; Ibarbalz et al., 2019). However, geographical structures, natural fluctuations and absence
374 of equilibrium state in marine plankton communities are not enough to avoid exclusion among
375 related organisms, as observed here, and would refute the existence of any plankton paradox under
376 phylogenetic niche conservatism.

377 There are three novel aspects to this study. The first novel aspect was the use of null models
378 to test the significance of phylogenetic relatedness structures in co-occurrence and co-exclusion
379 networks. So far, only the relation between co-occurring/co-excluding protistan OTUs and their
380 putative function or the change in network topology among habitats were tested in marine (Guidi et
381 al., 2016; Lima-Mendez et al., 2015; Milici et al., 2016; Steele et al., 2011), freshwater (Debroas et
382 al., 2017; Posch et al., 2015) and terrestrial environments (Lentendu et al., 2014; Ma et al., 2016;
383 Xiong et al., 2017). In a network-based study on human microbiome combining analyses of
384 phylogenetic relatedness and co-occurrence/co-exclusion networks, it was shown that co-occurrence
385 between human bacterial OTUs were uniformly distributed among phylogenetic distances while co-
386 exclusions were mainly among phylogenetically distant OTUs (Faust et al., 2012). The lack of null
387 model and/or statistical test on these observations, however, did not allow to determine whether
388 biologic or random processes were responsible of the patterns. In a more recent study, global gut
389 microbiome co-occurrence networks were found to have significant higher phylogenetic
390 assortativity than in randomize networks overall (Tackmann, Matias Rodrigues, & von Mering,
391 2019), while size effect was not quantified at distinct distance classes and no interpretation was
392 provided on these observations. Our new approach has the potential to uncover inter-dependencies
393 between phylogenetic relatedness and co-occurrence and co-exclusion of any micro-organisms in
394 any environment.

395 The second novel aspect is that we showed that both phylogenetic distance and pairwise
396 sequence distance can both be used as measure of phylogenetic relatedness when applied to the

analysis of protistan community assembly patterns. Previous protist studies used phylogenetic relatedness of protist to assess phylogenetic diversity based macroecological and biogeographical patterns (Bates et al., 2013; Lentendu et al., 2018; Singer et al., 2018), while pairwise sequence distances were only used during bioinformatic procedure for sequence clustering or sequence similarity networks (Forster et al., 2019; Mahé et al., 2015).

The third novel aspect was the decomposition of the co-occurrence and co-exclusion signals along phylogenetic distance classes. By using traditional index of phylogenetic divergence (e.g., net relatedness index), only one type of divergence could be assessed per sample or pair of samples, that is either clustering or overdispersion. By using the co-occurrence and co-exclusion patterns over all samples, here we investigated the multiple signals hold by communities over increasing phylogenetic distances for the whole analyzed regions. In the marine surface environment, at the SES inversion zone, both phylogenetic clustering and overdispersion take place at the same time. Independent to the origin of these patterns (competition could also lead to phylogenetic clustering, Mayfield & Levine, 2010), phylogenetic relatedness play a strong role in determining the assembly of marine and terrestrial protists.

There are three major assumptions to this study. The first major assumption was that there is phylogenetic niche conservatism between the OTUs (Wiens & Donoghue, 2004). This assumption allowed us to infer that phylogenetic close OTUs share more niche space than phylogenetically distant OTUs. This assumption allows us to interpret the significant excess in co-occurrence among phylogenetically close as a signal of environmental filtering and the absence of significant effect size in co-exclusion among phylogenetically close OTUs as signal for lack of environmental filtering and competitive exclusion. However, the assumption that evolutionary close OTUs share the same niche may not be true and it could be misleading to deduce pattern from process (Gerhold, Cahill, Winter, Bartish, & Prinzing, 2015). In such large dataset, there is a multitude of niche evolution scenarios which lead to the current distribution of protist in marine waters and

Neotropical soils, and the apparent environmental filtering deducted here from the co-occurrence patterns could hide other processes at play which are not necessarily linked to phylogenetic niche conservatism. A modeling approach could also help to test for the reality of phylogenetic niche conservatism by protists (Münkemüller, Boucher, Thuiller, & Lavergne, 2015) but remains inapplicable for large datasets as analyzed here for which a large proportion of organisms are unknown (de Vargas et al., 2015; Mahé et al., 2017). Considering that current knowledge on traits and function is not sufficient to determine functional niche of most protists (Ramond et al., 2019), relating phylogeny to assembly patterns with the phylogenetic niche conservatism assumption is the most precise approach we can apply yet to find clues about large scale and whole community processes at play in protist community assembly.

The second major assumption to this study is that the OTUs are accurately estimating protistan species diversity. This assumption, which is made by most metabarcoding studies (Bik et al., 2012; Blaxter et al., 2005; Taberlet, Bonin, Zinger, & Coissac, 2018), allowed us to infer relative occurrences of each protist taxonomic unit among all samples of each datasets and allowed to infer the co-occurrence and co-exclusion networks. However, all clustering programs used to construct OTUs make assumptions about the best ways to handle the environmental sequencing data (Callahan et al., 2016; Caron & Hu, 2018; Mahé et al., 2015; Nebel, Pfabel, Stock, Dunthorn, & Stoeck, 2011; Rognes, Flouri, Nichols, Quince, & Mahé, 2016; Zhang, Kapli, Pavlidis, & Stamatakis, 2013) and these assumptions, along with the choice of molecular markers, may or may not lead to under- or over-estimations of species diversity. Here the reads were clustered into OTU with the program Swarm (Mahé et al., 2015; Mahé, Rognes, Quince, Vargas, & Dunthorn, 2014), which uses local clustering thresholds and a breaking phase to construct the OTUs. Swarm can partition the data into finer OTUs than programs that use global clustering thresholds, which may lead to over-splitting of species (Mahé et al., 2015); this over-splitting could potential explain the

high positive SES in the smallest pairwise sequence and phylogenetic distance classes of co-occurrence networks.

The third assumption is that phylogenetic relatedness is correctly assessed with the analyzed genes. This assumption allowed us to infer strong interrelationship between phylogenetic distance and co-occurrence and co-exclusion patterns. The short and hyper-variable V4 and V9 fragments only provide partial phylogenetic signal of the full SSU-rRNA locus (Dunthorn et al., 2014), which is in-turn, only an approximation of the real protistan phylogenetic relatedness as assessed with whole genome sequencing (Burki, 2014). Besides, the genetic distances estimated between these two hyper-variable regions can be the same or drastically different depending on which taxa are being compared (Dunthorn, Klier, Bunge, & Stoeck, 2012; Hu et al., 2015; Tragin, Zingone, & Vaulot, 2018). The congruent results for protistan co-occurrences and co-exclusions derived from both pairwise sequence distances and phylogenetic distances shows that both type of distances can be used to infer phylogenetic relatedness. The congruent co-occurrence results for both global marine and Neotropical soil protists shows that both V4 and V9 fragments could deliver similar phylogenetic related assembly structure so that could be equally applied for large scale datasets.

By demonstrating the strong phylogenetic signals in co-occurrence and co-exclusion patterns of protists, we showed that global and regional assembly mechanisms are directly related to phylogenetic relatedness and are dominated by environmental filtering. We could not conclude that the simultaneous excess of co-occurrence and co-exclusion of phylogenetic related OTUs in the SES inversion zone of the marine surface communities is the result of intra-clade competitive exclusion, but we could only suspect it. Indeed, multiple other processes could lead to such pattern, like facilitation of phylogenetically distant species (Cahill et al., 2008; Gerhold et al., 2015; Kraft, Cornwell, Webb, & Ackerly, 2007). The co-exclusion discrepancy between marine and terrestrial protists highlights the difference in mechanisms involved in community assembly between these two environments. The novel network-phylogeny approach presented in this study have potential to

unravel phylogenetic-driven assembly patterns in large scale datasets for which little is known about the taxonomy and function of the target organisms in other environments. The interplay between phylogeny and co-occurrence/co-exclusion networks remain to be disclosed in other microbial taxonomic groups, like Bacteria and Fungi, and among functional groups, like autotrophs, heterotrophs and associated microbes.

References

Adl, M. S., & Gupta, V. S. (2006). Protists in soil ecology and forest nutrient cycling. *Canadian Journal of Forest Research*, *36*(7), 1805–1817. doi: 10.1139/x06-056

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., … Zhang, Q. (2019). Revisions to the classification, nomenclature, and diversity of eukaryotes. *Journal of Eukaryotic Microbiology*, *66*(1), 4–119. doi: 10.1111/jeu.12691

Azovsky, A. I., Chertoprud, E. S., Garlitska, L. A., Mazei, Y. A., & Tikhonenkov, D. V. (2020). Does size really matter in biogeography? Patterns and drivers of global distribution of marine micro- and meiofauna. *Journal of Biogeography*, *00*, 1–13. doi: 10.1111/jbi.13771

Bates, S. T., Clemente, J. C., Flores, G. E., Walters, W. A., Parfrey, L. W., Knight, R., & Fierer, N. (2013). Global biogeography of highly diverse protistan communities in soil. *The ISME Journal*, *7*(3), 652–659. doi: 10.1038/ismej.2012.147

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach. *Journal of the Royal Statistical Society Series B-Methodological*, *57*(1), 289–300. doi: 10.2307/2346101

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., & Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, *27*(4), 233–243. doi: 10.1016/j.tree.2011.11.010

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1935–1943. doi: 10.1098/rstb.2005.1725

Boenigk, J., Wodniok, S., Bock, C., Beisser, D., Hempel, C., Grossmann, L., … Jensen, M. (2018). Geographic distance and mountain ranges structure freshwater protist communities on a European scale. *Metabarcoding and Metagenomics*, *2*, e21519. doi: 10.3897/mbmg.2.21519

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harbor Perspectives in Biology*, *6*(5), a016147. doi: 10.1101/cshperspect.a016147

Burns, J. H., & Strauss, S. Y. (2011). More closely related species are more ecologically similar in an experimental test. *Proceedings of the National Academy of Sciences*, *108*(13), 5302–5307. doi: 10.1073/pnas.1013003108

506 Cahill, J. F., Kembel, S. W., Lamb, E. G., & Keddy, P. A. (2008). Does phylogenetic relatedness
507 influence the strength of competition among vascular plants? *Perspectives in Plant Ecology,*
508 *Evolution and Systematics*, *10*(1), 41–50. doi: 10.1016/j.ppees.2007.10.001

509 Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P.
510 (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature*
511 *Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869

512 Caron, D. A., & Hu, S. K. (2018). Are we overestimating protistan diversity in Nature? *Trends in*
513 *Microbiology*. doi: 10.1016/j.tim.2018.10.009

514 Cavender-Bares, J., Kozak, K. H., Fine, P. V. A., & Kembel, S. W. (2009). The merging of
515 community ecology and phylogenetic biology. *Ecology Letters*, *12*(7), 693–715. doi: 10.1111/
516 j.1461-0248.2009.01314.x

517 Chow, C.-E. T., Kim, D. Y., Sachdeva, R., Caron, D. A., & Fuhrman, J. A. (2014). Top-down
518 controls on bacterial community structure: microbial network analysis of bacteria, T4-like
519 viruses and protists. *The ISME Journal*, *8*(4), 816–829. doi: 10.1038/ismej.2013.199

520 Chung, F., & Lu, L. (2002). Connected components in random graphs with given expected degree
521 sequences. *Annals of Combinatorics*, *6*(2), 125–145. doi: 10.1007/PL00012580

522 Connor, N., Barberán, A., & Clauset, A. (2017). Using null models to infer microbial co-occurrence
523 networks. *PLOS ONE*, *12*(5), e0176751. doi: 10.1371/journal.pone.0176751

524 Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research.
525 *InterJournal Complex Systems*, (1695). Retrieved from http://igraph.sourceforge.net/

526 Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of*
527 *favoured races in the struggle for life.* London: John Murray.

528 de Menezes, A. B., Prendergast-Miller, M. T., Richardson, A. E., Toscas, P., Farrell, M.,
529 Macdonald, L. M., … Thrall, P. H. (2014). Network analysis reveals that bacteria and fungi
530 form modules that correlate independently with soil parameters. *Environmental Microbiology*,
531 n/a-n/a. doi: 10.1111/1462-2920.12559

532 de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., … Velayoudon, D. (2015).
533 Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605. doi: 10.1126/
534 science.1261605

535 Debroas, D., Domaizon, I., Humbert, J.-F., Jardillier, L., Lepère, C., Oudart, A., & Taïb, N. (2017).
536 Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available
537 metabarcoding data. *FEMS Microbiology Ecology*, *93*(4). doi: 10.1093/femsec/fix023

538 del Campo, J., Mallo, D., Massana, R., de Vargas, C., Richards, T. A., & Ruiz Trillo, I. (2015).
539 Diversity and distribution of unicellular opisthokonts along the European coast analysed using
540 high-throughput sequencing. *Environmental Microbiology*, *17*(9), 3195–3207. doi:
541 10.1111/1462-2920.12759

542 Dunthorn, M., Klier, J., Bunge, J., & Stoeck, T. (2012). Comparing the hyper-variable V4 and V9
543 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *Journal*
544 *of Eukaryotic Microbiology*, *59*(2), 185–187. doi: 10.1111/j.1550-7408.2011.00602.x

Dunthorn, M., Otto, J., Berger, S. A., Stamatakis, A., Mahé, F., Romac, S., … Stoeck, T. (2014). Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Molecular Biology and Evolution*, *31*(4), 993–1009. doi: 10.1093/molbev/msu055

Faust, K., Sathirapongsasuti, J. F., Izard, J., Segata, N., Gevers, D., Raes, J., & Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol*, *8*(7), e1002606. doi: 10.1371/journal.pcbi.1002606

Forster, D., Lentendu, G., Filker, S., Dubois, E., Wilding, T. A., & Stoeck, T. (2019). Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environmental Microbiology*, *21*(11), 4109–4124. doi: 10.1111/1462-2920.14764

Gause, G. F. (1934). *The struggle for existence*. Baltimore: Williams and Wilkins.

Gerhold, P., Cahill, J. F., Winter, M., Bartish, I. V., & Prinzing, A. (2015). Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). *Functional Ecology*, *29*(5), 600–614. doi: 10.1111/1365-2435.12425

Giner, C. R., Pernice, M. C., Balagué, V., Duarte, C. M., Gasol, J. M., Logares, R., & Massana, R. (2020). Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *The ISME Journal*, *14*(2), 437–449. doi: 10.1038/s41396-019-0506-9

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, *8*. doi: 10.3389/fmicb.2017.02224

Gotelli, N. J., & McCabe, D. J. (2002). Species co-occurrence: A meta-analysis of J. M. Diamond's assembly rules model. *Ecology*, *83*(8), 2091–2096.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., … Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*, 465.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. doi: 10.1093/nar/gks1160

Hardy, O. J. (2008). Testing the spatial phylogenetic structure of local communities: statistical performances of different null models and test statistics on a locally neutral community. *Journal of Ecology*, *96*(5), 914–926. doi: 10.1111/j.1365-2745.2008.01421.x

Horner-Devine, M. C., & Bohannan, B. J. M. (2006). Phylogenetic clustering and overdispersion in bacterial communities. *Ecology*, *87*(sp7), S100–S108. doi: 10.1890/0012-9658(2006)87[100:PCAOIB]2.0.CO;2

Hu, S. K., Liu, Z., Lie, A. A. Y., Countway, P. D., Kim, D. Y., Jones, A. C., … Caron, D. A. (2015). Estimating protistan diversity using high-throughput sequencing. *Journal of Eukaryotic Microbiology*, *62*(5), 688–693. doi: 10.1111/jeu.12217

Huisman, J., & Weissing, F. J. (1999). Biodiversity of plankton by species oscillations and chaos. *Nature*, *402*(6760), 407–410. doi: 10.1038/46540

584 Humboldt, A. V., & Bonpland, A. (1805). *Essai sur la géographie des plantes*. Retrieved from
585 https://www.scribd.com/document/288843692/Essai-Sur-La-Geographie-Des-Plantes-
586 Humboldt-Bonpland-1805

587 Hutchinson, G. E. (1961). The Paradox of the Plankton. *The American Naturalist*, *95*(882), 137–
588 145. doi: 10.1086/282171

589 Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., … Zinger, L.
590 (2019). Global trends in marine plankton diversity across kingdoms of life. *Cell*, *179*(5),
591 1084-1097.e21. doi: 10.1016/j.cell.2019.10.008

592 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7:
593 improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–
594 780. doi: 10.1093/molbev/mst010

595 Kembel, S. W., & Hubbell, S. P. (2006). The phylogenetic structure of a Neotropical forest tree
596 community. *Ecology*, *87*(sp7), S86–S99. doi: 10.1890/0012-
597 9658(2006)87[86:TPSOAN]2.0.CO;2

598 Khomich, M., Kauserud, H., Logares, R., Rasconi, S., & Andersen, T. (2017). Planktonic protistan
599 communities in lakes along a large-scale environmental gradient. *FEMS Microbiology*
600 *Ecology*, *93*(4). doi: 10.1093/femsec/fiw231

601 Kraft, N. J. B., Adler, P. B., Godoy, O., James, E. C., Fuller, S., & Levine, J. M. (2015).
602 Community assembly, coexistence and the environmental filtering metaphor. *Functional*
603 *Ecology*, *29*(5), 592–599. doi: 10.1111/1365-2435.12345

604 Kraft, N. J. B., Cornwell, W. K., Webb, C. O., & Ackerly, D. D. (2007). Trait evolution,
605 community assembly, and the phylogenetic structure of ecological communities. *The*
606 *American Naturalist*, *170*(2), 271–283. doi: 10.1086/519400

607 Lamb, E. G., & Cahill Jr., J. F. (2008). When competition does not matter: grassland diversity and
608 community composition. *The American Naturalist*, *171*(6), 777–787. doi: 10.1086/587528

609 Lauber, C. L., Strickland, M. S., Bradford, M. A., & Fierer, N. (2008). The influence of soil
610 properties on the structure of bacterial and fungal communities across land-use types. *Soil*
611 *Biology and Biochemistry*, *40*(9), 2407–2415. doi: 10.1016/j.soilbio.2008.05.021

612 Lennon, J. T., Aanderud, Z. T., Lehmkuhl, B. K., & Schoolmaster, D. R. (2012). Mapping the niche
613 space of soil microorganisms using taxonomy and traits. *Ecology*, *93*(8), 1867–1879. doi:
614 10.1890/11-1745.1

615 Lentendu, G., Mahé, F., Bass, D., Rueckert, S., Stoeck, T., & Dunthorn, M. (2018). Consistent
616 patterns of high alpha and low beta diversity in tropical parasitic and free-living protists.
617 *Molecular Ecology*, *27*(13), 2846–2857. doi: 10.1111/mec.14731

618 Lentendu, G., Wubet, T., Chatzinotas, A., Wilhelm, C., Buscot, F., & Schlegel, M. (2014). Effects
619 of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a
620 multiple barcoding approach. *Molecular Ecology*, *23*(13), 3341–3355. doi:
621 10.1111/mec.12819

622 Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., … Raes, J. (2015).
623 Determinants of community structure in the global plankton interactome. *Science*, *348*(6237),
624 1262073. doi: 10.1126/science.1262073

625 Ma, B., Wang, H., Dsouza, M., Lou, J., He, Y., Dai, Z., … Gilbert, J. A. (2016). Geographic
626 patterns of co-occurrence network topological features for soil microbiota at continental scale
627 in eastern China. *The ISME Journal*, *10*(8), 1891–1901. doi: 10.1038/ismej.2015.261

628 Macarthur, R., & Levins, R. (1967). The limiting similarity, convergence, and divergence of
629 coexisting species. *The American Naturalist*, *101*(921), 377–385. doi: 10.1086/282505

630 Mahé, F., de Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., … Dunthorn, M. (2017).
631 Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature*
632 *Ecology & Evolution*, *1*, 0091. doi: 10.1038/s41559-017-0091

633 Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2015). Swarm v2: highly-
634 scalable and high-resolution amplicon clustering. *PeerJ*, *3*, e1420. doi: 10.7717/peerj.1420

635 Mahé, F., Rognes, T., Quince, C., Vargas, C. de, & Dunthorn, M. (2014). Swarm: robust and fast
636 clustering method for amplicon-based studies. *PeerJ*, *2*, e593. doi: 10.7717/peerj.593

637 Manel, S., Couvreur, T. L. P., Munoz, F., Couteron, P., Hardy, O. J., & Sonké, B. (2014).
638 Characterizing the phylogenetic tree community structure of a protected tropical rain forest
639 area in Cameroon. *PLOS ONE*, *9*(6), e98920. doi: 10.1371/journal.pone.0098920

640 Martiny, A. C., Treseder, K., & Pusch, G. (2013). Phylogenetic conservatism of functional traits in
641 microorganisms. *The ISME Journal*, *7*(4), 830–838. doi: 10.1038/ismej.2012.160

642 Martiny, J. B. H., Jones, S. E., Lennon, J. T., & Martiny, A. C. (2015). Microbiomes in light of
643 traits: A phylogenetic perspective. *Science*, *350*(6261), aac9323. doi:
644 10.1126/science.aac9323

645 Massana, R., Campo, J. del, Sieracki, M. E., Audic, S., & Logares, R. (2014). Exploring the
646 uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within
647 stramenopiles. *The ISME Journal*, *8*(4), 854–866. doi: 10.1038/ismej.2013.204

648 Mayfield, M. M., & Levine, J. M. (2010). Opposing effects of competitive exclusion on the
649 phylogenetic structure of communities. *Ecology Letters*, *13*(9), 1085–1093. doi:
650 10.1111/j.1461-0248.2010.01509.x

651 McMurdie, P. J., & Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is
652 inadmissible. *PLOS Computational Biology*, *10*(4), e1003531. doi:
653 10.1371/journal.pcbi.1003531

654 Mercier, C., Boyer, F., Bonin, A., & Coissac, É. (2013). *SUMATRA and SUMACLUST: fast and*
655 *exact comparison and clustering of sequences*. Retrieved from
656 http://metabarcoding.org/sumatra

657 Milici, M., Deng, Z.-L., Tomasch, J., Decelle, J., Wos-Oxley, M. L., Wang, H., … Wagner-Döbler,
658 I. (2016). Co-occurrence analysis of microbial taxa in the Atlantic ocean reveals high
659 connectivity in the free-living bacterioplankton. *Frontiers in Microbiology*, *7*. doi:
660 10.3389/fmicb.2016.00649

Morriën, E., Hannula, S. E., Snoek, L. B., Helmsing, N. R., Zweers, H., de Hollander, M., … van der Putten, W. H. (2017). Soil networks become more connected and take up more carbon as nature restoration progresses. *Nature Communications*, *8*, 14349. doi: 10.1038/ncomms14349

Müller, J. P., Hauzy, C., & Hulot, F. D. (2012). Ingredients for protist coexistence: competition, endosymbiosis and a pinch of biochemical interactions. *Journal of Animal Ecology*, *81*(1), 222–232. doi: 10.1111/j.1365-2656.2011.01894.x

Münkemüller, T., Boucher, F. C., Thuiller, W., & Lavergne, S. (2015). Phylogenetic niche conservatism – common pitfalls and ways forward. *Functional Ecology*, *29*(5), 627–639. doi: 10.1111/1365-2435.12388

Nebel, M., Pfabel, C., Stock, A., Dunthorn, M., & Stoeck, T. (2011). Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. *Environmental Microbiology Reports*, *3*(2), 154–158. doi: 10.1111/j.1758-2229.2010.00200.x

Olff, H., & Ritchie, M. E. (1998). Effects of herbivores on grassland plant diversity. *Trends in Ecology & Evolution*, *13*(7), 261–265. doi: 10.1016/S0169-5347(98)01364-0

Philippot, L., Andersson, S. G. E., Battin, T. J., Prosser, J. I., Schimel, J. P., Whitman, W. B., & Hallin, S. (2010). The ecological coherence of high bacterial taxonomic ranks. *Nature Reviews Microbiology*, *8*(7), 523–529. doi: 10.1038/nrmicro2367

Posch, T., Eugster, B., Pomati, F., Pernthaler, J., Pitsch, G., & Eckert, E. M. (2015). Network of interactions between ciliates and phytoplankton during spring. *Frontiers in Microbiology*, *6*. doi: 10.3389/fmicb.2015.01289

R Core Team. (2017). R: a language and environment for statistical computing (Version 3.4.1). Retrieved from http://cran.r-project.org/

Ramond, P., Sourisseau, M., Simon, N., Romac, S., Schmitt, S., Rigaut Jalabert, F., … Siano, R. (2019). Coupling between taxonomic and functional diversity in protistan coastal communities. *Environmental Microbiology*, *21*(2), 730–749. doi: 10.1111/1462-2920.14537

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). *VSEARCH: a versatile open source tool for metagenomics* (No. e2409v1). Retrieved from PeerJ Preprints website: https://peerj.com/preprints/2409

Saleem, M., Fetzer, I., Dormann, C. F., Harms, H., & Chatzinotas, A. (2012). Predator richness increases the effect of prey diversity on prey yield. *Nature Communications*, *3*, 1305. doi: 10.1038/ncomms2287

Singer, D., Kosakyan, A., Seppey, C. V. W., Pillonel, A., Fernández, L. D., Fontaneto, D., … Lara, E. (2018). Environmental filtering and phylogenetic clustering correlate with the distribution patterns of cryptic protist species. *Ecology*, *0*(0). doi: 10.1002/ecy.2161

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Steele, J. A., Countway, P. D., Xia, L., Vigil, P. D., Beman, J. M., Kim, D. Y., … Schwalbach, M. S. (2011). Marine bacterial, archaeal and protistan association networks reveal ecological linkages. *The ISME Journal*, *5*(9), 1414–1425.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: for biodiversity research and monitoring*. Oxford, New York: Oxford University Press.

Tackmann, J., Matias Rodrigues, J. F., & von Mering, C. (2019). Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. *Cell Systems*, *9*(3), 286-296.e8. doi: 10.1016/j.cels.2019.08.002

Tedersoo, L., Bahram, M., Cajthaml, T., Põlme, S., Hiiesalu, I., Anslan, S., … Abarenkov, K. (2016). Tree diversity and species identity effects on soil fungi, protists and animals are context dependent. *The ISME Journal*, *10*(2), 346–362. doi: 10.1038/ismej.2015.116

Tragin, M., Zingone, A., & Vaulot, D. (2018). Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. *Environmental Microbiology*, *20*(2), 506–520. doi: 10.1111/1462-2920.13952

Tucker, C. M., Cadotte, M. W., Carvalho, S. B., Davies, T. J., Ferrier, S., Fritz, S. A., … Mazel, F. (2017). A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, *92*(2), 698–715. doi: 10.1111/brv.12252

Violle, C., Nemergut, D. R., Pu, Z., & Jiang, L. (2011). Phylogenetic limiting similarity and competitive exclusion. *Ecology Letters*, *14*(8), 782–787. doi: 10.1111/j.1461-0248.2011.01644.x

Weißbecker, C., Wubet, T., Lentendu, G., Kühn, P., Scholten, T., Bruelheide, H., & Buscot, F. (2018). Experimental evidence of functional group-dependent effects of tree diversity on soil Fungi in subtropical forests. *Frontiers in Microbiology*, *9*. doi: 10.3389/fmicb.2018.02312

Wetzel, C. E., Bicudo, D. de C., Ector, L., Lobo, E. A., Soininen, J., Landeiro, V. L., & Bini, L. M. (2012). Distance decay of similarity in Neotropical diatom communities. *PLOS ONE*, *7*(9), e45071. doi: 10.1371/journal.pone.0045071

Wiens, J. J., Ackerly, D. D., Allen, A. P., Anacker, B. L., Buckley, L. B., Cornell, H. V., … Stephens, P. R. (2010). Niche conservatism as an emerging principle in ecology and conservation biology. *Ecology Letters*, *13*(10), 1310–1324. doi: 10.1111/j.1461-0248.2010.01515.x

Wiens, J. J., & Donoghue, M. J. (2004). Historical biogeography, ecology and species richness. *Trends in Ecology & Evolution*, *19*(12), 639–644. doi: 10.1016/j.tree.2004.09.011

Xiong, W., Jousset, A., Guo, S., Karlsson, I., Zhao, Q., Wu, H., … Geisen, S. (2017). Soil protist communities form a dynamic hub in the soil microbiome. *The ISME Journal*, ismej2017171. doi: 10.1038/ismej.2017.171

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, *29*(22), 2869–2876. doi: 10.1093/bioinformatics/btt499

737 Zinger, L., Amaral-Zettler, L. A., Fuhrman, J. A., Horner-Devine, M. C., Huse, S. M., Welch, D. B.
738 M., … Ramette, A. (2011). Global patterns of bacterial beta-diversity in seafloor and seawater
739 ecosystems. *PLoS ONE*, *6*(9), e24570. doi: 10.1371/journal.pone.0024570

740

741 Author contributions

742 GL and MD conceived the ideas; GL conducted the analyses; GL and MD wrote the manuscript.

743 Acknowledgement

744 This work was supported by the Deutsche Forschungsgemeinschaft grants DU1319/5-1 to Micah
745 Dunthorn. The authors are grateful to the High Performance Computer Elwetritsch at the Technical
746 University of Kaiserslautern and the Centre for Computation of the Science Faculty of the
747 University of Neuchâtel for computing support.

748 Competing Interests

749 The authors declare that they have no conflict of interest.

Tables

Table 1 Network parameters

network	dataset	samples	candidate OTUs*	reads	Spearman's *rho* threshold	network OTUs §	network reads	candidate correlations $	significant correlations &	average network degree	average network path length
co-occurrence	marine surface	47	8274	1.1e+8	0.58	4351 (52.6 %)	8.2e+7	3.4e+7	49616 (0.14 %)	22.8	4.7
	marine DCM	32	10760	6.5e+7	0.68	3575 (33.2 %)	3.6e+7	5.8e+7	25306 (0.04 %)	14.2	6.2
	Neotropical soil	114	687	1.8e+7	0.45	83 (12.1 %)	5.2e+6	2.4e+5	373 (0.16 %)	9.0	2.2
co-exclusion	marine surface	47	8274	1.1e+8	-0.52	4265 (51.5 %)	8.1e+7	3.4e+7	29873 (0.09 %)	14.0	4.0
	marine DCM	32	10760	6.5e+7	-0.64	3478 (32.3 %)	3.5e+7	5.8e+7	13760 (0.02 %)	7.9	5.1
	Neotropical soil	114	687	1.8e+7	-0.24	41 (6 %)	4.8e+6	2.4e+5	54 (0.02 %)	2.6	3.4

* OTU of the original dataset occurring in at least 30 % of marine or 10 % of terrestrial samples

§ percentage of candidate OTUs in brackets

$ total number of potential edges between network OTUs

& number of edges in the network; percentage of potential edges in brackets

Figures

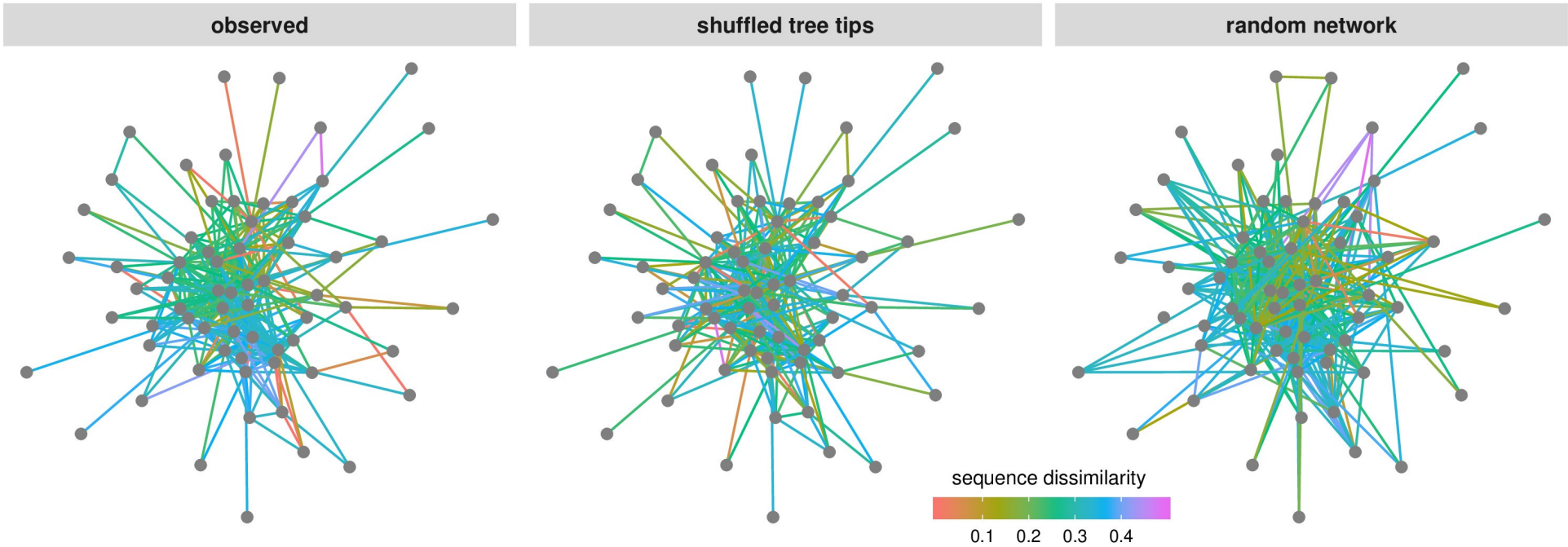
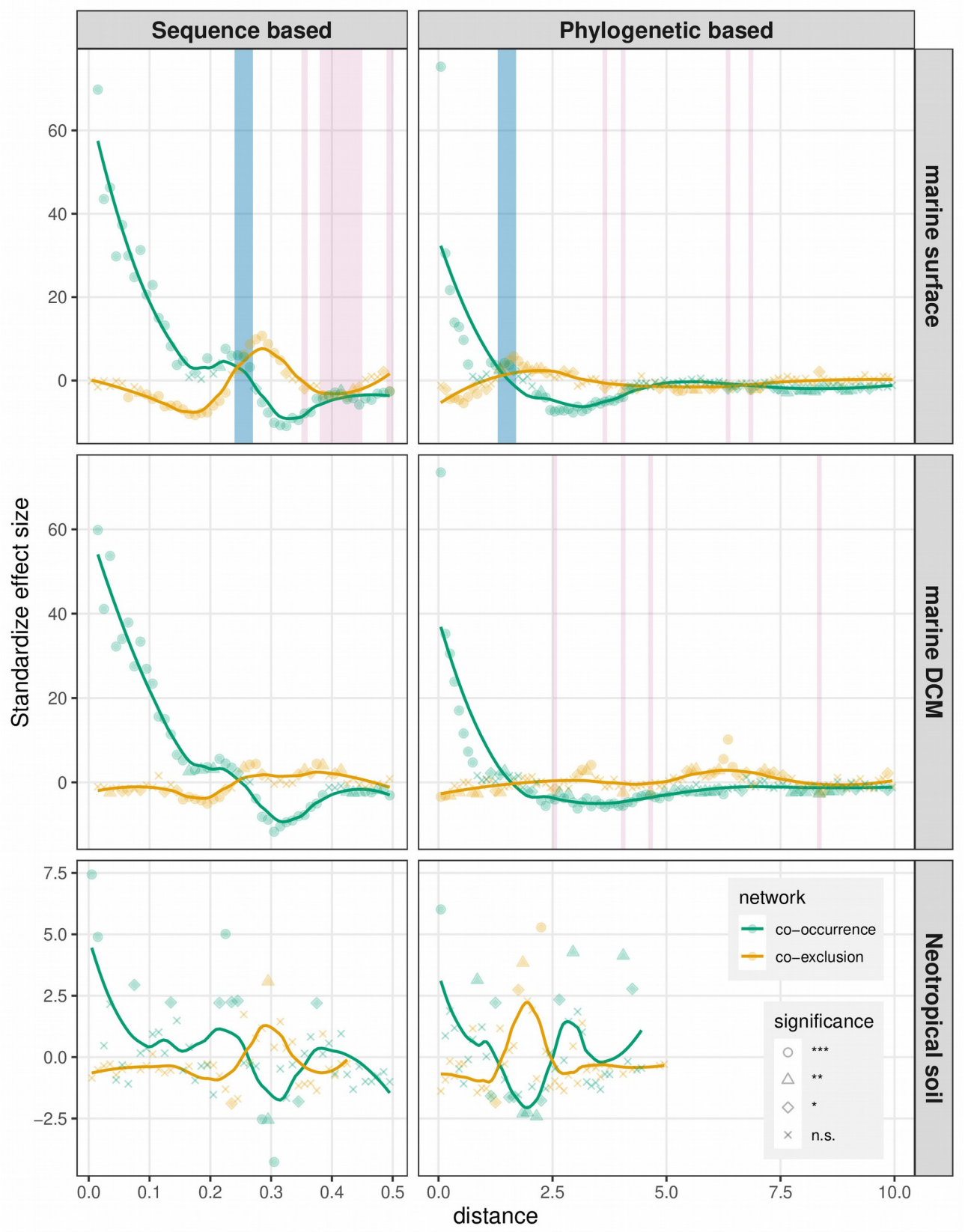
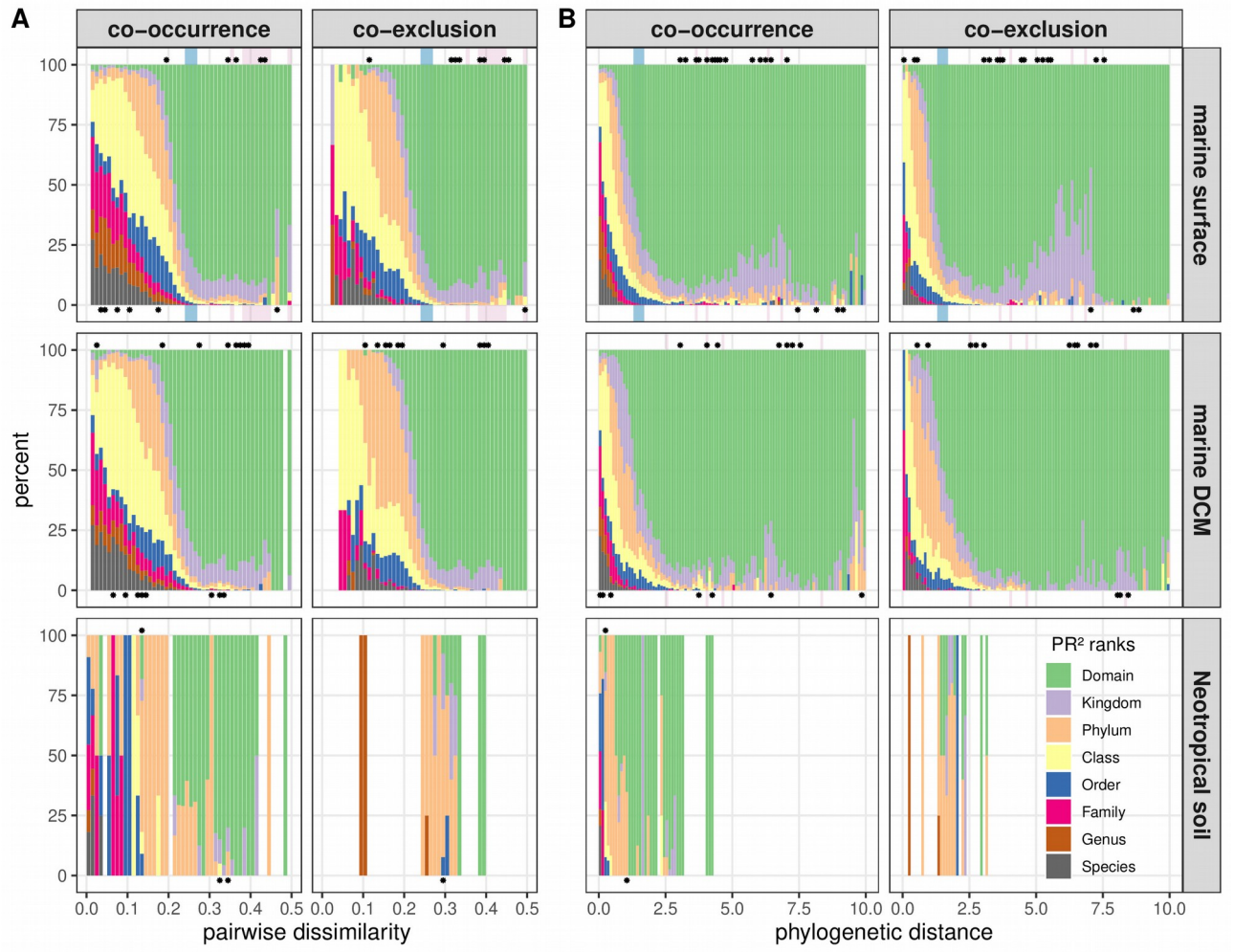


Figure 1 Null models effects on co-occurrence networks. Using the terrestrial protists co-occurrence network (observed) in which nodes are OTUs, edges are significant co-occurrences and edge colors are pairwise sequence dissimilarity. The first null model shuffle the pairwise sequence distance matrix (shuffled tree tips) while the second null model randomized the edges with a probability model (random network). The same approach was used for phylogenetic distance with phylogenetic tree tips shuffling in the first null model. The same computations were conducted on co-exclusion networks in which edges are significant co-exclusions.

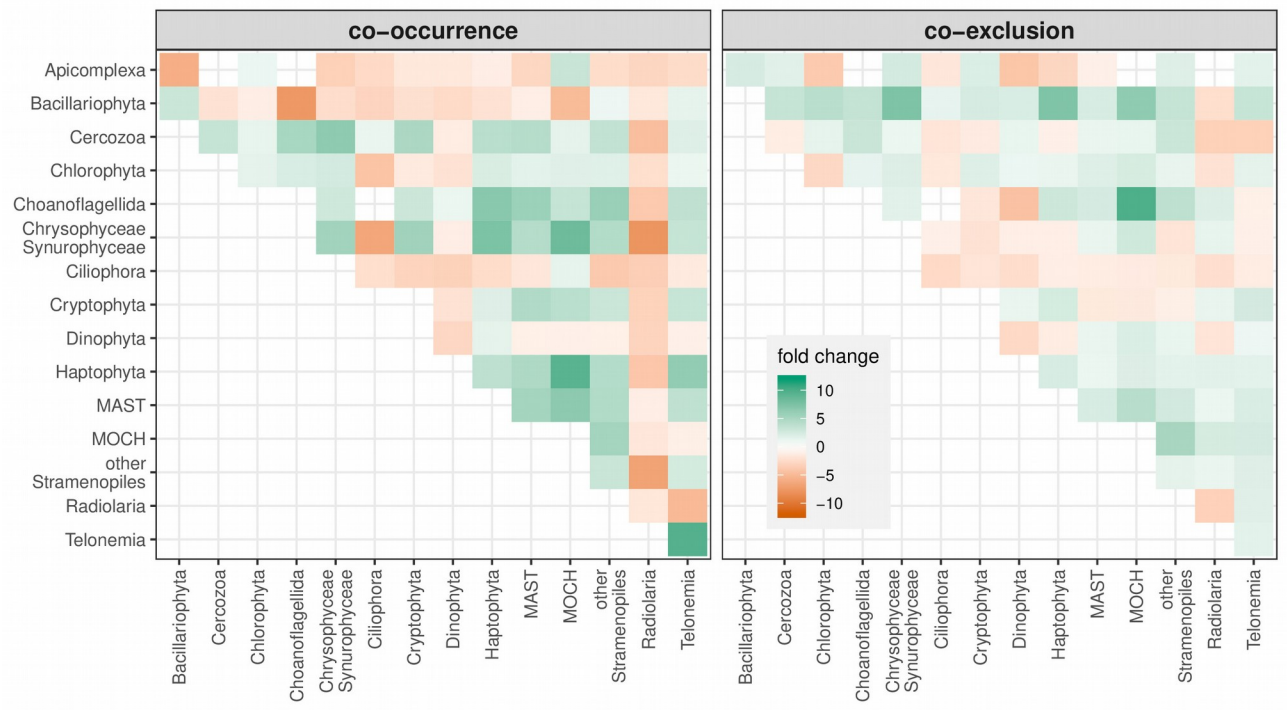


762 **Figure 2** Standardize effect sizes (SES) in co-occurrence and co-exclusion networks compared to
763 null models with shuffled phylogenetic tree tips (null model 1). SES were calculated separately for
764 stepwise increased pairwise sequence genetic distances and phylogenetic distances. The number of

OTU pairs connected by an edge in the observed networks was accounted for each distance class (from 0 to 0.5 with a 0.01 step for sequence based; from 0 to the maximum phylogenetic distance with a 0.1 step for phylogenetic based) and compared to the corresponding distance class reported from the randomized networks. Two-sided non-parametric p.values are inversely proportional to the amount of null models with a higher (for positive SES) or lower (for negative SES) amount of co-occurrence than in the observed network for each distance class. P.values below or equal to 0.05 were considered significant (* ≤ 0.05; ** ≤ 0.01; *** ≤ 0.001). Distance ranges highlighted in blue or red are for distances with excess (significant positive SES) or lack (significant negative SES) of edges in both co-occurrence and co-exclusion networks simultaneously.



774 **Figure 3** Distribution of taxonomic relationships between network connected OTUs for each
775 pairwise sequence distance (a) and phylogenetic distance (b) classes. Blue and red shaded areas in
776 the background are the distance classes with simultaneous positive or negative SES in both co-
777 occurrence and co-exclusion networks using null model 1, as in Figure 2. Stars at the bottom of the
778 bars indicate classes with significant deeper (toward species level) taxonomic ranks distribution
779 compared to all candidate edges (Figure S7), stars at the top of the bars indicate classes with
780 significant higher (toward domain level) taxonomic ranks distribution (Mann-Whitney test, $p<0.05$).



781 **Figure 4** Fold changes in proportion of edges connecting the main clades in the marine surface
782 dataset compared to all candidate edges in the pairwise sequence distance range of 0.24-0.27 (*i.e.*
783 the largest range of distance with simultaneous positive SES in co-occurrence and co-exclusion
784 networks when using the null model 1). The fold change color scale is identical to the one use for
785 the marine DCM dataset (Figure S9).

Supporting information

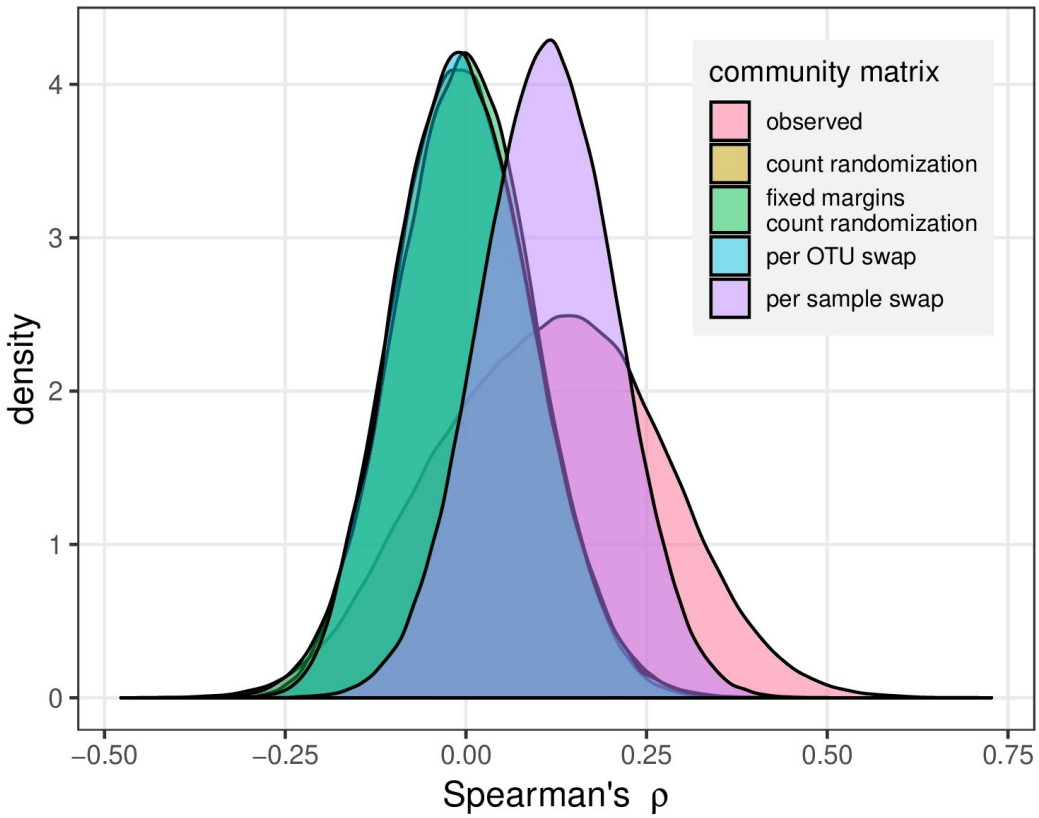


Figure S1 Distribution of Spearman's *rho* correlations among OTUs normalized relative abundance calculated using the observed Neotropical soil community matrix (red) and four randomization of it: all counts were randomly drawn over the community matrix without constrain (yellow), all counts were randomly drawn while keeping OTU and samples sum fixed (green), abundance values were randomly swap within each OTU (blue), abundance values were randomly swap within each sample (purple). Overlapping yellow, green and blue areas produced a dark blue area with median and mean Spearman's *rho* of zero. Observed and per sample swap matrices had median and mean Spearman's *rho* of 0.11. The observed matrix had a standard deviation of 0.15, while this value was 0.09 for all randomized matrices.

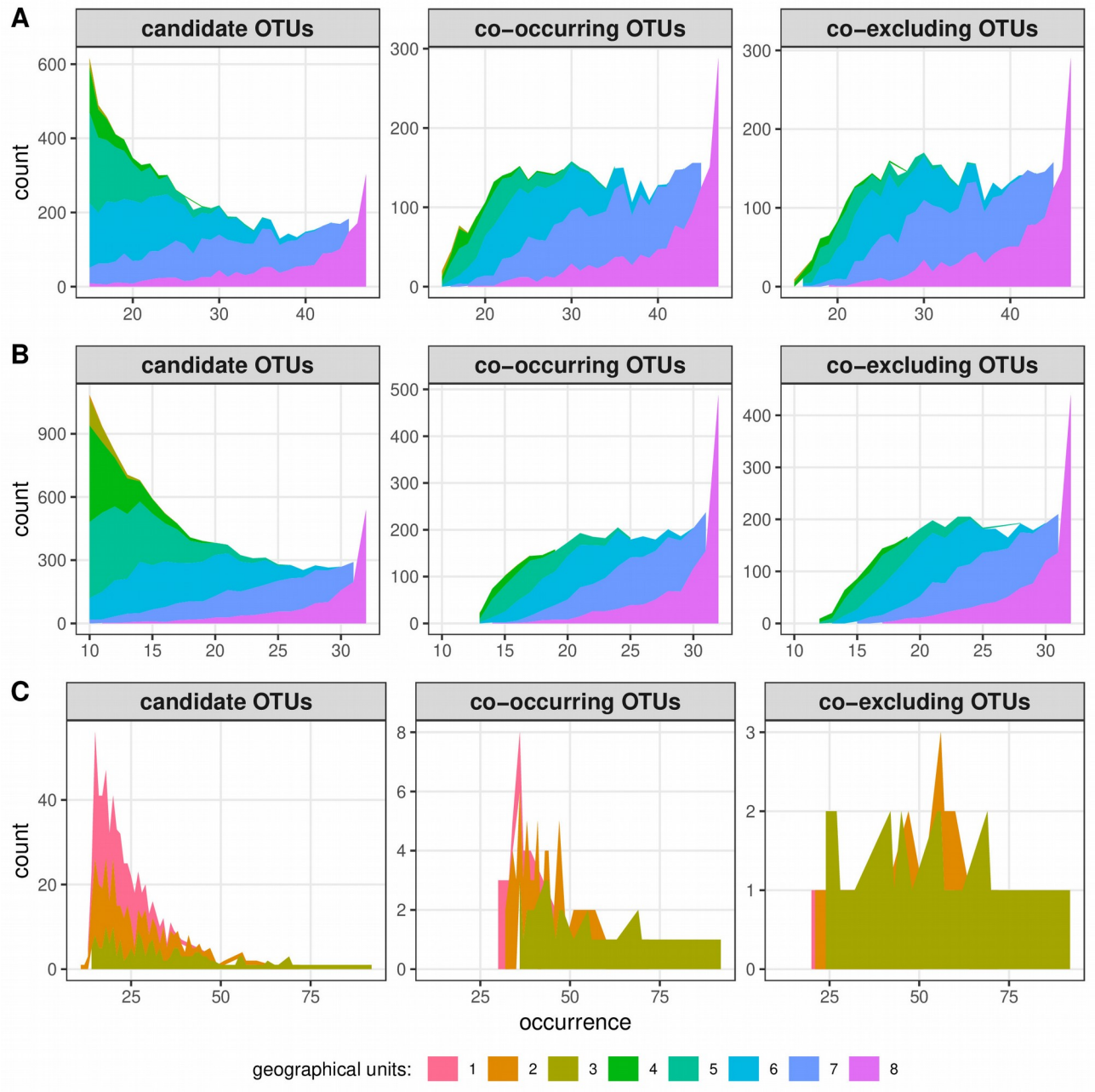


Figure S2 Distribution of candidate OTUs (OTUs occurring in at least 30 % of all samples for marine protists and in at least 10 % for terrestrial protists) and OTUs integrated into the co-occurrence or co-exclusion networks among the different geographical units: a. surface and b. DCM marine protist OTUs among eight different oceans and seas worldwide; c. Neotropical soil protist OTUs among three forests. Colored areas are for OTUs occurring in increasing amount of geographical units. Areas are stacked on each other (i.e. non-overlapping), so that the upper limit of the upper area is the cumulative amount of OTU occurring in the same number of samples.

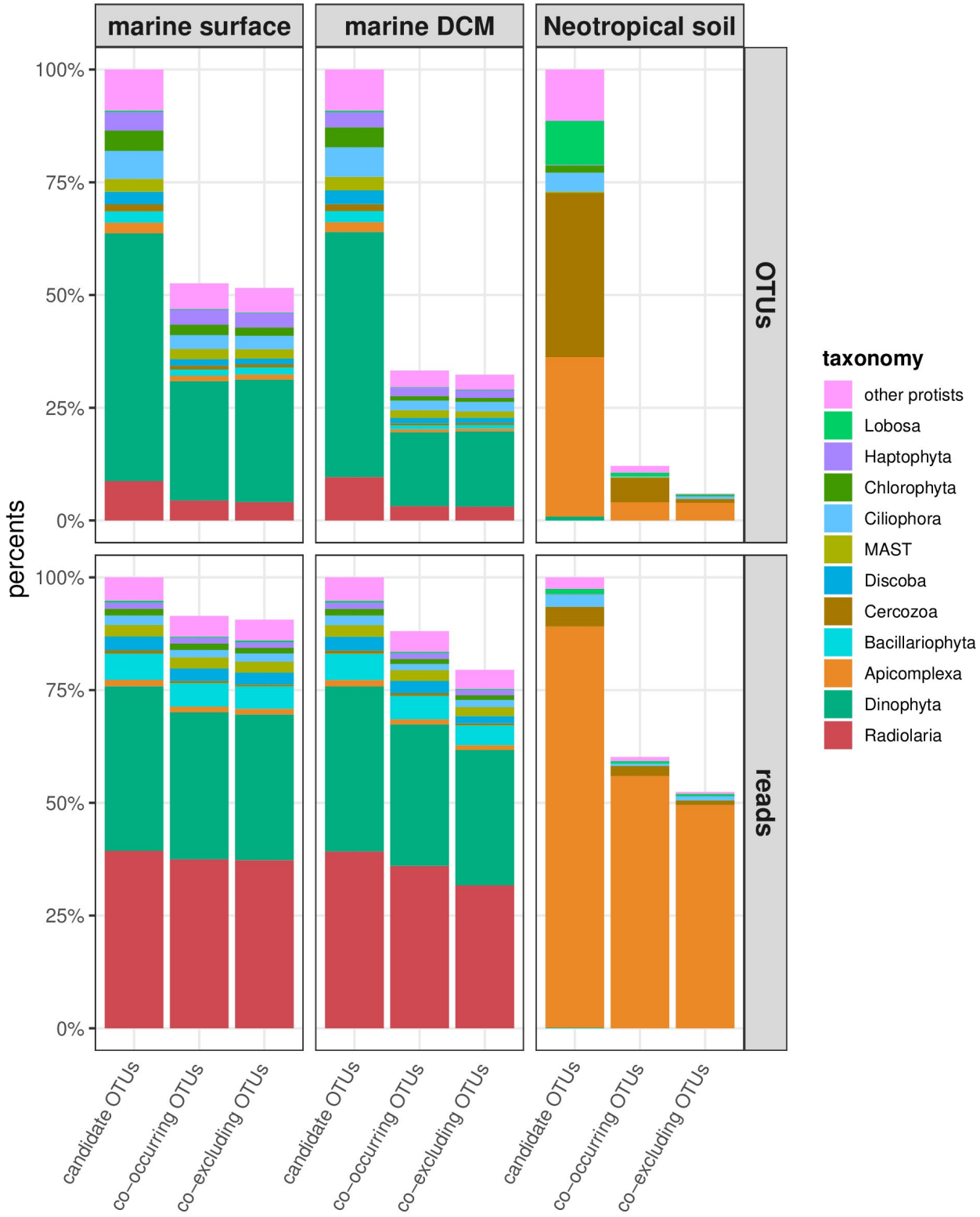


Figure S3 Taxonomy of candidate OTUs and OTUs integrated into the co-occurrence or co-exclusion networks for each dataset, expressed in term of OTUs percentages and log-ratio transformed relative abundance percentages.

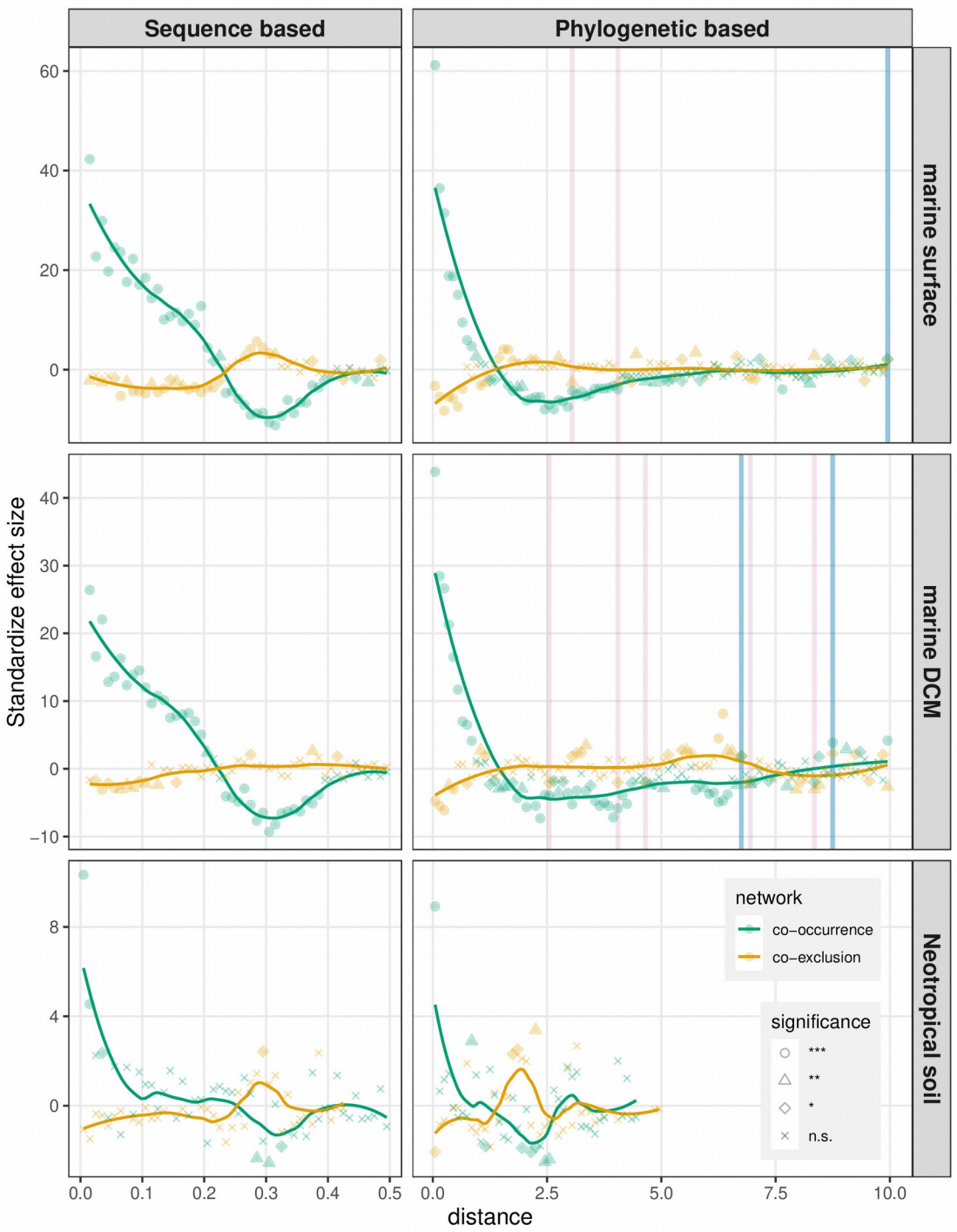
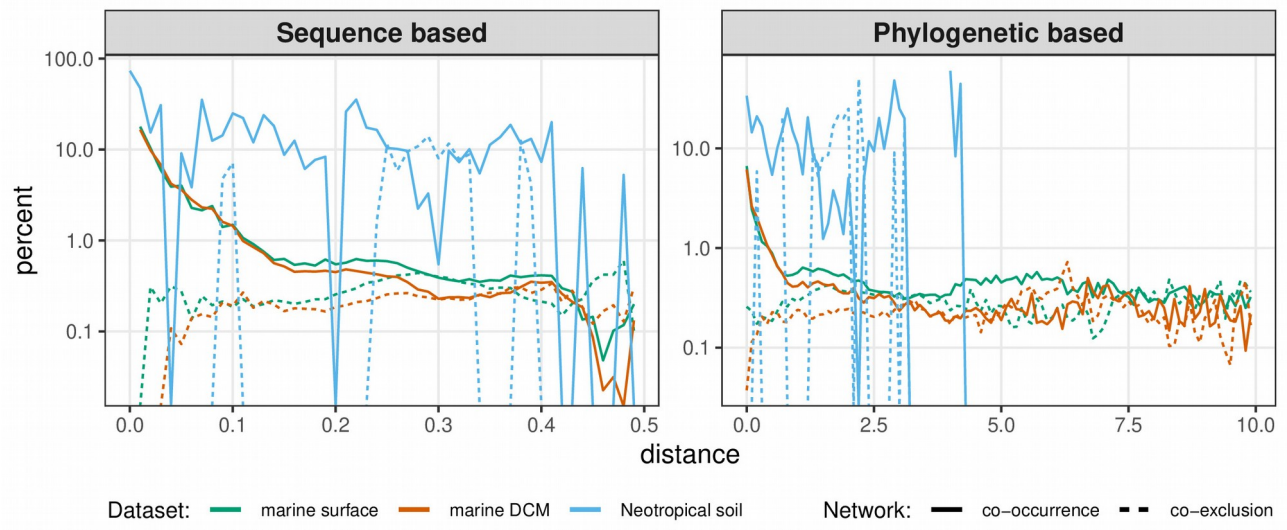
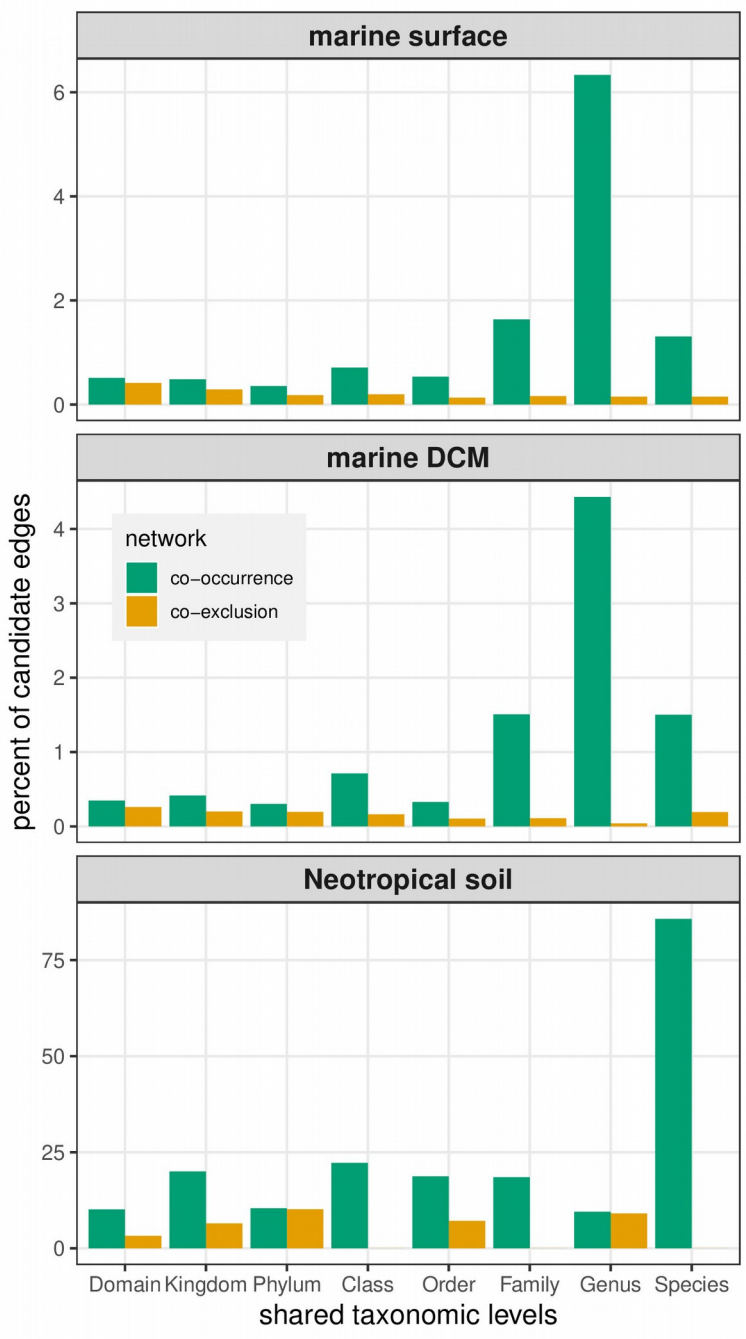


Figure S4 Standardize effect sizes (SES) in co-occurrence and co-exclusion networks compared to random networks with shuffled edges (null model 2). SES were calculated separately for stepwise increased pairwise sequence genetic distances and phylogenetic distances. The number of OTU pairs connected by an edge in the observed networks was accounted for each distance class (from 0 to 0.5 with a 0.01 step for sequence based; from 0 to the maximum phylogenetic distance with a 0.1 step for phylogenetic based) and compared to the corresponding distance class reported from the randomized networks. Two-sided non-parametric p.values are inversely proportional to the amount

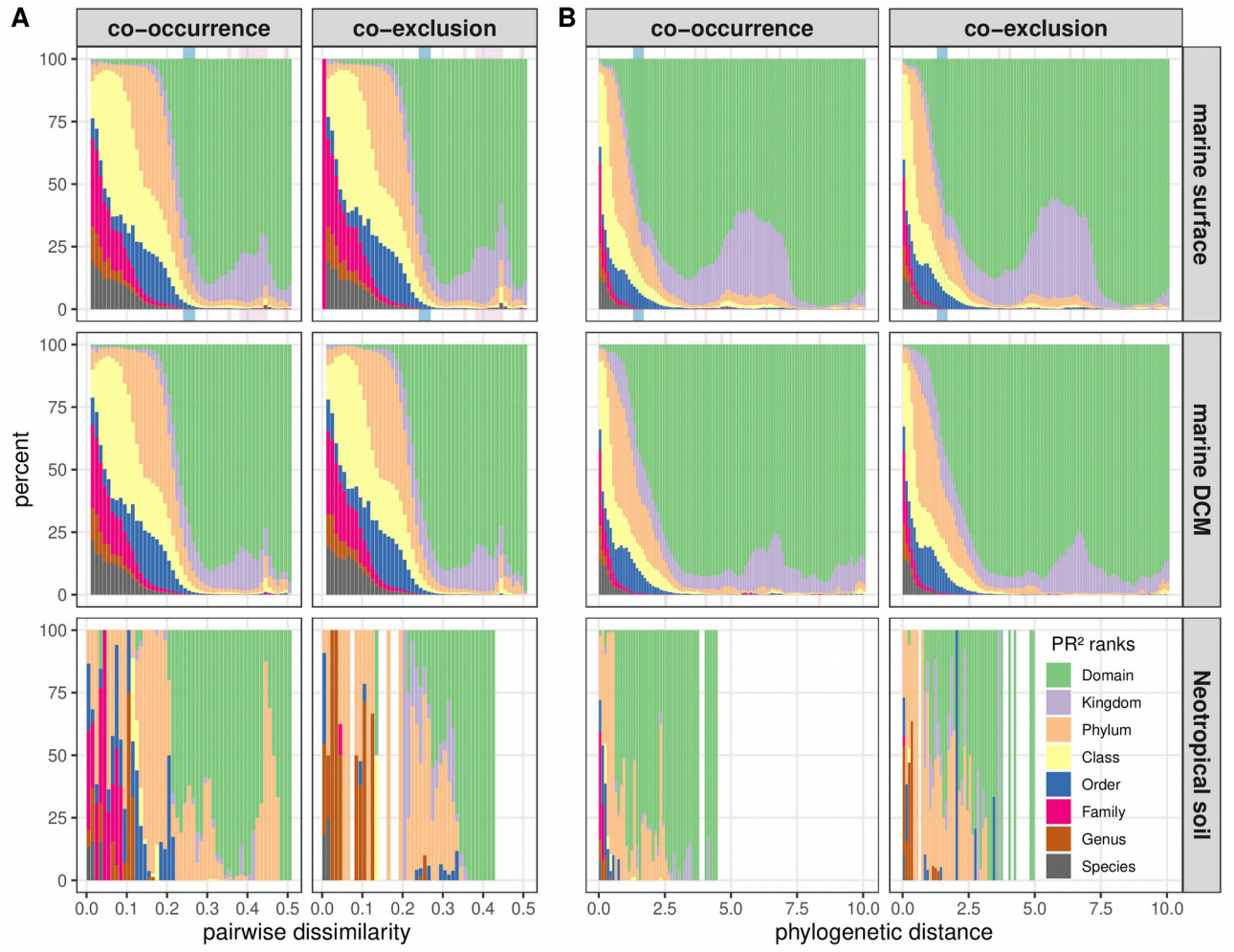
of null models with a higher (for positive SES) or lower (for negative SES) amount of co-occurrence than in the observed network for each distance class. P.values below or equal to 0.05 were considered significant (* ≤ 0.05; ** ≤ 0.01; *** ≤ 0.001). Distance ranges highlighted in blue or red are for distances with excess (significant positive SES) or lack (significant negative SES) of edges in both co-occurrence and co-exclusion networks simultaneously.



819 **Figure S5** Percent of total candidate edges in the observed networks arranged by distance classes.
820 Y-axis is square-root transformed to improve readability. The highest phylogenetic distance among
821 candidate edges for Neotropical soil is 4.5. No slope were drawn for distance classes not covered in
822 the observed networks.



824 **Figure S6** Percent of candidate edges sampled in the observed networks arranged by amount of
825 shared taxonomic levels between co-occurring or co-excluding OTUs.



826 **Figure S7** Distribution of taxonomic relationships between OTUs of all candidate edges for each
827 pairwise sequence distance (a) and phylogenetic distance (b) classes. Blue and red shaded areas in
828 the background are the distance classes with simultaneous positive or negative SES in both co-
829 occurrence and co-exclusion networks using null model 1, as in Figure 2.

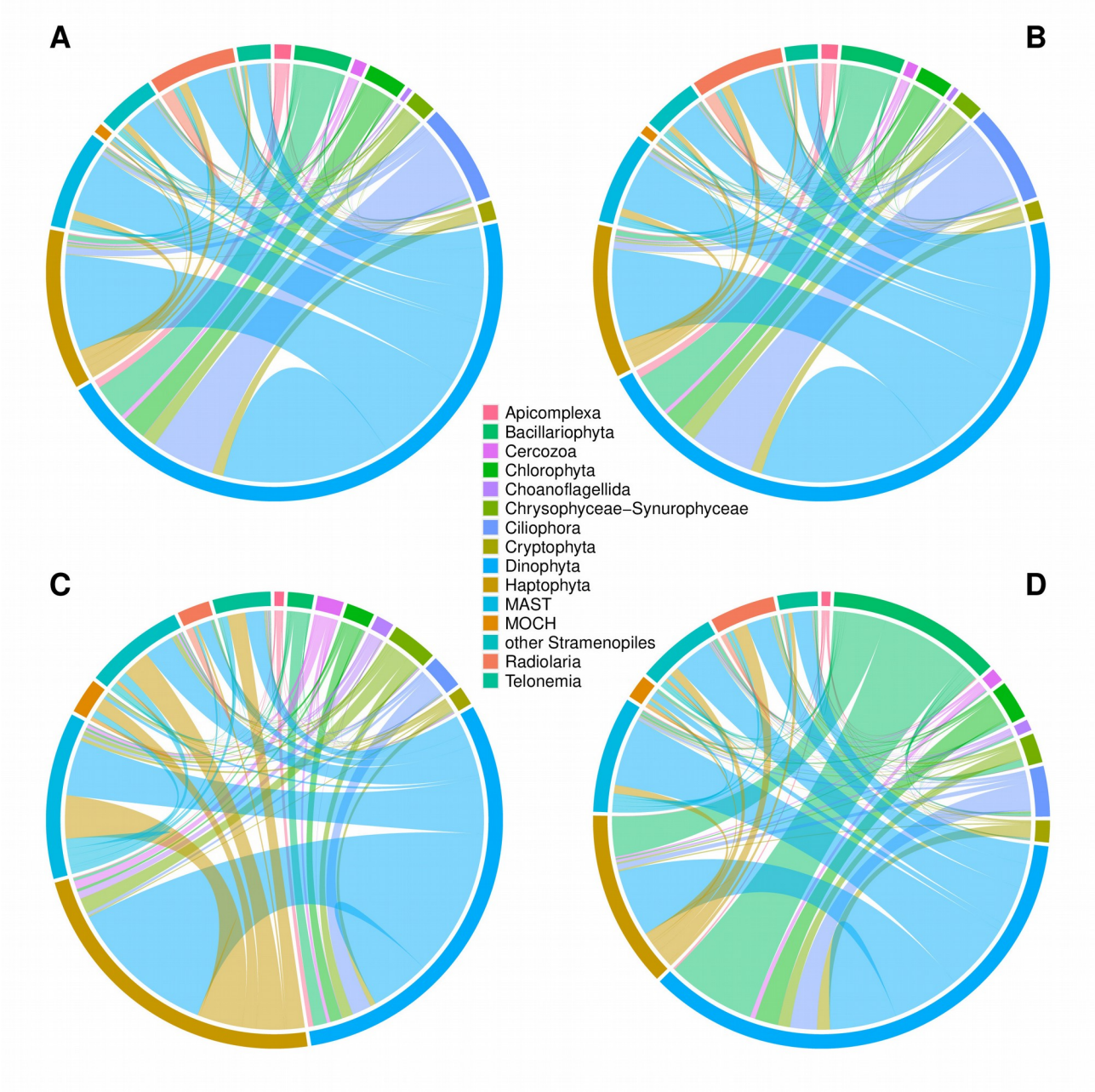


Figure S8 Proportion of edges between the different clades in the pairwise sequence genetic distance range 0.24-027 of the marine surface datasets. The two first chord diagrams represent all candidate edges for the co-occurrence (A) and co-exclusion (B) networks. The two last chord diagrams represent the observed distribution of edges in the co-occurrence (C) and co-exclusion (D) networks. Fold changes between observed and candidate edges ratio for each pair of clades are presented in Figure 4.

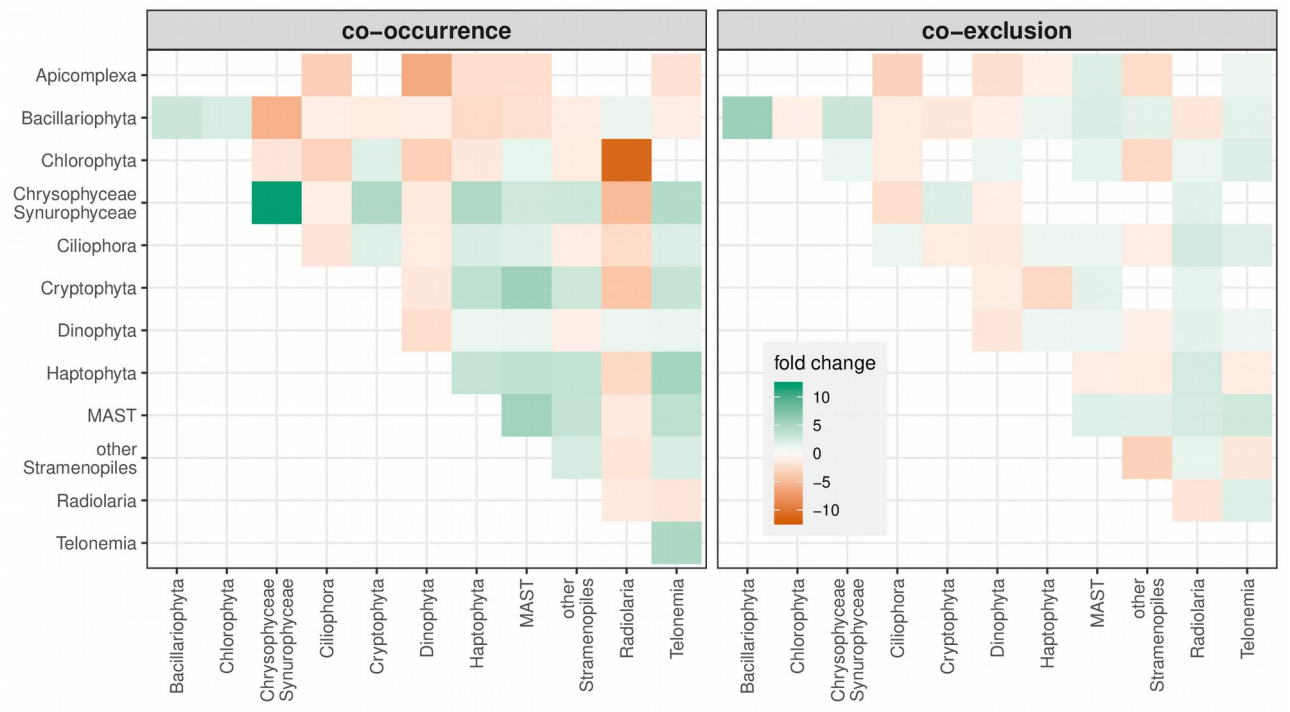


Figure S9 Fold changes in proportion of edges connecting the main clades in the marine DCM dataset compared to all candidate edges in the pairwise sequence distance range of 0.24-0.27 (*i.e.* the largest range of distance with simultaneous positive SES in co-occurrence and co-exclusion networks of the marine surface dataset when using the null model 1). The fold change color scale is identical to the one use for the marine surface dataset (Figure 4).