Gene similarity networks unveil a potential novel unicellular group closely related to animals from the *Tara* Oceans expedition

Alicia S. Arroyo[1*], Romain Iannes[2*], Eric Bapteste[2] & Iñaki Ruiz-Trillo[1,3,4◆]

[1] Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Passeig Marítim de la Barceloneta, 37-49, 08003 Barcelona, Spain

[2] Institut de Systématique, Evolution, Biodiversité (ISYEB), Sorbonne Université, CNRS, Museum National d'Histoire Naturelle, EPHE, Université des Antilles, Paris, France.

[3] Departament de Genètica, Microbiologia I Estadística, Institut de Recerca de la Biodiversitat, Universitat de Barcelona, Avinguda Diagonal 643, 08028 Barcelona, Spain

[4] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

*corresponding author: Iñaki Ruiz-Trillo; inaki.ruiz@multicellgenome.org

ABSTRACT

The Holozoa clade comprises animals and several unicellular lineages (choanoflagellates, filastereans and teretosporeans). Understanding their full diversity is essential to address the origins of animals and other evolutionary questions. However, they are poorly known. To provide more insights into the real diversity of holozoans and check for undiscovered diversity, we here analysed 18S rDNA metabarcoding data from the global *Tara* Oceans expedition. To overcome the low phylogenetic information contained in the metabarcoding dataset (composed of sequences from the short V9 region of the gene), we used similarity networks by combining two datasets: unknown environmental sequences from *Tara* Oceans and known reference sequences from GenBank. We then calculated network metrics to compare environmental to reference sequences. These metrics reflected the divergence between both types of sequences and provided an effective way to search for evolutionary relevant diversity, further validated by phylogenetic placements. Our results showed that the percentage of unicellular holozoan diversity remains hidden. We found novelties in several lineages, especially in Acanthoecida choanoflagellates. We also identified a potential new holozoan group that could not be assigned to any of the described extant clades. Data on geographical distribution showed that, although ubiquitous, each unicellular holozoan lineage exhibits a different distribution pattern. We also identified a positive association

between new animal hosts and the ichthyosporean symbiont *Creolimax fragrantissima*, as well as for other holozoans previously reported as free-living. Overall, our analyses provide a fresh perspective into the diversity and ecology of unicellular holozoans, highlighting the amount of undescribed diversity.

Keywords

networks, metabarcoding, 18S, molecular diversity, unicellular Holozoa, novelty

INTRODUCTION

The origin of animals from their unicellular ancestor is, undoubtedly, an important evolutionary question. To address this question in the most effectively way, we first need to have a well-resolved phylogenetic framework as well as a good understanding of the diversity of the closest unicellular relatives to animals (Ruiz-Trillo *et al.*, 2007). Thanks to phylogenomic analyses, a well-resolved phylogenetic framework of animal origins is now in place. We know that animals are closely related to several unicellular lineages, namely Choanoflagellatea, Filasterea, and Teretosporea (Ichthyosporea and Corallochytrea), all together forming the Holozoa clade (Lang *et al.*, 2002; Ruiz-Trillo *et al.*, 2004, 2008; Shalchian-Tabrizi *et al.*, 2008; Torruella *et al.*, 2012, 2015; Grau-Bové *et al.*, 2017). In contrast, environmental data show us that we still do not have a full understanding of the diversity of Holozoa (del Campo *et al.*, 2015; Arroyo *et al.*, 2018). Therefore, current interpretations on the evolutionary transition towards animal multicellularity may be challenged by improving our knowledge about Holozoa diversity (Ruiz-Trillo *et al.*, 2007; del Campo *et al.*, 2014).

To fill this gap and provide a more accurate perspective on Holozoa diversity and their geographical distribution, we analysed the longest and largest metabarcoding marine dataset: the *Tara* Oceans expedition, which is based on the 18S ribosomal RNA gene (hereafter 18S or 18S rDNA) (de Vargas *et al.*, 2015; Pesant *et al.*, 2015). *Tara* Oceans comprise thousands of reads from hundreds of sampling stations around the globe, with a third of those reads not matching any reference in databases (de Vargas *et al.*, 2015). However, a drawback of this dataset is the absence of full-length 18S sequences, being composed by the relatively small V9 region (around 130 bp long), located at the end of the 18S (Hugerth *et al.*, 2014).

To overcome the issue of the limited phylogenetic signal, we decided to analyse the *Tara* Oceans dataset using gene similarity networks. Networks have been preferentially applied to study ecological interactions, such as predator-prey, parasite-host or mutualism (Logares *et al.*, 2014; Krabberød *et al.*, 2017; Layeghifard *et al.*, 2017; Pilosof *et al.*, 2017; Valverde *et al.*, 2018). Networks are now becoming widely adopted to explain complex evolutionary processes, such as horizontal gene transfer, gene domain fusion, and gene or genome introgression (Corel *et al.*, 2016; Pathmanathan *et al.*, 2018; Ocaña-Pallarès *et al.*, 2019). To our knowledge, there are very few metabarcoding studies that used networks to describe novelty in metabarcoding datasets (Forster *et al.*, 2015; Forster *et al.*, 2019), even though this methodology offers a structure to test evolutionary questions in massive high-throughput data and to mine large datasets for sequences of interest.

Our analyses showed novel unicellular Holozoa diversity, in particular within Choanoflagellatea and Ichthyosporea. Specifically, we found unicellular Holozoa Operational Taxonomic Units (OTUs) branching off several acanthoecid subgroups (for example Choanoflagellate H), *Syssomonas multiformis* and *Creolimax fragrantissima*. We also retrieved 15 Filasterea-related OTUs, detecting this clade for very first time in an environmental survey. Interestingly, we also identified a putative novel unicellular Holozoa group, composed of 21 OTUs (6,244 reads in total)., that could not be located within any other known lineage and may represent a novel lineage (here tentatively named as MASHOL, for marine small Holozoa clade). We also observed that the freshwater environmental group FRESCHO3 could have diverged from a marine clade, showing another marine-to-freshwater transition in choanoflagellates. Finally, our co-occurrence analyses suggested potential novel associations between animals and ichthyosporeans. For example, the ichthyosporean *C. fragrantissima* could be associated with a broader range of animal hosts than previously described.

RESULTS AND DISCUSSION

Initial datasets & network construction

To look for potential new diversity of unicellular Holozoa and to address their geographical distribution we combined two 18S rRNA datasets: an environmental

dataset of OTUs (Operational Taxonomic Units) and a reference dataset with known holozoan sequences. The environmental dataset came from the worldwide *Tara* Oceans expedition (de Vargas 2015), which included metabarcoding data from the V9 region of the 18S rRNA gene from a total of 1,086 samples from 210 oceanic stations, 3 water column layers and 10 size fractions (further details about sampling procedures can be found in Pesant *et al.*, 2015). The reference dataset was built by collecting sequences from both GenBank Nucleotide and PR2 databases (see Materials and Methods).

The initial unicellular Holozoa network was built from 2,426 sequences (2,197 from *Tara* Oceans, 229 from the reference dataset). In the network, each node represented either an environmental OTU from *Tara* Oceans (hereafter ENV) or a sequence from the reference database (hereafter REF) (Figure 1). The basic structure of the network consisted of Connected Components (CCs): subgraphs of the network in which there is always a path between all nodes (Figure 2). The initial network was subsequently partitioned using increasing percentages of sequence similarity thresholds (≥85%, ≥87%, ≥90%, ≥95% and ≥97%), resulting in more fragmented networks (Figure 2). In each of these networks, CCs could be classified in three types: CCs in which all nodes were environmental (CC_{ENV}), CC in which all nodes were reference (CC_{REF}) and CC in which there were both types of nodes (CC_{MIX}) (Figure 1).

Networks produced at all thresholds displayed a similar trend: the number of CC_{ENV} was always the largest, followed by a CC_{MIX} and CC_{REF} (Supplementary Figure 1), which indicated the presence of abundant divergent groups of environmental sequences, independently of the stringency level considered.

Definition of novelty

To find potential novelty, we then explored the structure of the sequence similarity networks to search for molecular diversity. To do so, we calculated different metrics that are grouped into four categories:

I. ***Closeness centrality*** (Figure 2 and Supplementary Material 1): It defines to which extent a node (sequence) is central in a network. Typically, a peripheral sequence in a CC is more divergent than the rest of the nodes in this CC because it has less direct neighbours, meaning that peripheral sequences share less similarity with the majority of the sequences with which they cluster. Therefore, we tested whether and which environmental sequences (ENV) were significantly

more peripheral than reference sequences (REF) as a way to test whether ENV sequences extends the current known diversity of Holozoa, as well as to identify significantly peripheral ENV nodes.

II. ***Preferential association*** (***Assortativity,*** Figure 2 and Supplementary Material 1): Assortativity quantifies whether nodes that belong to the same category (e.g. ENV or REF) are more connected with each other rather than with nodes from other categories. For example, a significant preferential association between ENV nodes in a network would indicate the existence of groups of similar environmental sequences, distinct from sequences from already described Holozoa.

III. ***Network comparison (path analyses by BRIDES,*** (Figure 2 and Supplementary Figure 2***)***: It quantifies the new paths created in an augmented network when new sequences (e.g. ENV) are added to an original network (with only REF), as in Lord *et al.*, 2016. In particular, this allows the evaluation of whether newly added ENV sequences fill in some gaps between the original REF sequences. Typically, Breakthroughs (B paths) and Shortcuts (S paths) indicate that added ENV sequences decrease the topological distance (hence by assumption the putative phylogenetic distance) between known REF sequences. By contrast, Impasses (I paths) indicate that added ENV sequences locate outside short paths between REF sequences in the augmented network.

IV. ***Shortest-path distance*** (Figure 2): Shortest paths describe the minimal number of edges to connect any pairs of nodes in a network. We used these metrics to quantify a topological distance between ENV and REF nodes in the graph. By definition, increasingly divergent ENV sequences will be located increasingly far from REF sequences. If ENV and REF sequences are located in distinct CCs, there is even no path between them; thus the shortest path distance for such pairs of nodes is infinite.

All these steps of graph-mining were used to detect ENV sequences that could potentially indicate novelty, for which phylogenetic placement could be finally computed.

The structure of the unicellular Holozoa network shows potential undiscovered diversity

The general structure of the network provided an overview of the unicellular Holozoa diversity and highlighted potential new diversity (Figure 1). First, we computed the closeness of all nodes (Figures 2 and 3 and Supplementary Material 1) to test whether the distribution of closeness values for REF nodes was (i) significantly different and (ii) significantly higher than the distribution of closeness values for ENV nodes, using Wilcoxon signed-rank test. The results showed that ENV nodes were significantly more peripheral than REF nodes (Wilcoxon signed-rank test, p-value<0.01**) (Figure 3A) in all networks. This result indicates a high amount of potential new diversity in our unicellular Holozoa dataset from *Tara* Oceans. Not only the closeness distributions for REF nodes were significantly higher than that for ENV nodes, but also their shapes were different. At ≥85, ≥87 and ≥90% identity similarity thresholds, most closeness values of both ENV and REF distributions were low (95% confident interval between 0.2-0.4, approximately), and only few nodes presented a closeness value of 1. On the other hand, at ≥95 and ≥97% identity thresholds, when the network was more disconnected into divergent clusters of similar sequences, the distributions of closeness values for ENV nodes were scattered along a wider range of higher closeness values (~0.2-1). This change reflected the fragmentation of the network into more but smaller CCs.

Next, we analysed the assortativity, which showed significant preferential connections between ENV sequences. For every network, we computed (i) the distribution of null assortativity values by randomly shuffling the ENV and REF node labels, and we contrasted these values with (ii) the assortativity values of all our real networks (see Materials and Methods). All networks were significantly assortative (one sample t-test, p-value<0.01**) (Figure 3B). This tendency for intra-group preferential linkage suggests a lack of representation of oceanic Holozoa in the reference dataset before the *Tara* Ocean expedition, stressesing the high level of potential new diversity present in *Tara* Oceans data.

Overall, these metrics (closeness and assortativity) indicated that our environmental dataset of unicellular holozoans from *Tara* Oceans was different from the reference dataset, expanding the current known diversity of this group.

New molecular diversity in Holozoa, including a potential novel clade.

To identify new groups of interest, we first performed network comparisons using BRIDES software (Figure 2, Supplementary Figure 2) (see Materials and Methods and Lord 2016). This allowed us to contrast the topologies of networks built exclusively from REF nodes (original networks) with that in which ENV nodes had been included (augmented networks). BRIDES analysis showed that ENV sequences of unicellular Holozoa created numerous new paths in the augmented similarity networks (Figure 3C), guiding the discovery of evolutionary relevant novel sequences. First, despite the enhanced molecular diversity provided by the *Tara* Oceans dataset, some REF nodes remained disconnected from other REF nodes, indicating that the diversity of most ENV sequences was not close enough to fill the gaps between REF sequences. This was especially noticeable for networks built at high similarity thresholds. At ≥97% threshold, the vast majority of paths were *impasses* (I), meaning that ENV sequences did not create bridges between REF sequences in the augmented network (Supplementary Figure 2). This is logical because, given this high level of stringency, only sequences from the closest related holozoan lineages would connect in a given CC, confirming the general divergent nature of most ENV sequences with respect to sequences from sequenced holozoan taxa. Interestingly, when lowering the similarity threshold required to connect sequences in the networks, the proportion of *impasses* decreased, showing that some of these divergent ENV sequences started to connect some REF sequences. Still, at ≥85% identity, some Holozoa REF sequences remained disconnected, suggesting that the *Tara* Oceans dataset did not provide evidence for ENV groups bridging phylogenetic gaps between some known Holozoan clades. Possible explanations to this amount of *impasses* may be: (i) a lack of sufficient sampling effort, (ii) the absence of intermediate ENV sequences in marine water columns (there may be in other habitats), (iii) the nature of the Holozoa clade, which may be comprised of some significantly divergent lineages without extant intermediate diversity between them, or (iv) that most ENV sequences belong to groups branching outside currently described Holozoans.

On the other hand, *breakthroughs* (B) and *shortcuts* (S) were increasingly observed in networks at lower thresholds (Figure 3C). These two types of paths correspond to sequences that introduce either new paths between known Holozoan groups (B) or new ENV sequences closely related to known groups, and likely belonging to known clades (S). Thus, under the hypothesis that an intermediate position in the network reflected an intermediate phylogenetic position in the corresponding sequence phylogeny (Atkinson *et al.*, 2009; Méheust *et al.*, 2018), we assumed B paths could potentially indicate ENV

sequences branching in between two phylogenetically distant groups of Holozoans in a phylogenetic tree, whereas S paths may potentially indicate ENV sequences branching within a less divergent group of sampled holozoans (Supplementary Figure 2). Overall, the presence of a high proportion of B and S paths (36.93% at ≥85%, 33.22% at ≥87%, 45.42% at ≥90% ID) suggested that *Tara* Oceans data hinted at the existence of novel, phylogenetically relevant, holozoan diversity.

To corroborate the potential novelty of those sequences and have a better understanding of their phylogenetic position within Holozoa, we performed phylogenetic placement analyses (see Materials and Methods). In particular, we analysed the OTUs that created *breakthroughs* and *shortcuts* in the network at 85% similarity threshold (Figure 3D). These OTUs unravelled novelty within Acanthoecida, one of the two subgroups of Choanoflagellatea. A group of 6 sequences or OTUs (with a total of 1,675 reads) branched off Choanoflagellate H, suggesting a potential novel environmental group of acanthoecids. Another group of 3 sequences (including one of the most abundant OTUs in the whole *Tara* Oceans dataset: OTU 2703, with more than 28,000 reads) appeared to be the sister group of Choanoflagellate G. The importance of this result lies in the fact these OTUs did not cluster together with the already morphologically described Choanoflagellate G species (i.e. *Acanthocorbis unguiculata, Acanthoeca spectabilis*, *Savillea micropora*, *Helgoeca nana*), but branched at an internal node, showing their divergent nature. We also recovered the second earliest diverging acanthoecid (OTU 5953, with 7,448 reads), splitting apart from the reference sequence JQ223245, which had already been identified as a divergent choanoflagellate (del Campo 2015). Finally, several OTUs clustered within freshwater environmental choanoflagellate groups, such as FRESCHO3 or FRESCHO1, which shows a wider ecosystem range in which these species can inhabit. We confirmed the good quality of these phylogenetic placements gauging the likelihood and distance between placements (Supplementary Figure 3A,B). Alignments and the full tree of Figure 3D can be found in Supplementary Material 2.

Our second approach to examine in detail the novelty in unicellular Holozoa was to perform a shortest-path distance analyses between every ENV node and its closest REF node in the network (Figure 4). The longer the topological distance between REF and ENV nodes, the more divergent the ENV sequence is, because many steps are required to reach the nearest REF sequence. The most extreme case is the infinite distance, shown by ENV nodes belonging to exclusively environmental CCs. Our results showed

that indirect connections to REF (when there are more than 1 step from ENV to REF) were the most abundant, ranging from 92.5% of all ENV nodes at ≥85% ID similarity network to 69.83% at ≥97% identity (Figure 4A). In addition, networks at higher similarity thresholds (≥95% identity and ≥97% identity) exhibited a high proportion of infinite distances (15.39% of ENV nodes at ≥95% similarity threshold; 30.56% at ≥97% similarity threshold) (Figure 4A). We then extracted those distant ENV OTUs to perform phylogenetic placement against a curated reference Holozoa tree (see Materials and Methods). The deepest novelty (understood as the diversity that lays in deeper, more internal nodes in the Holozoan tree) was observed in the networks at ≥95% and ≥97% thresholds. We performed a specific phylogenetic placement of this deep novelty, shown in Figure 4B. A group of 21 OTUs with a total abundance of 6,244 reads was located in the most internal branch outside Choanoflagellata, specifically scattered across the internal branches of choanoflagellates and *Syssomonas multiformis*. These OTUs were mainly recovered in the pico (0.8-3/5 µm) and nano (3/5-20 µm) fraction sizes from the Indian Ocean and Mediterranean Sea. Inspired by its uncertain phylogenetic position and the small size, we tentatively named this group as MASHOL (standing for MArine Small HOLozoa). The quality of the placement test revealed that the placements had very low Likelihood Weight Ratios (Supplementary Figure 3D), although all of them were located around the same internal branches in the tree. As Mahé *et al.*, 2017 pointed out, these low-probability placements do not necessarily mean that they are incorrect, but they hold a high molecular distance with the reference sequences in the tree. This result indicates that these OTUs do not really belong to any of the already known unicellular holozoan lineages, although its exact position remain uncertain. In any case, they probably represent a novel clade among Holozoa.

Unicellular holozoans are globally distributed, with some lineages showing specific geographical patterns

There is no data on the geographical distribution of unicellular Holozoa. Thus, we decided to take the most of the Tara Oceans dataset and evaluate the geographical distribution of the different unicellular holozoan lineages across oceans, layers of the water column, and size fractions. In general, all lineages of unicellular Holozoa were widely distributed across the world's oceans (Figure 5A). Ichthyosporeans were the most homogeneously dispersed group across all oceans. There were, however, some

exceptions. For example, Acanthoecida choanoflagellates were more abundant in the Arctic samples (60.29% of total abundance), and in contrast to Craspedida (4.5%) (Figure 5A). These results are consistent with previous morphological studies of choanoflagellates in sea ice (Thomsen *et al.*, 1997). OTUs assigned to Filasterea were widely distributed, but their abundance was higher in the samples coming from the South Pacific Ocean (43.37%), Red Sea (24.7%) and Indian Ocean (16.97%) (Figure 5A). OTUs related to Corallochytrea group were widely distributed, although the OTU with the highest abundance (OTU 30781, 248 reads) was mainly located in the North Pacific Ocean (Figure 5A). Both the Indian Ocean and the Arctic Ocean held 30% of the reads of corallochytreans (Figure 5A). On the contrary, the presence of corallochytreans in the Atlantic Ocean seemed to be insignificant. Regarding the environmental groups Marine Opisthokonts 1 and 2 (MAOP1 and MAOP2, respectively), they showed a pattern of distribution similar to Choanoflagellata. MAOP2 appeared to be most abundant and with more OTUs than MAOP1, in contrast to what had been found in European coastal waters (del Campo *et al.*, 2015). Moreover, while MAOP1 was not found in the Arctic or Antarctic Oceans, MAOP2 exhibited 36% of its abundance in the Arctic, expanding to the maximum the range of geographical locations in which this environmental group has been found up to now (Figure 5A) (Romari and Vaulot, 2004; Amacher *et al.*, 2009; Edgcomb *et al.*, 2011; Marshall and Berbee, 2011). Assortativity coefficients of geographical distribution across oceans and oceanic provinces showed positive values in all networks (Supplementary Table). Even though these values were not very high (a range from 0.016 in the network at ≥85% identity similarity threshold to 0.046 in the network at ≥97%identity), it shows a tendency of OTUs from the same geographical region to be more associated between them, hence genetically more similar, than with OTUs from other regions.

Regarding the depth in the water column, the majority of the unicellular Holozoans were preferentially located in the surface or the Deep Chlorophyll Maximum (DCM) layers (Figure 5B). This tendency indicates that holozoan sequences in the upper layers were more similar than those sampled at lower depths (positive assortativity, Supplementary Table). Even though these are low positive numbers, they were significantly different from the random shuffled distribution (one sample t-test, p-value<0.01**), which supported the tendency for a shallower preference location.

Finally, unicellular holozoans were recovered from a wide range of size fractions (Figure 5C). For example, within Choanoflagellata, the majority of Acanthoecida abundance (69.37%) was present in the nano fraction (3/5-20 µm), followed by 19.4% in the pico fraction (0.8-3/5 µm). Filasterean reads were mainly found in meso (43.18%) and nano (46.21%) fractions. Ichthyosporeans had a different pattern of sizes (Figure 5C). The distribution of Dermocystida reads was shifted towards the largest fractions (10.96%, 19.98% and 57.73% in meso, micro and nano fractions, respectively). On the contrary, the distribution of Ichthyophonida reads was shifted towards the smallest fractions (24.46% in nano and 61.97% in pico fractions). OTUs associated with Corallochytrea were preferentially found in the pico, nano and pico-nano fractions (0.8-20 µm). Finally, both MAOP groups were more present in the smallest fractions: nano (54.94%) and pico (37.81%), which differ from previous findings that showed MAOP dominating the micro fraction (del Campo and Ruiz-Trillo, 2013). Nevertheless, these results are consistent with these authors, who already suggested that MAOP group might be composed by species with different sizes. The MAOP group might also undergo a life cycle with several stages that include different cell sizes. The preferential location of different holozoan lineages in different size fractions can be seen in the assortativity values (Supplementary Table). In all networks, assortativity coefficients of fraction sizes were the highest among all elements considered (depths, oceanic provinces, oceans and size). These values were also significant compared to the distribution of randomly shuffled labels (one sample t-test, p-value<0.01**), indicating a tendency for similar Holozoa sequences to be found in specific size fraction, compared to other sizes.

Co-occurrence of *Creolimax fragrantissima* and its animal hosts

Some of these unicellular species, especially the Ichthyosporea, have been previously described as animal parasites or symbionts (Mendoza *et al.*, 2002; Glockling *et al.*, 2013). To see whether our data could illuminate us on this aspect, we checked if there was any association between the presence of unicellular Holozoa and animals.

Our results showed that there were indeed significant positive and negative correlations between unicellular Holozoa and animals (Figure A). The strongest positive correlation (Spearman's rank correlation coefficient, ρ_S=0.6-0.8, p<0.01**) was shown between OTUs associated with *Creolimax fragrantissima* and several animal phyla such as Entoprocta (Barentsiidae), Mollusca (Polyplacophora), Tardigrada, and Porifera

(Homoscleromorpha, Calcarea and Demospongiae). To see if we could detect other associations but monotonic and linear (as Spearman and Pearson describe, respectively), we used a bipartite network (Figure 6B). We corroborated the previous finding of *C. fragrantissima* with several animal phyla, specifically with Polyplacophora (ρ_S=0.465), Calcarea (ρ_S=0.352) and Demospongiage (ρ_S=0.311). *C. fragrantissima* was isolated 27 times from invertebrate guts, mostly from a sipunculid species, but also one tunicate, sea cucumber and chiton (Marshall *et al.*, 2008). Thus, our results corroborated some symbiotic relationships (with Polyplacophora, commonly known as chiton) and suggested some other putative hosts (Entoprocta, Tardigrada and Porifera).

We also found that the environmental group Marine Ichthyosporea 1 (MAIP1) was associated with Acoelomorpha, Arthropoda (Hexapoda, Crustacea), Bryozoa, Cnidaria, Nematoda (Enoplea) and Chordata (Tunicata, Craniata). This result suggests that the environmental group MAIP1 may be associated with animal phyla and not being exclusively free-living. Another interesting result was the interaction between MAOP2 and Ctenophora (ρ_S=0.409) or Mollusca (Cephalopoda) (ρ_S=0.317), which could imply that these taxa use the same resources or have some ecological interaction, as it was found for other environmental groups (Lima-Mendez *et al.*, 2015; Lambert *et al.*, 2019). Regarding MASHOL, the potential new Holozoa group described here, no strong correlations could be found with any animal group, suggesting that this environmental group might be free-living or not have a strong association with any particular animal phyla.

Overall, these results suggest more complex ecological interactions between parasitic/symbiotic unicellular holozoans and animals than what it is currently known. These biotic effects (grazing, pathogenicity and parasitism) have been reported to explain 82% of the variability in the *Tara* Oceans interactome, giving a greater importance to these interspecific connections (Lima-Mendez *et al.*, 2015). This also implies that sampling within animal phyla may still be a useful method to isolate new species from unicellular holozoans. However, we refuse to claim that correlation implies causation. What is certain though is that metabarcoding has a great power to assess diversity in its multiple forms, from pure ecological and evolutionary studies to applied conservationism, which is of vital importance in a world of threat to biodiversity.

CONCLUSIONS

Our analysis of metabarcoding data from *Tara* Oceans using sequence similarity networks shows a greater diversity of unicellular holozoans than previously sampled, including a potential novel clade. Our data also demonstrates global geographical distribution from most unicellular holozoans and pinpoints to potential associations with different animal phyla.

MATERIALS & METHODS

Datasets

The initial environmental dataset was provided by the *Tara* Oceans consortium, which contained a total of 474,303 Operational Taxonomic Units (OTUs) from all eukaryotic clades. Note that this is the full dataset generated in the expedition, not the one used in de Vargas *et al.*, 2015, as the latter is a subsample of the former. The *Tara* Oceans consortium provided us with this dataset already cleaned, filtered and clustered. During the first steps of the bioinformatic pipeline, they merged, dereplicated and quality filtered the original V9 barcodes. A chimera detection analysis was carried out using the usearch program (Edgar *et al.*, 2011). After a filtering process to discard possible spurious reads, barcodes were clustered using Swarm approach (Mahé *et al.*, 2014). For further details on the OTU table generation, see http://taraoceans.sb-roscoff.fr/EukDiv/.

Our reference database was obtained by merging three different databases: GenBank, PR2-Opistho and PR2_V9. First, we downloaded two databases from GenBank: nucleotide (nt) and environmental nucleotide (env_nt) by January 25th 2018. We retrieved 18S rDNA sequences from these databases by searching them using the human 18S sequence as a query (AC139250, positions 551,257 to 553,055). This sequence had been previously confirmed to contain the *Tara* Oceans V9 primer sequences. BLASTn parameters were: E-value <1E-10, percentage of identity ≥60% and maximum target sequences of $9{,}9 \cdot 10^7$ (for nt) and $9{,}9 \cdot 10^8$ (for env_nt). From the BLASTn output, we implemented two filtering processes. In the first one, we retrieved the sequences that contained both *Tara* Oceans V9 primer sequences. We then trimmed the sequences to have only the V9 region. In the second step, we kept those sequences whose length was comprised between 80 and 120 base pairs to keep the most frequent length range of this region (Amaral-Zettler *et al.*, 2009). The second database, PR2-Opistho, was a well-curated and updated version of the original PR2 database for Opisthokonta clade. This database (PR2-Opistho) was also trimmed with the *Tara*

Oceans primer sequences to keep only the V9 region. The third database, PR2_V9, was generated by the *Tara* Oceans consortium (de Vargas *et al.*, 2015). Because both PR2-Opistho and PR2_V9 were originally generated from PR2 database, we eliminated redundancies and kept the taxonomical annotation from the PR2-Opistho database. Finally, we combined all databases, producing a global reference database of 49,379 eukaryotic sequences.

To retrieve the unicellular Holozoa sequences, we performed a phylogenetic placement of both environmental and reference datasets against a eukaryotic reference tree and took those that branched within Holozoa and outside animals. A phylogenetic placement consists of mapping short amplicons (in this case, *Tara* Oceans OTUs) into a fixed reference tree made from full-length 18S rDNA sequences. This reference was constructed using 130 full 18S sequences that covered all eukaryotic groups. We performed the phylogenetic placement using the RAxML-EPA algorithm (Berger *et al.*, 2011) and we selected the sequences that were placed into unicellular Holozoa using the C++ script extract_clade_placements from Genesis software v0.18.1 (Czech and Stamatakis, 2016). Therefore, the starting dataset of unicellular Holozoa contained 2,426 sequences (2,197 were environmental from *Tara* Oceans while 229 were reference sequences). This dataset can be found in Supplementary Material 4.

Similarity Network construction

We built the initial similarity network based on a blast all-against-all of the unicellular Holozoa dataset. We used BLASTn v2.7.1+ (Camacho *et al.*, 2009), with the following options: E-value <1E-10, percentage of identity ≥85%, maximum number of HSPs 1 and maximum target sequences 3,000.

We used the cleanblastp script from CompositeSearch software to filter the output in order to remove auto-loops and reciprocal connections (A-B would be the same as B-A) (Pathmanathan *et al.*, 2018). Final networks were obtained by setting up a mutual cover threshold of ≥95% and increasing sequence similarity thresholds: ≥85%, ≥87%, ≥90%, ≥95% and ≥97% identity threshold, respectively. These networks can be found in Supplementary Material 4.

Network node annotation

In order to annotate taxonomically every node in the network, we performed a BLAST of the initial 2,426 holozoan sequences against the PR2-Opistho database, using the

following parameters: E-value <1E-50 and ≥97% percentage of identity. Under these conditions, only 438 sequences could be annotated. Thus, we decided to use a phylogenetic method to taxonomically assign the rest of the unannotated OTUs: tax2tree algorithm (McDonald *et al.*, 2012). This software requires the structure of the phylogenetic tree of both reference and unannotated sequences. Then, it assigns the taxonomy to the unannotated tips, given a file with the taxonomical information of the annotated tips. We could successfully annotate 1,503 additional sequences. Thus, a total of 1,941 sequences (78.8% of the initial dataset) could be taxonomically annotated.

Sequence similarity Network analyses

To address the molecular diversity and novelty of unicellular Holozoa, we analysed topological metrics, as well as closeness and assortativity using NetworkX v2.1 library on python 3.5.1 (Hagberg *et al.*, 2008).

Novelty assessment: preferential connection

Assortativity is a property of the network that measures the preferential connection between nodes belonging to the same group (Newman, 2003; Forster *et al.*, 2015) (Figure 2). To compute its significance, we first calculated a distribution of null assortativity values for each network, randomly distributing the same amount of node labels to the ones existing (e.g. REF and ENV) under test. The reason is that a random null assortativity value may be different from 0, given the structure of the graph and the group sizes of the tested labels (Figure 2 and Supplementary Material 1). Next, in the standard protocol, we randomly shuffled the labels of the nodes 100 times while keeping the same network topology. For example, one ENV node (i.e., a node composed of an environmental sequence) could turn out to be ENV or REF (i.e., a node composed of a reference sequence) after the shuffling. For all these 100 random networks, we computed the assortativity, generating the distribution of assortativity values for random networks. We next computed the actual value of assortativity in the networks (Figure 3B and Supplementary Table), for each tested pairwise comparison of categories to calculate the p-values of our observations (ENV vs REF; IND vs MEDIT vs ARCTIC vs ANTAR vs NPAC vs SPAC vs NATL vs SATL vs REDS; SURF vs DCM vs MES vs MIX vs ZZZ; MESO vs MICRO_MESO vs MICRO vs NANO vs PICO_NANO vs PICO_MICRO vs PICO).

Novelty assessment: BRIDES

BRIDES software characterizes new paths that are created when extra nodes are added to an original network (Lord *et al.*, 2016). For every sequence similarity network, we first used only the REF nodes (original network), and then we added the ENV nodes of unicellular Holozoa (augmented networks) to compute BRIDES using the default parameters.

Novelty assessment: phylogenetic placement

In order to validate the putative novel diversity previously obtained with BRIDES and shortest-path analyses, we performed a phylogenetic placement of the OTUs into our curated reference Holozoa tree, which can be found in Supplementary Material 5. We aligned the sequences using PaPaRa with default parameters (Berger and Stamatakis, 2011) and manually examined the alignment and corrected wrong positions in Geneious v9.0.5 (Kearse *et al.*, 2012). We then trimmed the non-homologous positions with trimAl 1.4.rev15, setting the gap threshold option at 0.2 for the alignment of selected sequences found on B and on S paths by ourBRIDES analysis (Capella-Gutiérrez *et al.*, 2009). Regarding the alignment of divergent sequences identified by our shortest-path analyses, the trimming was done manually, removing those positions with a mean pairwise identity over all pairs below 30%. We performed the phylogenetic placement using the RAxML-EPA algorithm (Berger *et al.*, 2011). The final tree in figure 4B was enhanced using iTOL (Letunic and Bork, 2016).

We validated the quality of the phylogenetic placement using the placement_histograms script from Genesis package v0.18.1 (Czech and Stamatakis, 2016). The first parameter computed was the EDPL (Expected Distance between Placement Locations). For every OTU, it calculates the weighted distance between all placement positions. In other words, EDPL quantifies to which extent all placements from an OTU are scattered over the tree. In both groups, EDPL values were extremely small (<0.05) (Supplementary Figure 2A,C). Considering that most branches in the tree had less than 0.05 nucleotide substitutions per site, it meant that the majority of the OTUs were located within the same branch. However, the quality of these placements was not high, measured as the distribution and frequency of Likelihood Weight Ratio values (LWR). This was especially drastic in the placements of MASHOL OTUs (Supplementary Figure 2D), which shows the uncertainty in the location of the group.

Geographical distribution

We described the geographical distribution of unicellular Holozoa lineages, as well as the distribution along the water column and size fractions, through circular layouts using "circlize" package in Rstudio (Gu *et al.*, 2014; RStudio, 2017)

Co-occurrence patterns

To test the association between unicellular Holozoa and animal OTUs, we carried out a co-occurrence analysis. First, we filtered the dataset to keep those OTUs that were present in at least 3 samples (out of 1,086 total samples in *Tara* Oceans). Then, we summed up OTU abundances if these OTUs belonged to the same class in animals or the same genus/species in unicellular Holozoa. We used "corrplot" and "Hmisc" libraries in Rstudio v.1.1.383 to perform the analyses (RStudio, 2017; Wei *et al.*, 2017; Harrell, 2019). These consist of building a correlation matrix among all pairwise comparisons and then extract the significant relationships (Spearman's significance<0.01**), which finally were plotted in a heatmap.

There was a possibility, however, that some associations could be neither monotonic nor linear. In that case, we would not be able to detect them using Spearman's or Pearson's correlation coefficients. We used instead MICtools package (Albanese *et al.*, 2018), which is able to identify a wider range of relationships in large datasets and assess their statistical significance. Final networks were created using Cytoscape 3.3.0 (Shannon *et al.*, 2003).

ACKNOWLEDGMENTS

We thank Ramón Massana, Philippe Lopez, and Ramiro Logares for discussion on the manuscript. This work was supported by grants (BFU2014-57779-P and BFU2017-90114-P) from Ministerio de Economía y Competitividad (MINECO), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER) to I.R.-T.

FIGURE LEGENDS

Figure 1. Network metrics. Upper panel: once the unicellular Holozoa network was constructed, different similarity thresholds were applied to gain a more detailed structure of their diversity. Lower panels: network metrics computed in this study to address

molecular novel diversity in unicellular Holozoa. A more technical explanation of closeness and Assortativity can be found in Supplementary Material 1, and of BRIDES in Supplementary Figure 2.

Figure 2. Unicellular Holozoa network at ≥85% similarity threshold. Environmental nodes from Tara Oceans are depicted with triangles that are coloured according to the distance to their shortest reference sequences (right panel). Reference nodes from GenBank dataset are depicted with circles that are coloured according to the taxonomy (left panel). Connected Components composed of only reference nodes are located in the top right corner. The novel Holozoa group described in this paper, MASHOL (for MArine Small HOLozoa), is shown in red triangles and pointed in the network with a black circle. Raw network data can be found in Supplementary Material 4.

Figure 3. Network approach to the analysis of novel diversity of unicellular Holozoa. (A) Closeness distribution of reference nodes was significantly higher than that of environmental nodes. This showed that environmental nodes were located at the periphery of the connected components because they were more divergent. Two asterisks mark the significance of the Wilcoxon signed-rank test when p-value<0.01. **(B) Assortativity** values were significantly positive in all networks, meaning that environmental nodes tended to connect preferentially together rather than with reference nodes. **(C) BRIDES analysis**. Environmental OTUs from unicellular Holozoa created new paths with respect to the original reference network, as green bars show (see Supplementary Figure 2 for details about each type of path). **(D) New molecular groups in Choanoflagellatea**. Phylogenetic placement of the OTUs that created breakthroughs and shortcuts at ≥85% similarity threshold in (C; in red) against a curated reference tree of unicellular Holozoa. We computed the placement using the RAxML-EPA algorithm with the GTR+CAT+I evolutionary model (Berger *et al.*, 2011). Several OTUs branched off some acanthoecid clades, such as Choanoflagellate I, G and H, showing a different diversity from the extant known species. This novel molecular diversity is well supported by the high abundance of some OTUs (shown as the number in brackets) and the good quality of their placement (Supplementary Figure 3 A,B). Alignments and the full phylogenetic tree can be found in Supplementary Material 2.

Figure 4. Potential new group of unicellular Holozoa (MASHOL) found branching off Choanoflagellatea. (A) Shortest path analysis showed that a considerable

proportion of environmental nodes have infinite distance with their closest reference node (15.39% in the network at ≥95% similarity threshold; 30.56% in the network at ≥97%). These ENV nodes were not connected to any reference node whatsoever, suggesting a substantial amount of diverging diversity. **(B)** Phylogenetic placement of the 21 OTUs that exhibited infinite distance in the networks at ≥95% and ≥97% similarity threshold in (A). All OTUs were allocated in internal branches, outside Choanoflagellatea and *Syssomonas multiformis*, depicted as a thick magenta line. The lack of high support (measured as Likelihood Weight Ratio or LWR) in the placements suggests a deep uncertainty about the exact placement of these sequences in the Holozoa tree of life (Supplementary Figure 3D). However, their narrow scattering over the tree and their clear position in internal rather than external branches open the possibility for these OTUS to be a potential new Holozoa group that we tentatively named as MASHOL (for Marine Small HOLozoa). Phylogenetic placement was carried out using RAxML-EPA algorithm (Berger *et al.*, 2011) under the GTR+CAT+I evolutionary model. Alignments and the full phylogenetic tree can be found in Supplementary Material 3.

Figure 5. Geographical distribution of unicellular Holozoa OTUs from the *Tara* Oceans expedition. As depicted in the example (bottom left panel), chord diagrams show OTUs on the bottom half of the circle, and oceanic regions, depths and fraction sizes on the upper half. Each OTU is represented by a line, whose thickness depicts the OTU's abundance in that particular place. In general, all unicellular holozoans were widespread and located in surface or DCM layers of the water column. However, some had different preferential geographical location (i.e., MAOP1 *vs* MAOP2, or Craspedida *vs* Acanthoecida), or fraction sizes (i.e., Ichthyophonida *vs* Dermocystida, or Craspedida *vs* Acanthoecida). Note that the thickness of each OTU is relative to the amount of OTUs in each group, so comparisons between lineages are not possible. Numbers below group names indicate the number of OTUs.

Figure 6. Co-occurrence analysis between unicellular Holozoa OTUs and animal classes from *Tara* Oceans. **(A)** Heatmap representing the Spearman's rank correlation coefficient (ρ). The ichthyosporean symbiont *Creolimax fragrantissima* had the strongest correlation coefficient (ρ_S=0.6-0.8, p<0.01**) with several animal phyla, suggesting a wider diversity of animal hosts in which this organism can dwell. Full heatmap can be found in Supplementary Figure 4A. **(B)** Network depicting other possible associations, besides monotonic and linear. The environmental clades marine ichthyosporea 1

(MAIP1) and marine opisthokonta 2 (MAOP2) were connected with several animal phyla, suggesting non-exclusive free-living lifestyles, or coincidence due to the use of same ecological resources. Full network can be found in Supplementary Figure 4B.

Supplementary Figure 1. Topological metrics of each network. Connected Components (CCs) with only environmental nodes exceeds the rest of CCs because of the unequal amount of environmental sequences compared to reference sequences in the original database (2,197 environmental sequences; 230 reference sequences). Number of nodes reflects only the nodes that are connected, not singletons. This is the reason why the number of nodes decreases as the similarity threshold increases.

Supplementary Figure 2. BRIDES paths. An illustration of all BRIDES paths, together with the possible biological interpretation. Blue nodes and edges are generated by the environmental sequences (ENV), which are added to the original network only made from reference sequences (REF), depicted by black nodes and edges. We focused on the path highlighted with a red box because they were the simplest to interpret from a biological standpoint.

Supplementary Figure 3. Phylogenetic placement validation. (A,C) The Expected Distance between Placement Locations (EDPL) indicates whether one OTU is scattered over the tree or not. The smaller the EDPL, the better is the placement because it is located in a specific area of the tree. **(B,D)** Barplot represents the first three most probable Likelihood Weight Ratios (LWR) of each OTU. In (D) the distribution of the placements was left-tailed, showing the uncertainty of the placement.

Supplementary Figure 4. Co-occurrence analysis of unicellular Holozoa OTUs and animal classes from *Tara* Oceans. **(A)** Significant correlations (Spearman's significance<0.01**) range from negative values (brown) to positive ones (blue). "_X" sign after a taxa means "unknown". Unicellular Holozoa are depicted in red. **(B)** Significant correlations (Maximal Information Coefficient, MICe, between 0.08-0.638) displayed among unicellular Holozoa.

REFERENCES

Albanese, D., Riccadonna, S., Donati, C., and Franceschi, P. (2018) A practical tool for maximal information coefficient analysis. *Gigascience* **7**: 1–8.

Amacher, J., Neuer, S., Anderson, I., and Massana, R. (2009) Molecular approach to determine contributions of the protist community to particle flux. *Deep Sea Res. Part I* **56**: 2206–2215.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLoS One* **4**: e6372.

Arroyo, A.S., López-Escardó, D., Kim, E., Ruiz-Trillo, I., and Najle, S.R. (2018) Novel Diversity of Deeply Branching Holomycota and Unicellular Holozoans Revealed by Metabarcoding in Middle Paraná River, Argentina. *Front. Ecol. Evol.* **6**: 99.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., Babbitt, P.C. (2009) Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS One*. **4**: e4345.

Berger, S.A., Krompass, D., and Stamatakis, A. (2011) Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Syst. Biol.* **60**: 291–302.

Berger, S.A. and Stamatakis, A. (2011) Aligning short reads to reference alignments and trees. *Bioinformatics* **27**: 2068–2075.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

del Campo, J., Mallo, D., Massana, R., de Vargas, C., Richards, T. a., and Ruiz-Trillo, I. (2015) Diversity and distribution of unicellular opisthokonts along the European coast analysed using high-throughput sequencing. *Environ. Microbiol.* n/a-n/a.

del Campo, J. and Ruiz-Trillo, I. (2013) Environmental Survey Meta-analysis Reveals Hidden Diversity among Unicellular Opisthokonts. *Mol. Biol. Evol.* **30**: 802–805.

del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., and Ruiz-Trillo, I. (2014) The others: our biased perspective of eukaryotic genomes. *Trends Ecol. Evol.* **29**: 252–259.

Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016) Network-Thinking: Graphs

to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* **24**: 224–237.

Czech, L. and Stamatakis, A. (2016) Genesis. A Toolkit for Working with Phylogenetic Data. *https://github.com/lczech/genesis*.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., et al. (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**: 1344–1356.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., et al. (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol.* **13**: 1–16.

Forster D, Lentendu G, Filker S*, et al.* (2019) Improving eDNA-based protist diversity assessments using networks of amplicon sequence variants. *Environ. Microbiol.* **21**: 4109-4124.

Glockling, S.L., Marshall, W.L., and Gleason, F.H. (2013) Phylogenetic interpretations and ecological potentials of the Mesomycetozoea (Ichthyosporea). *Fungal Ecol.* **6**: 237–247.

Grau-Bové, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A., and Ruiz-Trillo, I. (2017) Dynamics of genomic innovation in the unicellular ancestry of animals. *Elife* **6**: e26036.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014) circlize implements and enhances circular visualization in R. *Bioinformatics* **30**: 2811–2812.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008) Exploring network structure, dynamics, and function using NetworkX. In, Varoquaux,G., Vaught,T., and Millman,J. (eds), *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Pasadena, CA USA, pp. 11–15.

Harrell, F.E. (2019) Hmisc: Harrell Miscellaneous. *https://github.com/harrelfe/Hmisc*.

Hugerth, L.W., Muller, E.E.L., Hu, Y.O.O., Lebrun, L.A.M., Roume, H., Lundin, D., et al. (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. *PLoS One* **9**:.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012) Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Krabberød, A.K., Bjorbækmo, M.F.M., Shalchian-Tabrizi, K., and Logares, R. (2017) Exploring the oceanic microeukaryotic interactome with metaomics approaches. *Aquat. Microb. Ecol.* **79**: 1–12.

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., and Galand, P.E. (2019) Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *ISME J.* **13**: 388–401.

Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., and Burger, G. (2002) The closest unicellular relatives of animals. *Curr. Biol.* **12**: 1773–1778.

Layeghifard, M., Hwang, D.M., and Guttman, D.S. (2017) Disentangling Interactions in the Microbiome: A Network Perspective. *Trends Microbiol.* **25**: 217–228.

Letunic, I. and Bork, P. (2016) Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics* **44**: 127–128.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., et al. (2015) Determinants of community structure in the global plankton interactome. *Science (80-.).* **348**: 1262073–1262073.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., et al. (2014) Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**: 813–821.

Lord, E., Le Cam, M., Bapteste, É., Méheust, R., Makarenkov, V., and Lapointe, F.J. (2016) BRIDES: A new fast algorithm and software for characterizing evolving similarity networks using breakthroughs, roadblocks, impasses, detours, equals and shortcuts. *PLoS One* **11**: 5–7.

Mahé, F., De Vargas, C., Bass, D., Czech, L., Stamatakis, A., Lara, E., et al. (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat. Ecol. Evol.* **1**:.

Marshall, W.L. and Berbee, M.L. (2011) Facing Unknowns: Living Cultures (Pirum gemmata gen. nov., sp. nov., and Abeoforma whisleri, gen. nov., sp. nov.) from Invertebrate Digestive Tracts Represent an Undescribed Clade within the Unicellular Opisthokont Lineage Ichthyosporea (Mesomycetozoea). *Protist* **162**: 33–57.

Marshall, W.L., Celio, G., McLaughlin, D.J., and Berbee, M.L. (2008) Multiple Isolations of a Culturable, Motile Ichthyosporean (Mesomycetozoa, Opisthokonta), Creolimax fragrantissima n. gen., n. sp., from Marine Invertebrate Digestive Tracts. *Protist* **159**: 415–433.

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., Desantis, T.Z., Probst, A., et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J.* **6**: 610–618.

Méheust, R., Watson, A.K., Lapointe, F.J., Papke, R.T., Lopez, P., Bapteste, E. (2018). Hundreds

of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. *Genome Biol*. **19**: 75

Mendoza, L., Taylor, J.W., and Ajello, L. (2002) The Class Mesomycetozoea: A Heterogeneous Group of Microorganisms at the Animal-Fungal Boundary. *Annu. Rev. Microbiol.* **56**: 315–344.

Newman, M.E.J. (2003) Mixing patterns in networks. *Phys. Rev. E* **67**: 1–13.

Ocaña-Pallarès, E., Najle, S.R., Scazzocchio, C., and Ruiz-Trillo, I. (2019) Reticulate evolution in eukaryotes: Origin and evolution of the nitrate assimilation pathway. *PLOS Genet.* **15**: e1007986.

Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2018) CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. *Mol. Biol. Evol.* **35**: 252–255.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., et al. (2015) Open science resources for the discovery and analysis of Tara Oceans data. *Sci. Data* **2**: 150023.

Pilosof, S., Porter, M.A., Pascual, M., and Kéfi, S. (2017) The multilayer nature of ecological networks. *Nat. Ecol. Evol.* **1**: 0101.

Romari, K. and Vaulot, D. (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. *Limnol. Oceanogr.* **49**: 784–798.

RStudio, T. (2017) Rstudio: Integrated Development for R. *http://www.rstudio.com*.

Ruiz-Trillo, I., Burger, G., Holland, P.W.H., King, N., Lang, B.F., Roger, A.J., and Gray, M.W. (2007) The origins of multicellularity: a multi-taxon genome initiative. *Trends Genet.* **23**: 113–118.

Ruiz-Trillo, I., Inagaki, Y., Davis, L.A., Sperstad, S., Landfald, B., and Roger, A.J. (2004) Capsaspora owczarzaki is an independent opisthokont lineage. *Curr. Biol.* **14**: R946–R947.

Ruiz-Trillo, I., Roger, A.J., Burger, G., Gray, M.W., and Lang, B.F. (2008) A Phylogenomic Investigation into the Origin of Metazoa. *Mol. Biol. Evol.* **25**: 664–672.

Shalchian-Tabrizi, K., Minge, M.A., Espelund, M., Orr, R., Ruden, T., Jakobsen, K.S., and Cavalier-Smith, T. (2008) Multigene Phylogeny of Choanozoa and the Origin of Animals. *PLoS One* **3**: e2098.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., et al. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular

Interaction Networks. *Genome Res.* **13**: 2498–2504.

Thomsen, H.A., Garrison, D.L., and Kosman, C. (1997) Choanoflagellates (Acanthoecidae, Choanoflagellida) from the Weddell sea, Antarctica, taxonomy and community structure with particular emphasis on the ice biota; with preliminary remarks on Choanoflagellates from Arctic sea ice (Northeast Water Polynya, G. *Arch. fur Protistenkd.* **148**: 77–114.

Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., and Ruiz-Trillo, I. (2012) Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Mol. Biol. Evol.* **29**: 531–544.

Torruella, G., de Mendoza, A., Grau-Bové, X., Antó, M., Chaplin, M.A., del Campo, J., et al. (2015) Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Curr. Biol.* **25**: 1–7.

Valverde, S., Piñero, J., Corominas-Murtra, B., Montoya, J., Joppa, L., and Solé, R. (2018) The architecture of mutualistic networks as an evolutionary spandrel. *Nat. Ecol. Evol.* **2**: 94–99.

de Vargas, C., Stephane, A., Nicolas, H., Johan, D., Frederic, M., Ramiro, L., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* **348**: 1–12.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., and Zemla, J. (2017) corrplot: Visualization of a Correlation Matrix. *https://github.com/taiyun/corrplot*.

FIGURES

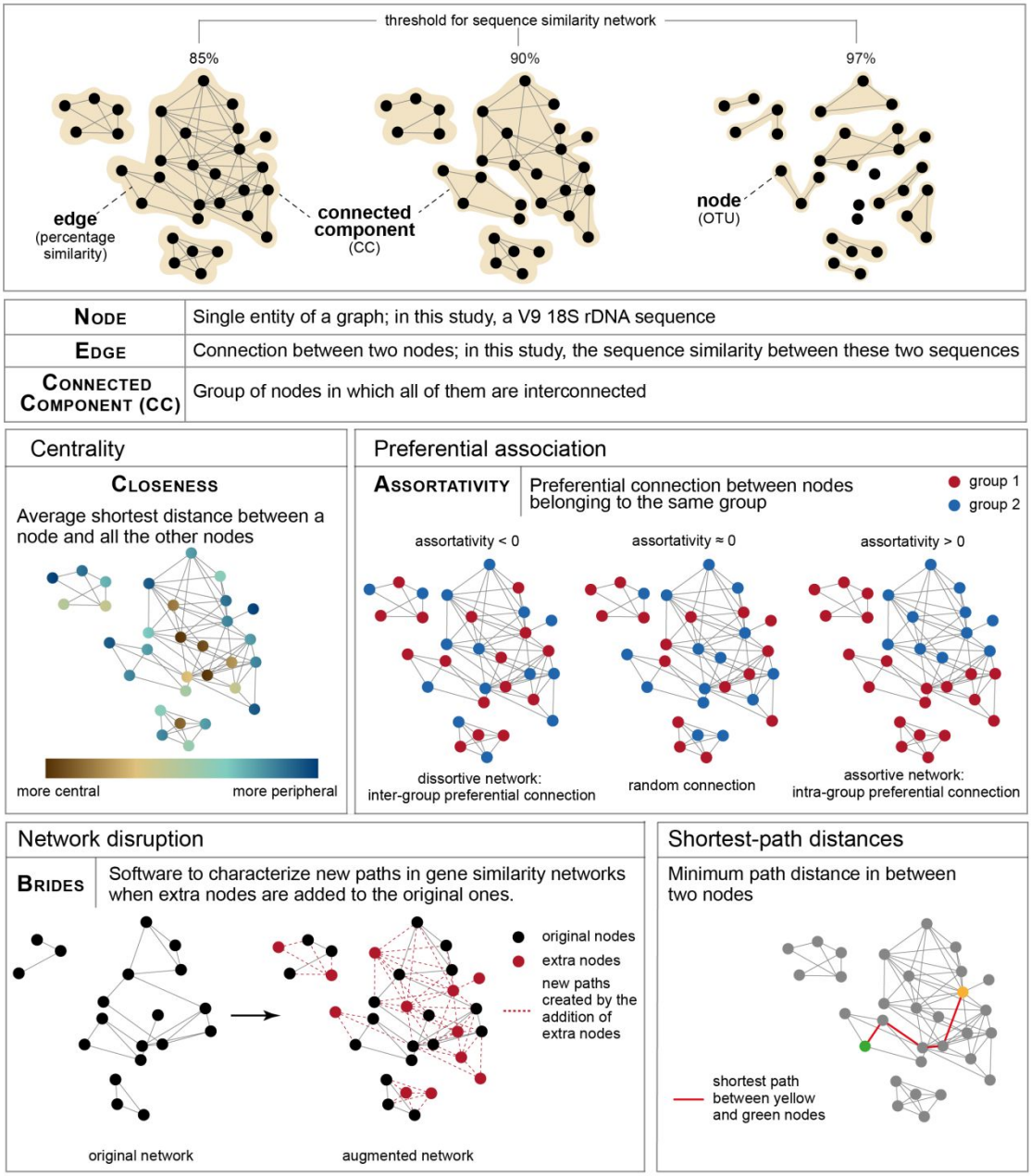

Figure 1

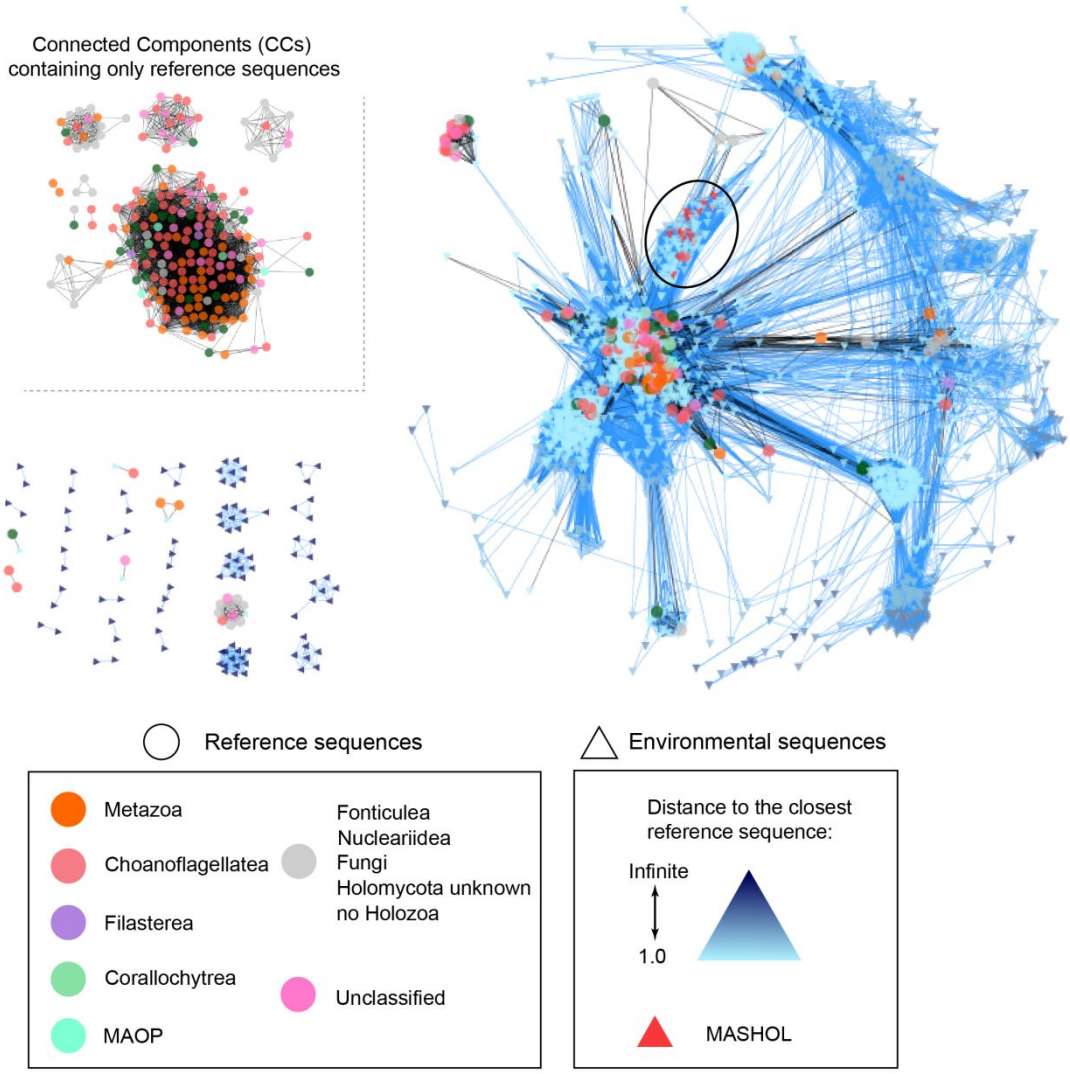

Figure 2

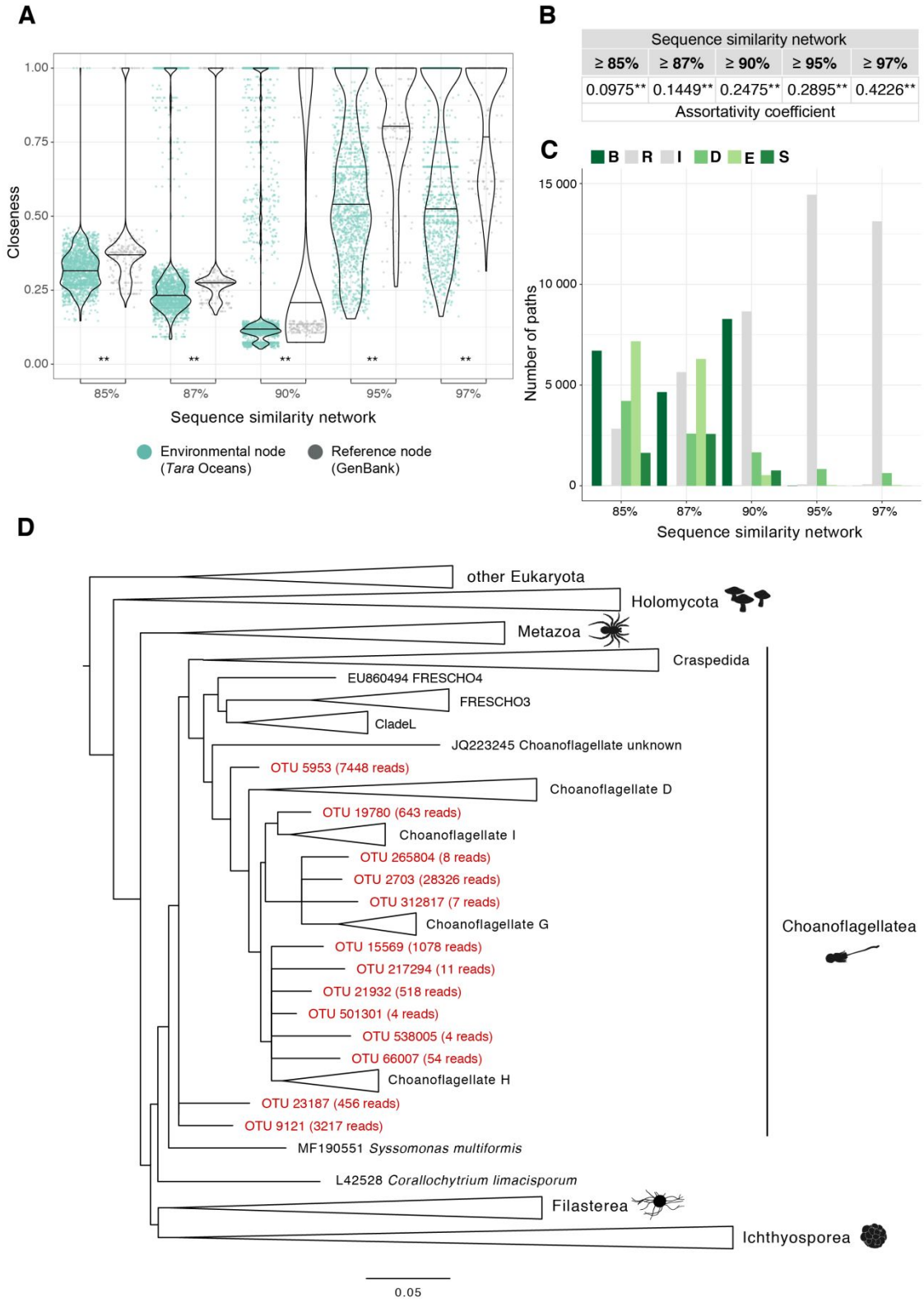

Figure 3

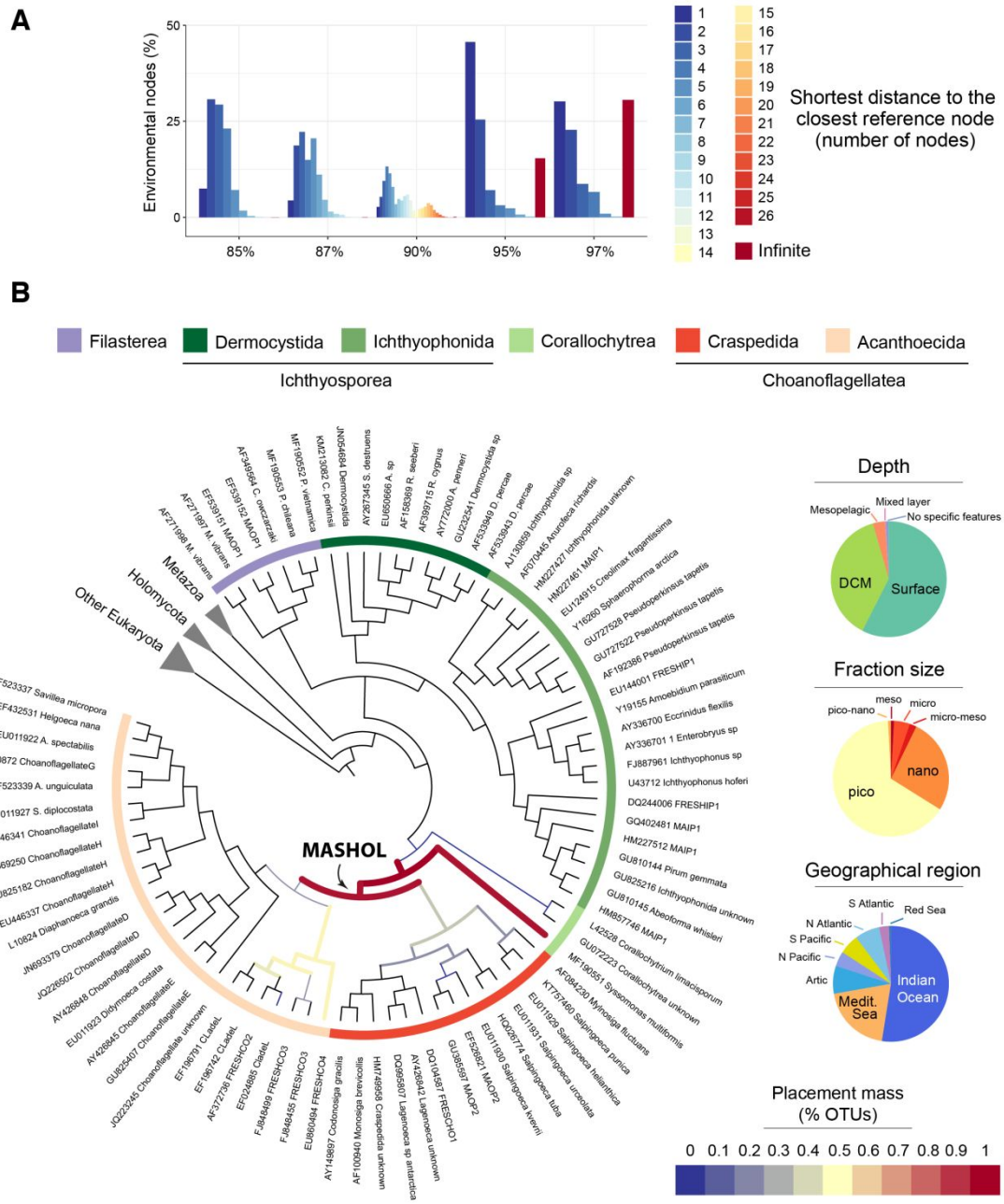

Figure 4

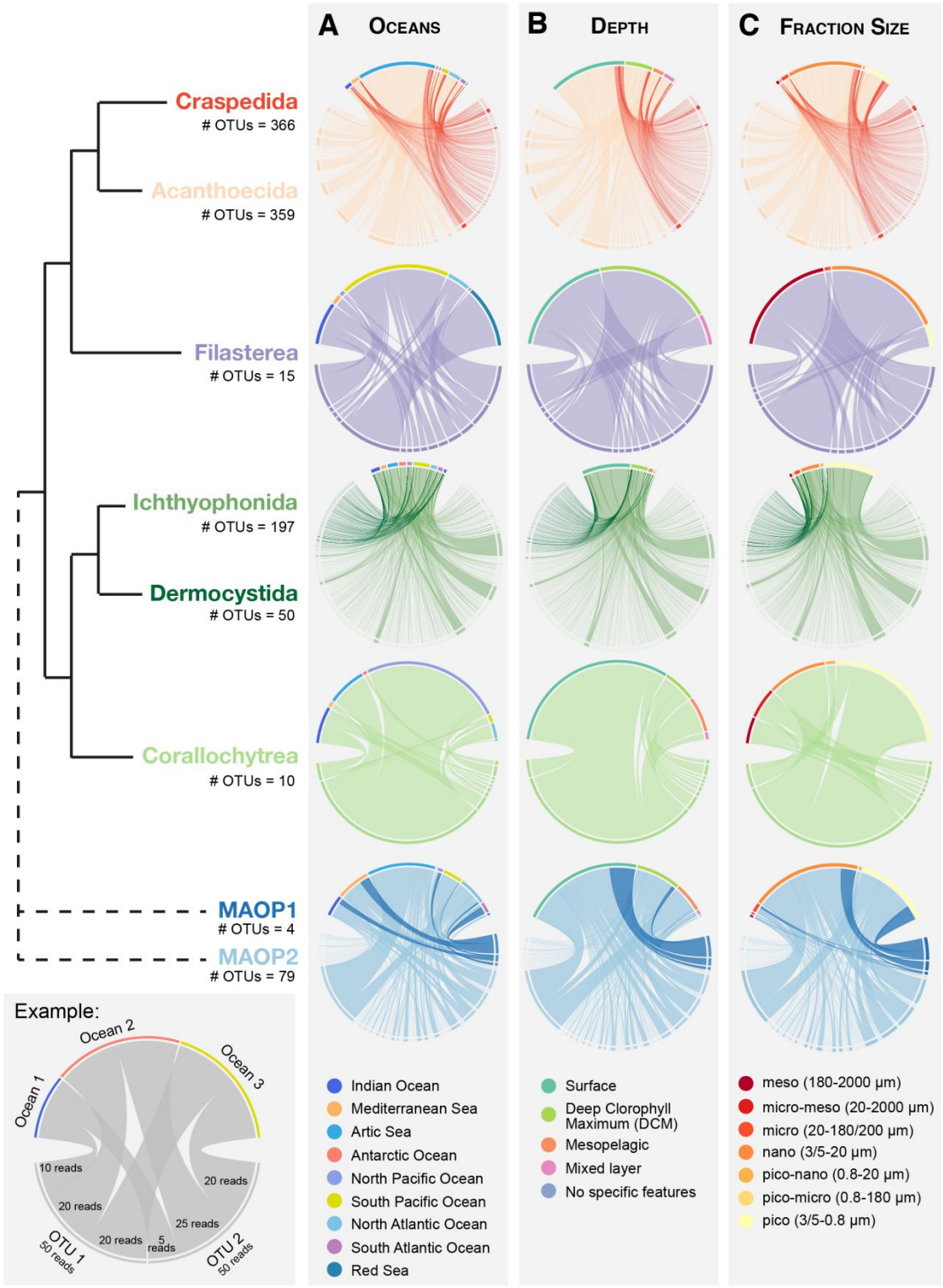

Figure 5

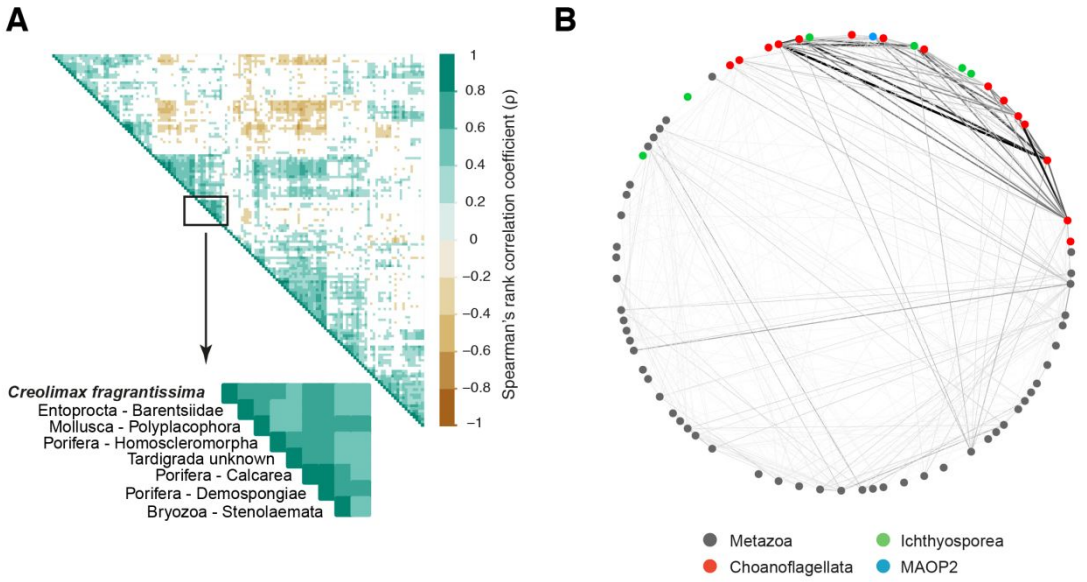

Figure 6