Determinants of community structure in the global plankton interactome.

AUTHOR(S):

Lima-Mendez, Gipsi; Faust, Karoline; Henry, Nicolas; Decelle, Johan; Colin, Sébastien; Carcillo, Fabrizio; Chaffron, Samuel; ... Bowler, Chris; de Vargas, Colomban; Raes, Jeroen

CITATION:

Lima-Mendez, Gipsi ...[et al]. Determinants of community structure in the global plankton interactome.. Science 2015, 348(6237): 1262073.

ISSUE DATE:

2015-05-22

URL:

http://hdl.handle.net/2433/197952

Title: Top-down determinants of community structure in the global plankton interactome

Authors: Gipsi Lima-Mendez[1,2,3,†], Karoline Faust[1,2,3,†], Nicolas Henry[4,5,†], Johan Decelle[4,5], Sébastien Colin[4,5,6], Fabrizio Carcillo[2,3,7], Samuel Chaffron[1,2,3], J. Cesar Ignacio-Espinosa[8], Simon Roux[8], Flora Vincent[2,6], Lucie Bittner[4,5,6], Youssef Darzi[2,3], Jun Wang[1,2], Stéphane Audic[4,5], Léo Berline[9,10], Ana M. Cabello[11], Laurent Coppola[9,10], Francisco M. Cornejo-Castillo[11], Francesco d'Ovidio[12], Luc De Meester[13], Isabel Ferrera[11], Marie-José Garet-Delmas[4,5], Lionel Guidi[9,10], Elena Lara[11], Stéphane Pesant[14,15], Marta Royo-Lonch[11], Guillem Salazar[11], Pablo Sánchez[11], Marta Sebastian[11], Caroline Souffreau[13], Céline Dimier[4,5,6], Marc Picheral[9,10], Sarah Searson[9,10], Stefanie Kandels-Lewis[16], *Tara* Oceans coordinators[‡], Gabriel Gorsky[9,10], Fabrice Not[4,5], Hiroyuki Ogata[17], Sabrina Speich[18,19], Jean Weissenbach[20,21,22], Patrick Wincker[20,21,22], Gianluca Bontempi[7], Silvia G. Acinas[11], Shinichi Sunagawa[16], Peer Bork[16], Matthew B. Sullivan[8], Chris Bowler[6,*], Eric Karsenti[6,16,*], Colomban de Vargas[4,5,*] and Jeroen Raes[1,2,3,*].

Affiliations:

[1]Department of Microbiology and Immunology, Rega Institute KU Leuven, Herestraat 49, 3000 Leuven, Belgium.
[2]VIB Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium.
[3]Laboratory of Microbiology, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.
[4]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
[5]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France.
[6] Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France.
[7]Interuniversity Institute of Bioinformatics in Brussels (IB)[2], ULB Machine Learning Group, Computer Science Department, Université Libre de Bruxelles.
[8]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA.
[9]CNRS, UMR 7093, LOV, Observatoire océanologique, 06230, Villefranche/mer, France.
[10]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire océanologique, 06230, Villefranche/mer, France.
[11]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Pg. Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain.
[12]Sorbonne Universités, UPMC, Univ Paris 06, CNRS-IRD-MNHN, LOCEAN Laboratory, 4 Place Jussieu, 75005, Paris, France.
[13]KU Leuven, Laboratory of Aquatic Ecology, Evolution and Conservation, Charles Deberiotstraat 32, 3000 Leuven.
[14]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Hochschulring 18, 28359 Bremen, Germany.
[15]MARUM, Center for Marine Environmental Sciences, University of Bremen, Hochschulring 18, 28359 Bremen, Germany.
[16]Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany.
[17]Institute for Chemical Research, Kyoto University, Gokasho, Uji, 611-0011 Kyoto, Japan.
[18]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France.
[19]Laboratoire de Physique des Océan, UBO-IUEM, Palce Copernic, 29820 Polouzané, France.
[20]CEA, Genoscope, 2 rue Gaston Crémieux, 91000 Evry France.
[21]CNRS, UMR 8030, 2 rue Gaston Crémieux, 91000 Evry, France.
[22]Université d'Evry, UMR 8030, CP5706 Evry, France.

[‡]*Tara* Oceans coordinators and affiliations are listed below.

[†]These authors contributed equally to this work

*Correspondence to: jeroen.raes@vib-kuleuven.be, vargas@sb-roscoff.fr,

cbowler@biologie.ens.fr, karsenti@embl.de

Abstract: Reconstructing global species interaction networks and identifying the abiotic and biotic factors that shape them are fundamental yet unsolved goals in ecology. Here, we integrate multi-kingdom organismal abundances and rich environmental measures from *Tara* Oceans and find that environmental factors are incomplete predictors of community structure. To study biotic effects, we reconstructed the first global photic-zone co-occurrence network. Interactions are non-randomly distributed across plankton functional types and phylogenetic groups, and show both local and global patterns. Known and novel interactions were identified among grazers, primary producers, viruses and (mainly parasitic) symbionts. We show how network-generated hypotheses guide confocal microscopy analyses towards discovery of symbiotic relationships. Together, this effort provides a foundational resource for ocean food web research and integrating biological components into ocean models.

One Sentence Summary: A species interaction network from the global ocean shows novel insights in top-down effects on community structure.

Introduction

The structure of oceanic ecosystems result from the complex interplay between resident organisms and their physico-chemical environment. In the world's largest ecosystem, oceanic plankton (composed of viruses, prokaryotes, microbial eukaryotes, and zooplankton) form intricate and dynamic trophic and symbiotic interaction networks (*1-4*) that are also influenced by environmental conditions. Ecosystem structure and composition are governed by abiotic as well as biotic control. The former includes environmental conditions and nutrient availability (*5*), while the latter encompasses grazing, pathogenicity and parasitism (*6, 7*). Like in terrestrial and intertidal ecosystems, determining the relative importance of both processes represents a grand challenge in ecology (*5*), but overall, abiotic effects have historically been considered to be the factor most strongly determining community structure (*8*). The challenge is to establish a quantitative understanding of biotic and abiotic interactions in natural systems where the organisms are taxonomically and

trophically diverse (*9*). Although experimental methods were developed to detect interactions encompassing virus-host associations (*10-13*) and competition and cooperation among bacteria (*14*), they are not sufficiently high-throughput yet to be applied community-wide in natural systems. However, sequencing technologies are now enabling community profiling across trophic levels, organismal sizes, and geographic ranges, providing the opportunity to predict organismal interactions across entire biomes based on co-occurrence patterns (*15*). Previous bioinformatics efforts addressing these issues have provided insights on the structure (*16, 17*) and dynamics of microbial communities at specific locations or organismal domains (*18-20*).

Here we analyze data from 313 plankton samples the *Tara* Oceans expedition (*21*) derived from 7 size-fractions covering collectively 68 stations at 2 depths across 8 oceanic provinces (Table S1), spanning organisms from viruses to small metazoans. For these samples, viral (*13*), prokaryotic and eukaryotic abundance profiles were derived from clusters of metagenomic contigs, $_{mi}$tags (*22*) and 18S rDNA V9 metabarcodes, respectively (*9, 23, 24*) (Table S1). In addition, rich environmental data from on-site and satellite measurements were collected (*21, 25, 26*). On this dataset, network inference methods and machine learning techniques are leveraged to disentangle biotic and abiotic signals shaping ocean plankton communities, and to construct a global-ocean cross-kingdom species interaction network (interactome). The interactome is then used to explore top-down relationships in the photic zone and validated using microscopic investigation of host/symbiont pairs and *in silico* analysis of phage-host pairings.

Evaluating the effect of abiotic and biotic factors on community structure

Given the breadth of the dataset we first re-assessed the effects of environment and geography on community structure. Using variation partitioning (*27*) we found that on average, the percentage of variation in community composition explained by environment alone was 18%, by environment combined with geography 13%, and by geography alone only 3% (*28*);(*29*). In addition, we built random forest-based models (*30*) to predict abundance profiles of the Operational Taxonomic Units (OTU) using a) OTUs alone, b) environmental variables alone, c) OTUs and environmental variables combined, and tested for each OTU whether one of the three approaches outcompeted the other (see Methods). These analyses revealed that 95% of the OTU-

only models are more accurate in predicting OTU abundances than environmental variable models, while combined models were no better than the OTU-only models (*31*);(*32*). This suggests that, unlike previously assumed (*8*), abiotic factors have a limited effect on community structure.

To study the role of biotic interactions, we developed a method to identify robust species associations in the context of environmental conditions. Twenty-three taxon-taxon and taxon-environment co-occurrence networks were constructed based on 9,292 taxa, representing the combinations of two depths, seven organismal size ranges and four organismal domains (Bacteria, Archaea, Eukarya, viruses) (*33*). To reduce noise and thus false positive predictions, we restricted our analysis to taxa present in at least 20% of the samples and used conservative statistical cutoffs (see Methods). A global network was obtained by performing the union of the individual networks. This network features a total of 127,995 unique edges, of which 92,633 are taxon-taxon edges and 35,362 are taxon-environment edges (Table 1). Node degree does not depend on the abundance of the node (OTU) (*33*). As such this network represents a novel, extensive resource to examine species associations in the global oceans (*33-36*).

Next, we assessed how many of the observed taxon links were indirect associations representing 'niche effects' driven by geographic or environmental parameters (i.e., associations between taxa that are only due to a common response to an environmental condition (*15*)). Motifs consisting of two correlated taxa that also correlate with at least one common environmental parameter ("environmental triplets") were examined using three approaches (interaction information, sign pattern analysis, and network deconvolution (*37*)) to identify associations that were driven by environment (*32, 34*); 27,868 such taxon-taxon-environment associations (30% of total) were detected. Among environmental factors, we found that PO_4 , temperature, NO_2 and mixed layer depth were frequent drivers of network connections (Figure 1A). Notably, while the three methodologies pinpoint indirect associations, only interaction information directly identifies synergistic effects in these biotic-abiotic triplets. Exploiting this property, we disentangled the 27,868 environment-affected associations into 8,961 edges driven solely by abiotic factors (*38*) (excluded from the network for the remainder of the study) and 18,907 edges whose dependencies result

from biotic-abiotic synergistic effects. This revealed that a minority of associations can be partly or completely explained by an environmental factor.

Evaluation of predicted interactions

Because co-occurrence techniques were thus far applied principally to bacteria, we assessed the sensitivity of the approach for detecting eukaryotic interactions based on V9 rDNA metabarcodes. We created, through extensive literature searches, a list of 573 known symbiotic interactions *sensu lato* (i.e., parasitism and mutualism) in marine eukaryotic plankton (*36, 39*). We extracted 42 genus level interactions for which both partners (OTUs) were present in the abundance pre-processed input matrices, and found that 40.5 % of these were predicted, and up to 49 % when only parasitism was considered – a considerable number, given the fact that the list is based on interactions that are from other locations and potentially transitive or facultative. The probability of having found each of these interactions by chance alone was <0.01 (Fisher exact test, average pval = 4e-3, median pval = 5e-7). Most of the false negative interactions were due to the strict filtering rules we determined to avoid false positives. Based on this sensitivity and a false discovery rate averaging to 9% (computed from null models; see Methods), we estimate the lower and upper limits for the number of interactions among eukaryotes present in our filtered input matrices to be 55,000 and 150,000.

Biotic interactions within and across kingdoms

We next focused on the integrated network containing 83,672 predicted biotic interactions (*31*) (*36*) that were non-randomly distributed within and between size fractions (Figure 1B, C) (*40*). Copresences (positive associations) outnumbered mutual exclusions (anticorrelations; 73% versus 27%), and a non-random edge distribution with regard to phylogeny was observed (Figure 2A), with most copresences derived from syndiniales and other dinoflagellates, and exclusions involving arthropods. On higher taxonomic ranks (e.g., Order), we found that although taxonomically related groups do co-occur (2,500 associations within the same order; (*15, 16*)), 32% (1,157) was found across different orders (*38, 41*). Certain combinations of phylogenetic groups are over-represented. For instance, a clade of syndiniales (the MALV-II Clade 1 belonging to Amoebophrya (*3*)) shows a significant enrichment in positive associations with tintinnids (P = 2e-4), amongst the

most abundant ciliates in marine plankton (*42*). Although this tintinnid/syndiniales interaction is not mentioned in current plankton ecology studies, the tintinnid *Xystonella lohmani* was described in 1964 to be infected by *Amoebophrya tintinnis* (*43*) and tintinnids can feed on *Amoebophrya* free-living stages (*44*). Other found host-parasite associations included the copepod parasites *Blastodinium*, *Ellobiopsis* and *Vampyrophrya* (*43, 45-47*).

On the other hand, *Maxillopoda*, anticorrelating with *Bacillariophyceae* and collodarians, contribute a large number of the negative associations (*38*), three groups of relatively large species whose biomass can dominate planktonic ecosystems. Collodarians and copepods occupy the same size range in, respectively, the oligotrophic tropical and eu/meso-trophic temperate systems (*9, 48*). The decoupling of phyto- and zooplankton in open oceans by diatoms anti-correlating to copepods (*49, 50*) is classically attributed to growth rate differences and to the diatom production of compounds harmful to their grazers (*51*). The combination of these effects could lie at the basis of this observation, which contrasts with other free-living autotrophs represented in the network (cyanobacteria and prymnesiophytes), which display primarily positive associations (Figure 2A).

Cross-kingdom associations between Bacteria and Archaea were limited to 24 mutual exclusions. Within Archaea, Thermoplasmatales (Marine Group II) co-occur with several phytoplankton clades. Links between Bacteria and protists recovered 5 out of 8 recently discovered interactions from protist single-cell sequencing (*52*). Significant copresences between Diatoms and Flavobacteria agreed with their described symbioses (*53*). We also observed co-ocurrence of uncultured dinoflagellates with members of Rhodobacterales (*Ruegeria*), in agreement with a symbiosis between *Ruegeria* sp. TM1040 and *Pfiesteria piscicida* around the ability of *Ruegeria* to metabolize dinoflaggelate-produced DMSP (*54*).

Global versus local associations

We further investigated whether our network was driven by global trends (e.g., whether species co-occur across oceanic regions) or is mainly local and limited to specific interaction ‘hotspots.’ To this aim, we divided our set of samples into 7 main regions: Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic (SAO), Southern Ocean (SO), South Pacific Ocean (SPO) and North Atlantic Ocean

(NAO), and assessed the 'locality' of associations by comparing the score with or without that region (see Methods). We found that association patterns were mostly driven by global trends as only 15% of edges were identified as local (Figure 2B, C). Approximately two thirds of local associations occur in MS (8,371) followed by SPO (1,119), while the rest are contributed by IO (946), with SO (901), SAO (123) and RS (891), and NAO (60) (Figures 2C-G). One should note that MS was the region with most sampling sites, which allowed us to recover more local patterns. Nevertheless, Figures 2C-G show that although the same major groups (order level) interact in both the global and local networks, each local site has its own specific interaction profile (Pval < 1e-8) (*38, 41, 55*).

Parasite impact on plankton functional types

Our approach being particularly suitable for predicting parasitic interactions, we assessed their potential impact on biogeochemical processes by exploring a functional sub-network (22,223 edges) of known and novel plankton parasites (*9*) together with classical 'plankton functional types' (PFTs (*56*)). PFTs group taxa by trophic strategy (e.g., autotrophs vs. heterotrophs) and role in ocean biogeochemistry (Figure 3A)(*39, 57*). Overall, the tight relationship between the different PFTs (network density of 0.71) highlights strong dependencies between phytoplankton and grazers. Furthermore, we find that PFTs are universally, but non-homogeneously associated to parasites. Most links involve syndiniales MALV-I and MALV-II clades associated to zooplankton and, to a lesser extent, to microphytoplankton (excluding diatoms). This emphasizes the important role of alveolate parasitoids as top-down affectors of zooplankton and microphytoplankton population structure and functioning (*3*) - although the latter group is also affected by grazing (*1*). The meso-planktonic networks contain several known syndiniales targets (Dinophyceae, Ciliophora, Acantharia and Metazoa; Figure 3B) (*58*). In large size fractions, interactions between known parasites and groups of organisms that in theory are too small to be their hosts were found (*59*); 32% of these associations involved the abundant and diverse marine stramenopiles (MAST) and diplonemids (other Discoba and Diplonema) (*9*). Although their eco-physiology is under-investigated, previous studies (*60, 61*) suggest a parasitic role for these lineages. The association of these groups with other parasites would be explained by putative co-infection of the same hosts. Contrasting with the above observations, we find phytoplankton silicifiers (diatoms) displaying a

significant number of mutual exclusions. One possible interpretation of this is that diatom silicate exoskeletons (*62*) and toxic compound production (*51*) could act as efficient barriers against top-down pressures (*63*).

Phage-microbe associations

We investigated phage-microbe interactions, another major top-down process affecting global bacterial/archaeal community structure (*7*). A major challenge in this area is to link viruses to their hosts (*64*). Here, surface and deep chlorophyll maximum virus-bacteria networks revealed 1,479 positive associations between viral populations and six of the 54 known bacterial phyla (specifically – Proteobacteria, Cyanobacteria, Actinobacteria, Bacteroidetes, Deferribacteres and Verrucomicrobia), and one archaeal phylum (Euryarchaeota). These 7 phyla represent most of abundant bacterial/archaeal groups across 43 investigated samples (Figure 3C), suggesting that the networks are detecting abundant virus-host interactions. Additionally, these interactions include phyla of microbes lacking viral genomes in RefSeq databases including Verrucomicrobia, and non-extremophile Euryarchaeota, hinting at some of the first viral genomic sequences for important yet understudied phyla (Figure 3D) (*41, 65, 66*).

In addition, these data help in more comprehensively evaluating viral "host range" breadth, fundamental for predictive modeling and thus far largely limited to observations of cultured virus-host systems that insufficiently map complex community interactions (*64*). While not without caveats, these virus-host interaction data suggest that viruses are very host specific with ~45% of the phage populations interacting with only a single host OTU, and the remaining 55% interacting with no more than a few, often closely related OTUs (Figure 3C; robustness tested using simulations, see Methods). Further, these networks are modular at large scales and nested within sub-modules at smaller scales (*67*), suggesting that viruses are host-range-limited across large sections of host space (the modularity), but that specialist and generalist phages prey on specific groups within sub-sections of this available host space (the nestedness). One should note that the eukaryotic parasite-host networks also show nestedness and indications for modularity, suggesting that similar rules apply across kingdoms. While these co-occurrence networks have the scale needed to more fully represent natural communities, they will obviously require

experimental validation to confirm predicted host-phage associations. Recent experimental (*10-12, 68*) and informatic (*13, 69*) innovations should help in this regard once these technologies (i.e. experimental protocols) are operational and required data (i.e. complete genomes or large genome fragments) are available (*70*).

Microscopic validation of predicted interactions

Finally, we assessed whether our global interactome can also be used to guide observation-based discovery of symbiotic interactions. Specifically, a predicted putative photosymbiotic interaction between an acoel flatworm (*Symsagittifera* sp.) and a green microalga (*Tetraselmis* sp.), was validated experimentally by combined Laser Scanning Confocal Microscopy (LSCM), 3D reconstruction, and reverse molecular identification on flatworm specimens isolated from *Tara* Oceans preserved morphological samples (see Materials and Methods). Using this approach, we observed numerous microalgal cells (5 to 10 µm diameter) within each of the 15 isolated acoel specimens (Figure 4);(*71*). The 18S sequence from several sorted holobionts matched the metabarcode pair identified in the co-occurrence global network. These results demonstrate that the combination of molecular ecology, microscopy and bioinformatics provide a powerful toolkit to unveil key symbioses in marine ecosystems.

Conclusions

Unraveling the global ocean interactome remains a grand challenge for developing predictive understanding of the dynamics and structure of ocean ecosystems. The interactome, reported here, spanning all three organismal domains and viruses, provides a foundational global ocean dataset towards this aim. The analyses presented place new emphasis on the role of top-down biotic interactions in the epipelagic zone, and present myriad hypotheses that will guide future research to understand how symbionts, pathogens, predators and parasites interact with their target organisms, and ultimately help elucidate the structure of the global food webs that drive nutrient and energy flow in the ocean.

Methods

Sampling

The sampling strategy used in the Tara Oceans expedition is described in (*72*) and samples used in the present study are listed in Table S1. The *Tara* Oceans nucleotide sequences are available at the European Nucleotide Archive (ENA) under project ID PRJEB402 (http://www.ebi.ac.uk/ena/data/view/PRJEB402).

Physical and environmental measurements

Physical and environmental measurements were carried out with a vertical profile sampling system (CTD-Rosette) and data collected from Niskin bottles. We measured temperature, salinity, chlorophyll, CDOM fluorescence (fluorescence of the colored dissolved organic matter), particles abundance, nitrate concentration and particle size distribution (using an Underwater Vision Profiler). In addition, mean mixed layer depth (MLD), maximum fluorescence, vertical maximum of the Brünt-Väisälä Frequency N (s−1), vertical range of dissolved oxygen, and of change of nitrates were determined. Satellite altimetry provided the Okubo-Weiss parameter, Lyapunov exponent, mesoscale eddie retention and sea surface temperature (SST) gradients at eddie fronts (*24*). Data are available at http://www.pangaea.de.

Abundance table construction

Prokaryotic 16S rDNA metagenomic reads were identified, annotated and quantified from Illumina sequenced metagenomes (hereafter $_{mi}$tags) as described in (*22*) using the SILVA v.115 database (*24, 73, 74*). The abundance table was normalized using the summed read count per sample (*24, 75*). Quality-checked V9 rDNA metabarcodes were clustered into swarms as in (*9, 76*), and annotated using the V9 PR2 database (*77*). PR2 barcodes were associated to fundamental trophic modes (auto- or hetero-trophy) and symbiotic interactions (parasitism and mutualism) based on literature. Swarm abundance and normalization was performed as in (*9, 76*). Bacteriophage metagenomes were obtained from the < 0.2µm fractions for 48 samples and contigs were annotated and quantified as in (*23*). The abundance matrix was normalized by total sample read count and contig length.

In all cases, only OTUs with relative abundance > 1e-8 and detected in at least 20 % of samples were retained. Since sample number in the input tables ranged from 17 to

63, prevalence thresholds varied (from 22% to 40%). The sum of all filtered OTU relative abundances was kept in the tables to preserve proportions. Abundance tables are available at http://psbweb07.psb.ugent.be/raeslab/supplemental/taraoceans.html

Random forest-based models

Eukaryotic, prokaryotic and environmental matrices were merged into two matrices (Deep Chlorophyll Maximum layer (DCM) and surface water layer (SRF)). For each of the three models (OTU *versus* other OTUs (M_{OTU}), environmental factors (M_{ENV}) or combined ($M_{OTU+ENV}$)), regressions were perfomed with OTU abundance as dependent and the abundances of other OTUs or environmental factors as independent variables. For each regression, up to 20 independent variables were selected using the minimum Redundancy Maximum Relevance (mRMR) filter-ranking algorithm. Regressions were performed using random forests (*30*) and followed by a leave-one-out cross-validation. The variable subset with the minimum leave-one-out NMSE (Normalized Mean Square Error) was selected. To identify the best model for a given target OTU, the significance of the NMSE difference was tested on the absolute error values (paired Wilcoxon test adjusted by Benjamini-Hochberg FDR estimation (*78*)). NMSE computed on random data are larger than those from original data. In addition, M_{ENV} outperformed M_{OTU} when OTU abundances were randomized.

Variance partitioning

Environmental variables were z-score-transformed; spatial variables (MEM eigenvectors) were calculated based on latitude and longitude (*79*). Forward selection (*80*) was carried out with function forward.sel in R-package packfor. Significance of the selected variables was assessed with 1000 permutations using functions rda and anova.cca in vegan. Variance partitioning (*81*) was performed using function varpart in vegan on Hellinger-transformed abundance data, the forward-selected environmental variables, and the forward-selected spatial variables and tested for significance with 1000 permutations.

Network inference

Taxon-taxon co-occurrence networks were constructed as in (*15*), selecting Spearman and Kullback-Leibler dissimilarity and discarding any edge not supported by both measures. Measure-specific p-values were merged using Brown's method (*82*) and

multiple-testing-corrected with Benjamini-Hochberg (*78*). Edges with p-values above 0.05 were discarded.

Taxon-environment networks were computed with the same procedure, starting with 8,000 initial positive and negative edges, each supported by both methods. For computational efficiency, we computed 23 taxon-taxon and taxon-environment networks separately, for two depths (DCM and SRF), four eukaryotic size fractions (0.8-5 µm, >0.8 µm, 20-180 µm and 180-2000 µm) and their combinations, the prokaryotic size fraction (0.2-1.6µm and 0.2-3.0µm) and its combination with each of the eukaryotic and virus (< 0.2 µm) size fractions. We then generated 23 taxon-environment union networks for environmental triplet detection and merged the taxon-taxon networks into a global network with 92,633 edges.

Estimation of false discovery rate

We estimated the false discovery rate (FDR) of network construction with two null models. The first shuffles counts while preserving total OTU count sums. For the second, we fitted a Dirichlet-Multinomial distribution to the input matrix using the dirmult package in R (*83*) and generated a null matrix by sampling from this distribution, preserving total sample count sums. Null matrices were generated from count matrices (0.8-5µm, 20-180 µm and 180-2000 µm eukaryotic and prokaryotic size fraction SRF and DCM). Network construction was performed with the 16 null matrices and thresholds applied to the original matrices (*33*). From edge numbers in the original and the null networks, we estimated an average FDR of 9% (*33*).

Indirect taxon edge detection

For each taxon-environment union network, node triplets consisting of two taxa and one environmental parameter were identified. For each triplet, interaction information II was computed as: $II = CI(X,Y \mid Z) - I(X,Y)$, where CI is the conditional mutual information between taxa X and Y given environmental parameter Z and I the mutual information between X and Y. CI and I were estimated using minet (*84*). Taxon edges in environmental triplets were considered indirect when II<0 and within the 0.05 quantile of the random II distribution obtained by shuffling environmental vectors (500 iterations). If a taxon was part of more than one environmental triplet, the triplet with minimum interaction information was selected.

For each environmental triplet, we also checked whether its sign pattern (the combination of positive and/or negative correlations) was consistent with an indirect interaction. From 8 possible patterns, 4 indicate indirect relationships (e.g. two negatively correlated taxa correlated with opposite signs to an environmental factor).

Network deconvolution (*37*) was carried out with beta=0.9. We considered an environmental triplet as indirect according to network deconvolution if any of its edges were removed.

All (8,961) negative interaction information triplets were consistent with an indirect relationship according to their sign patterns and a majority (6,711) was also supported by network deconvolution.

Influence of ocean regions on co-occurrence patterns

Samples were divided into groups according to region membership. The impact of each sample group on the Spearman correlation of each edge in the network was assessed by dividing the (absolute) omission score (OS; Spearman correlation without these samples) by the absolute original Spearman score. To account for group size, the OS was computed repeatedly for random, same-sized sample sets. Nonparametric p-values were calculated as the number of times random OSs were smaller than the sample group OS, divided by number of random OS (500 for each taxon pair). Edges were classified as region-specific when the ratio of OS and absolute original score was below 1 and multiple-testing-corrected p-values (Benjamini-Hochberg) were below 0.05.

Over-representation analysis

Significance of taxon–taxon counts at high taxonomic ranks was assessed with the hypergeometric distribution implemented in the R function phyper, as was the enrichment of parasites in the network by comparing the intersection between the parasites in the network and those in the community (pre-filtered) matrix.

Mutual exclusion vs copresence analysis was performed using the binomial distribution implemented in the R function pbinom, with the background probability estimated by the frequency of edges in the network.

Oceanic region analysis was also assessed using R's pbinom function, with the background probablility estimated by dividing total ocean-specific edge number by total edge number. The p-value was computed as the probability of obtaining the observed number of ocean-specific edges among the edges of a taxon pair. The same

procedure was repeated for each oceanic region separately, with region-specific success probabilities. Edges classified as indirect were discarded before the analysis. P-values were adjusted for multiple testing according to Benjamini, Hochberg and Yekutieli (BY), implemented in the R function p.adjust.

Extracting functional groups from the global plankton interactome

Functional groups consist of a mix of major monophyletic lineages of parasites, together with classical polyphyletic Plankton Functional Types (PFT), as defined in (*56*) (*9*). (*57*)Metabarcodes in the network were sorted into 15 parasite groups and 7 PFTs (*57*) based on their *(i)* taxonomical classification, *(ii)* membership in a given size fraction, *(iii)* trophic mode, and *(iv)* biogeochemical role in DMS production or silicification. After mapping the metabarcodes and their edges onto PFTs and parasites, edges are weighted by the number of links they represent. Over-represention of the number of links included in each edge was assessed with the hypergeometric distribution.

Parasite links in large fractions may point to parasite-host connections. We extracted all edges in the large fractions (20-180µm and 180-2000µm) between barcodes annotated as parasites and non-parasitic barcodes. Partners of parasites comprised potential hosts (Fig 3B) but also organisms that are either too small or without size information. The former may represent unknown parasites (e.g. co-infecting a host with known parasites) while the latter may represent novel hosts.

Nestedness and modularity analysis

Nestedness was quantified with the NTC (nestedness temperature calculator) algorithm (*85*) implemented in nestedtemp in vegan and assessed for significance with 25 permutations (also confirmed using NTC in BiMAT (*86*) using 100 permutations). The analysis was carried out for 1,869 positively correlated phage-prokaryotic pairs and 3013 putative parasite-host interactions in the 20-180 µm and 180-2000 µm size fractions. Modularity was computed with the LP (Label propagation) BRIM algorithm (*87*) in BiMAT (*86*) with 100 permutations. Nestedness of the host-phage network as quantified with the NODF (nestedness with overlap and decreasing fill) algorithm (*88*) in BiMAT with 100 permutations was strongly significant. To assess the robustness of the results for the host-phage network, we tested the impact of random removal or addition of 5%, 10%, 15% and 20% edges.

Random addition of edges reduced modularity but not nestedness (according to both NODF and NTC), whereas random edge removal did neither change modularity nor nestedness scores.

Bacteriophage sequence screening in genomes of predicted hosts

Contigs of predicted hosts were compared (BLAST) to a set of viral sequences detected in draft and single-cell genomes. One virome contig (TARA_36DCM_3902) displayed significant similarity to a sequence from a single-cell genome available in WGS (AA160P02DRAFT_scaffold_31.32). The taxonomy of this single-cell genome was evaluated based on the affiliation of 16S rRNA detected with meta-rRNA (*89*) from this single-cell genome. The comparison of gene content between TARA and WGS SAG contigs was generated with Easyfig (*90*), and genes from these contigs were functionally annotated based on a BLAST comparison to the NCBI nr database.

Evaluation of predicted interactions

A list of 573 known symbiotic interactions *sensu lato* (i.e., parasitism and mutualism, at least 1 protist partner) in marine eukaryotic plankton, covering 197 eukaryotic genera, described in 76 publications since 1971 was compiled in three steps: First, we manually screened publications linked to each PR2 db ((*91*);3170 publications) for marine eukaryotic phytoplankton interactions. Second, we screened 293 publications retrieved from Web of Science with the query: 'TOPIC:(plankton* AND (marin* OR ocean*)) AND (parasit* OR symbios* OR mutualis*)'. Finally, we screened GenBank 18S rDNA sequences of symbionts for which the 'host' field was known. We labeled these interactions as 'Unpublished'.

Experimental validation of a predicted interaction

V9 pairs were searched for organisms of suitable size to allow its isolation from morphological samples. This way, we targeted a predicted photosymbiosis between an acoel flatworm (V9 rDNA metabarcode 83% similar to *Symsagittifera psammophila (92))* and a photosynthetic microalga (*Tara* Oceans V9 metabarcode 100% similar to a *Tetraselmis* sp) (*93*)

Fifteen acoel specimens (hosts) were isolated from formaldehyde-4% microplankton samples of station 22 (A100000458), where both partner OTUs displayed high abundances. Prior to imaging, specimens were rinsed with artificial seawater, then

DNA and membrane structures were stained for 60 minutes with 10µM Hoechst 33342 and 1.4µM DiOC6(3) (Life Technologies). Microscopy was conducted using a Leica TCS SP8 (Leica Microsystems) confocal laser scanning microscope and a HC PL APO 40x/1.10 W motCORR CS2 objective. The DiOC6 signal (ex488nm/em500-520nm) was collected simultaneously with the chlorophyll signal (ex488nm/em670-710nm), followed by the Hoechst signal (ex405 nm/em420-470nm). Images were processed with Fiji (*94*), and 3D specimens were reconstructed with Imaris (Bitplane). To obtain the sequences of the metabarcodes of each partner, seven acoels were isolated from ethanol-preserved samples from station 22 (TARA_A100000451), individually rinsed in filtered seawater, and stored at -20°C in absolute ethanol. DNA was extracted with MasterPureTM DNA/RNA purification kit (Epicenter) and PCR amplified using the universal-eukaryote primers (forward 1389F and reverse 1510R) from (*9*). Chlorophyte-specific primers (Chloro2F: 5'-CGTATATTTAAGTTGYTGCAG-3' and Tetra2-rev 5'-CAGCAATGGGCGGTGGC GAAC-3') were designed to amplify the microalgae V9 rDNA as in (*4*). Purified amplicons were subjected to poly-A reaction and ligated in pCR®4-TOPO TA Cloning vector (Invitrogen), cloned using chemically competent *E. coli* cells and Sanger-sequenced with the ABI-PRISM Big Dye Terminator Sequencing kit (Applied Biosystems) using the 3130xl Genetic Analyzer, Applied Biosystems.

References.

1. F. T. Azam F, *Marine Ecology Progress Series*, (1983)
2. A. W. Thompson, R. A. Foster, A. Krupke, B. J. Carter, N. Musat *et al.*, *Science* **337**, 1546 (2012)
3. A. Chambouvet, P. Morin, D. Marie, L. Guillou, *Science* **322**, 1254 (2008)
4. J. Decelle, I. Probert, L. Bittner, Y. Desdevises, S. Colin *et al.*, *Proc Natl Acad Sci U S A* **109**, 18000 (2012)
5. V. Smetacek, *J Biosci* **37**, 589 (2012)
6. J. L. Sabo, L. R. Gerber, "Trophic ecology."
7. F. Rohwer, R. V. Thurber, *Nature* **459**, 207 (2009)
8. P. G. Verity, V. Smetacek, in *Marine Ecology-Progress Series*. (1996), vol. 130, pp. 277.
9. C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé *et al.*, *Science* **Submitted**, (2014)
10. E. Allers, C. Moraru, M. B. Duhaime, E. Beneze, N. Solonenko *et al.*, *Environ Microbiol* **15**, 2306 (2013)
11. A. D. Tadmor, E. A. Ottesen, J. R. Leadbetter, R. Phillips, *Science* **333**, 58 (2011)
12. L. Deng, J. C. Ignacio-Espinoza, A. C. Gregory, B. T. Poulos, J. S. Weitz *et al.*, *Nature* **513**, 242 (2014)
13. S. Roux, A. K. Hawley, M. Torres Beltran, M. Scofield, P. Schwientek *et al.*, *eLife*, (2014)
14. D. Sher, J. W. Thompson, N. Kashtan, L. Croal, S. W. Chisholm, *ISME J* **5**, 1125 (2011)
15. K. Faust, J. Raes, *Nat Rev Microbiol* **10**, 538 (2012)
16. S. Chaffron, H. Rehrauer, J. Pernthaler, C. von Mering, *Genome Res* **20**, 947 (2010)
17. J. Raes, I. Letunic, T. Yamada, L. J. Jensen, P. Bork, *Mol Syst Biol* **7**, 473 (2011)
18. J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbruck, J. Reeder *et al.*, *ISME J* **6**, 298 (2012)
19. J. M. Beman, J. A. Steele, J. A. Fuhrman, *ISME J* **5**, 1077 (2011)

20. C.-E. T. Chow, D. Y. Kim, R. Sachdeva, D. A. Caron, J. A. Fuhrman, *ISME J* **8**, 816 (2014)
21. E. Karsenti, S. G. Acinas, P. Bork, C. Bowler, C. De Vargas *et al.*, *PLoS Biol* **9**, e1001177 (2011)
22. R. Logares, S. Sunagawa, G. Salazar, F. M. Cornejo-Castillo, I. Ferrera *et al.*, *Environ Microbiol*, (2013)
23. J. R. Brum, J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas *et al.*, *Science* **submitted**, (2014)
24. S. Sunagawa, L. P. Coelho, S. Chaffron , J. R. Kultima, K. Labadie *et al.*, *Science* **Submitted**, (2014)
25. E. Villar, S. Audic, L. Bittner, B. Blanke, J. R. Brum *et al.*, *Science* **submitted**, (2014)
26. Supplementary table s2, http://www.raeslab.org/LimaMendez_etal_2014/SOM/ST2.xls
27. A. Meot, P. Legendre, D. Borcard, *Environmental and Ecological Statistics* **5**, 1 (1998)
28. Supplementary table s3, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST3.xls
29. Supplementary figure s1, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF1.pdf
30. L. Breiman, *Machine Learning* **45**, 5 (2001)
31. Supplementary table s4, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST4.xls
32. Supplementary figure s2, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF2.pdf
33. Supplementary table s5, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST5.xls
34. Supplementary figure s3, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF3.pdf
35. Supplementary table s6, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST6.xls
36. Supplementary table s7, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST7.xls
37. S. Feizi, D. Marbach, M. Medard, M. Kellis, *Nat Biotechnol* **31**, 726 (2013)
38. Supplementary table s8, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST8.xls
39. Supplementary table s9, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST9.xls
40. Supplementary figure s4, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF4.pdf
41. Supplementary table s10, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST10.xls
42. G. B. McManus, L. F. Santoferrara, in *The biology and ecology of tintinnid ciliates*. (John Wiley & Sons, Ltd, 2012), pp. 198.
43. J. Cachon, in *Ann sci nat b*. (Paris, 1964), vol. 6, pp. 1.
44. P. S. Salomon, Gran, xe, E. li, M. Neves *et al.*, *Aquatic Microbial Ecology* **55**, 143 (2009)
45. F. Gomez, P. Lopez-Garcia, A. Nowaczyk, D. Moreira, *Syst Parasitol* **74**, 65 (2009)
46. S. Ohtsuka, M. Hora, T. Suzaki, M. Arikawa, G. Omura *et al.*, *Marine Ecology Progress Series* **282**, 129 (2004)
47. A. Skovgaard, S. A. Karpov, L. Guillou, *Front Microbiol* **3**, 305 (2012)
48. L. Stemmann, M. Youngbluth, K. Robert, A. Hosia, M. Picheral *et al.*, *ICES Journal of Marine Science: Journal du Conseil* **65**, 433 (2008)
49. J. M. Gasol, P. A. Del Giorgio, C. M. Duarte, *Biomass distribution in marine planktonic communities*. (American Society of Limnology and Oceanography, Waco, TX, ETATS-UNIS, 1997), vol. 42.
50. B. A. Ward, S. Dutkiewicz, M. J. Follows, *Journal of Plankton Research* **36**, 31 (2014)
51. A. Ianora, A. Miralto, S. A. Poulet, Y. Carotenuto, I. Buttino *et al.*, *Nature* **429**, 403 (2004)
52. M. Martinez-Garcia, D. Brazel, N. J. Poulton, B. K. Swan, M. L. Gomez *et al.*, *ISME J* **6**, 703 (2012)
53. E. T. Jolley, A. K. Jones, *British Phycological Journal* **12**, 315 (1977)
54. T. R. Miller, R. Belas, *Appl Environ Microbiol* **70**, 3383 (2004)
55. Supplementary figure s5, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF5.pdf

56. C. Le Quere, Sandy P. Harrison, I. Colin Prentice,, O. A. Erik T. Buitenhuis, Laurent Bopp, Herve Claustrek,, R. G. Leticia Cotrim da Cunha, Xavier Giraud, Christine, K. E. K. Klaas , Louis Legendre, Manfredi Manizza,, R. B. R. Trevor Plat T, Shubha Sathyendranath, *et al.*, *Global Change Biology* **11**, 17 (2005)
57. Supplementary table s11, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST11.xls
58. A. Skovgaard, *Acta Protozoologica* **53**, 51 (2014)
59. Supplementary figure s6, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF6.pdf
60. S. Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, *The Journal of Eukaryotic Microbiology* **51**, 402 (2004)
61. F. Gomez, D. Moreira, K. Benzerara, P. Lopez-Garcia, *Environ Microbiol* **13**, 193 (2011)
62. C. E. Hamm, R. Merkel, O. Springer, P. Jurkojc, C. Maier *et al.*, *Nature* **421**, 841 (2003)
63. P. Assmy, V. Smetacek. (2009), pp. 27.
64. J. S. Weitz, T. Poisot, J. R. Meyer, C. O. Flores, S. Valverde *et al.*, *Trends Microbiol* **21**, 82 (2013)
65. Supplementary table 12, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalTables/ST12.xlsx
66. Supplementary figure s7, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF7.pdf
67. Supplementary figure s8, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF8.pdf
68. C. M. Mizuno, F. Rodriguez-Valera, N. E. Kimes, R. Ghai, *PLoS Genet* **9**, e1003987 (2013)
69. G. Lima-Mendez, A. Toussaint, R. Leplae, *Res Microbiol* **162**, 737 (2011)
70. D. T. Pride, T. M. Wassenaar, C. Ghose, M. J. Blaser, *BMC Genomics* **7**, 8 (2006)
71. Supplementary figure s9, http://psbweb07.psb.ugent.be/raeslab/data/tara/supplementalFigures/SF9.pdf
72. S. Pesant, F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot *et al.*, **submitted**,
73. E. Pruesse, C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig *et al.*, *Nucleic Acids Res* **35**, 7188 (2007)
74. R. C. Edgar, *Bioinformatics* **26**, 2460 (2010)
75. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, *Appl Environ Microbiol* **73**, 5261 (2007)
76. F. Mahé, T. Rognes, C. Quince, C. d. Vargas, M. Dunthorn, *PeerJ*, e593 (2014)
77. L. Guillou, D. Bachar, S. Audic, D. Bass, C. Berney *et al.*, *Nucleic Acids Research* **41**, D597 (2013)
78. Y. Benjamini, Y. Hochberg, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289 (1995)
79. D. Borcard, P. Legendre, C. Avois-Jacquet, H. Tuomisto, *Ecology* **85**, 1826 (2004)
80. F. G. Blanchet, P. Legendre, D. Borcard, *Ecology* **89**, 2623 (2008)
81. D. Borcard, P. Legendre, P. Drapeau, *Ecology* **73**, 1045 (1992)
82. M. B. Brown, *Biometrics* **31**, 987 (1975)
83. T. Tvedebrink, *Theoretical Population Biology* **78**, 200 (2010)
84. P. E. Meyer, F. Lafitte, G. Bontempi, *BMC Bioinformatics* **9**, 461 (2008)
85. W. Atmar, B. D. Patterson, *Oecologia* **96**, 373 (1993)
86. C. O. Flores, T. Poisot, S. Valverde, J. S. Weitz, *eprint arXiv:1406.6732*, (2014)
87. M. J. Barber, *Phys. Rev. E* **76**, 066102 (2007)
88. M. Almeida-Neto, P. Guimaraes, P. R. G. Jr, R. D. Loyola, W. Ulrich, *Oikos* **117**, 1227 (2008)
89. H. Ying, P. Gilna, W. Li, *Bioinformatics* **25**, 1338 (2009)
90. M. J. Sullivan, N. K. Petty, S. A. Beatson, *Bioinformatics* **27**, 1009 (2011)
91. L. Guillou, D. Bachar, S. Audic, D. Bass, C. Berney *et al.*, *Nucleic Acids Res* **41**, D597 (2013)
92. I. Ruiz-Trillo, M. Riutort, D. T. J. Littlewood, E. A. Herniou, J. Baguñà, *Science* **283**, 1919 (1999)
93. R. J. Gast, T. A. McDonnell, D. A. Caron, *Journal of Phycology* **36**, 172 (2000)
94. J. Schindelin, I. Arganda-Carreras, E. Frise, *Nature Methods* **9**, 676 (2012)
95. M. Newman, *Physical Review E* **74**, 036104 (2006)

Acknowledgements. We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research – Flanders (GLM, KF, SC, JR), Rega Institute (JR), KU Leuven (JR), The French Ministry of Research, the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001-02), ANR (projects POSEIDON/ANR-09-BLAN-0348, PHYTBACK/ANR-2010-1709-01, PROMETHEUS/ANR-09-PCS-GENM-217, TARA GIRUS/ANR-09-PCS-GENM-218, European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376, ERC Advanced Grant Awards to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790) to MBS, Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS to SGA, TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Reserca funded to SGA, JSPS KAKENHI Grant Number 26430184 to HO, FWO, BIO5, Biosphere 2, Agnès b., the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, Etienne Bourgois, the *Tara* schooner and its captain and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge the EMBL Advanced Light Microscopy Facility (ALMF), and in particular Rainer Pepperkok. All authors approved the final manuscript. This article is contribution number XXX of the *Tara* Oceans Expedition.

Table legends

Table 1. Properties of the merged taxon network. The positive sub-set of the network was clustered with the leading eigen vector algorithm (*95*).

Figure legends

Figure 1. Global oceanic taxon-environment interaction network properties. (A) Major environmental factors affecting abundance patterns. Phosphate concentration (PO4), Temperature and Nitrite concentration (NO2) are the top 3 parameters driving abiotic associations followed by Mixed Layer Depth (MLD), Particulate beam attenuation measured at 660 nm, Nitrite+Nitrate concentration (NO2NO3), Depth (here pressure), Silica concentration (Si), Nitracline, Eddy retention (in days) and others corresponds to the agglomerated contribution of the rest of parameters tested (see Table S1 for a complete listing). (B) Number of inter-domain and intra-domain copresences and mutual exclusions. (C) Distribution of edges across size fractions: 0.2-1.6(3), prokaryote-enriched fractions 0.2-1.6 μ m and 0.2-3 μ m; > 08, non-size-fractionated samples; 08_5, piconano-plankton; 20_180, micro-plankton; 180_2000, meso-plankton; interfrac, includes interfraction networks 08_5 versus 20_180, 08_5 versus 180_2000, 20_180 versus 180_2000 and 0.2-1.6(3) versus ≤0.2 (virus-enriched fraction).

Figure 2. Taxonomic and geographic patterns within the co-occurrence network. (A) Top 15 interacting taxon groups depicted as colored segments in a CIRCOS plot, where ribbons connecting two segments indicate copresence and exclusion links, on left and right, respectively. Size of the ribbon is proportional to the number of links (copresences and exclusions) between the OTUs assigned to the respective segments, and color is the one of the segment (of the two involved) with more total links. Links are dominated by the obligate parasites syndiniales and by Arthropoda and Dinophyceae. (B) *Tara* Oceans sampling stations grouped by oceanic provinces. (C) Frequency of local co-occurrence patterns across the oceanic provinces, showing that most local patterns are located in MS. (D-G) Taxonomic patterns of co-occurrences across MS (D), SPO (E), IO (F) and RS (G). Edges are represented as ribbons between barcodes grouped into their taxonomic order as in A. Links sharing the same segment are affiliated to the same taxon (Order), showing that the connectivity patterns across taxa are conserved at high taxonomic ranks. The local specificity of interactions at higher resolution (OTUs) is apparent by thin ribbons (edge resolution) with different starts and end positions (different OTUs) within the shared (taxon) segment, section color and ordering correspond to those in panel A.

Figure 3. Top-down interactions in plankton. (A) Subnetwork of metanodes that encapsulate barcodes affiliated to parasites or Plankton Functional Ttypes (PFTs). The PFTs mapped onto the network are: phytoplankton dimethyl sulfide (DMS) producers, mixed phytoplankton, phytoplankton silicifiers, pico-eukaryotic heterotrophs, proto-zooplankton and meso-zooplankton. Edge width reflects the number of edges in the taxon graph between the corresponding metanodes. Over-

represented links (multiple-test corrected Pval < 0.05, Fisher's exact test) are colored in green if they represent copresences and in red if they represent exclusions; grey means non-overrepresented combinations. When both copresences and exclusions were significant, the edge is shown as copresence. (B) Parasite connections within micro- and zooplankton groups. (C) Number of hosts per phage (inset: phage associations to bacterial (target) phyla). **(D)** Putative Bacteroidetes viruses detected by co-occurence and detection in a single-cell genome (SAG). On the left, viral sequences from a Flavobacterium SAG (top) and *Tara* Oceans virome (bottom), displaying an average of 89% nucleotide identity. On the right is the correspondence between the ribosomal genes detected in the same SAG (top) and the 16S sequence associated to the *Tara* Oceans contig based on co-occurence (79% nucleotide identity). For clarity, a subset of contig ARTD0100013 only (from 10,000 to 16,000 nucleotides) is displayed. This sequence was also reverse-complemented. PurM: Phosphoribosylaminoimidazole synthetase, DNA Pol. A: DNA polymerase A.

Figure 4. Experimental validation of network-predicted interaction (photosymbiosis).

Guided by the predictions from the co-occurrence network and abundance patterns, acoel flatworms (*Symsagittifera* sp.) together with their photosynthetic green microalgal endosymbionts (*Tetraselmis* sp.) were collected in microplankton samples from *Tara* Oceans Station 22 in the Mediterranean Sea. Pictures show a 3D reconstructed specimen from LSCM images (Green channel: cellular membranes (DiOC6); Blue channel: DNA and the nuclei (Hoechst33342); Red channel: chlorophyll autofluorescence). **(A)** Co-occurrence plot of *Symsagittifera* and *Tetraselmis* related OTUs along *Tara* Oceans stations, showing the relatively high abundance of the holobiont at Station 22. (B) Dorsal view of the entire acoel flatworm specimen (~300µm). The epidermis (green) is completely covered with cilia and displays some pore holes. **(C)** The removal of the green channel reveals the widespread distribution of small unicellular algae (red areas) inside the acoel body. The worm's nuclei display a clear signal (compact round blue shapes) while the algal nuclei are dimmer. A dinoflagellate theca (arrow head) is located in the central syncytium likely indicating predation. **(D)** Cross-section along a ZY plane allows localization of the algae, beneath the epidermis in the parenchyma. Only the external cell layer (green signal) from the dorsal view is visible, due to the thickness and opacity of the worm. Scale bar: 50 µm.

Table 1. Network properties

Nodes	Edges	Positive edges (per-centage)	Negative edges	Average clustering coefficient	Average path length	Diameter	Average betweenness	Modularity of positive network	Number of modules in positive network
9169	92,633	68,856 (74.33)	23,777	0.229	3.43	12	11024	0.51	51

Tara Oceans Coordinators

Silvia G. Acinas[1], Peer Bork[2], Emmanuel Boss[3], Chris Bowler[4], Colomban De Vargas[5,6], Michael Follows[7], Gabriel Gorsky[8,9], Nigel Grimsley[10,11], Pascal Hingamp[12], Daniele Iudicone[13], Olivier Jaillon[14,15,16], Stefanie Kandels-Lewis[2], Lee Karp-Boss[3], Eric Karsenti[17,18], Uros Krzic[19], Fabrice Not[5,6], Hiroyuki Ogata[20], Stephane Pesant[21,22], Jeroen Raes[23,24,25], Emmanuel G. Reynaud[26], Christian Sardet[8], Mike Sieracki[27], Sabrina Speich[28,29], Lars Stemmann[8], Matthew B. Sullivan[30], Shinichi Sunagawa[2], Didier Velayoudon[31], Jean Weissenbach[14,15,16], Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)-CSIC, Barcelona, Spain.
[2]Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany.
[3]School of Marine Sciences, University of Maine, Orono, USA.
[4]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France.
[5]CNRS, UMR 7144, Station Biologique de Roscoff, Roscoff, France.
[6]Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Roscoff, France.
[7]Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA.
[8] CNRS/UPMC, UMR 70O9, BioDev, Observatoire océanologique, Villefranche/mer, France.
[9]Sorbonne Universités, UPMC Univ Paris 06, UMR 7093, LOV, Observatoire océanologique, Villefranche/mer, France.
[10]CNRS UMR 7232, BIOM, Banyuls-sur-Mer, France.
[11]Sorbonne Universités, OOB, UPMC Paris 06, Banyuls-sur-Mer, France.
[12]Aix Marseille Université, CNRS, IGS UMR 7256, Marseille, France.
[13]Laboratory of Ecology and Evolution of Plankton, Stazione Zoologica Anton Dohrn, Naples, Italy.
[14]CEA, Genoscope, Evry France.
[15]CNRS, UMR 8030, Evry, France.
[16]Université d'Evry, UMR 8030, Evry, France.
[17]Environmental and Evolutionary Genomics Section, Institut de Biologie de l'Ecole Normale Supérieure, CNRS, UMR 8197, Institut National de la Santé et de la Recherche Médicale U1024, Ecole Normale Supérieure, Paris, France.
[18]Directors' Research, European Molecular Biology Laboratory, Heidelberg, Germany.
[19]Cell Biology and Biophysics, European Molecular Biology Laboratory, Heidelberg, Germany.
[20]Institute for Chemical Research, Kyoto University, Kyoto, Japan.
[21]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany.

[22]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany.
[23]Department of Microbiology and Immunology, Rega Institute KU Leuven, Leuven, Belgium.
[24]VIB Center for the Biology of Disease, VIB, Leuven, Belgium.
[25]Laboratory of Microbiology, Vrije Universiteit Brussel, Brussels, Belgium.
[26]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland.
[27]Bigelow Laboratory for Ocean Sciences, East Boothbay, USA.
[28]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, Paris, France.
[29]Laboratoire de Physique des Océan, UBO-IUEM, Polouzané, France.
[30]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, USA.
[31]DVIP Consulting, Sèvres, France.