Plankton networks driving carbon export in the oligotrophic ocean

Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi, Stéphane Audic, L. Berline, Jennifer R. Brum, et al.

Title:
Plankton networks driving carbon export in the oligotrophic ocean

Authors: Lionel Guidi[1,2,*], Samuel Chaffron[3,4,5,*], Lucie Bittner[6,7,8,*], Damien Eveillard[9,*], Abdelhalim Larhlimi[9], Simon Roux[10,11], Youssef Darzi[3,4], Stephane Audic[8], Léo Berline[1,12], Jennifer Brum[10,11], Luis Pedro Coelho[13], Julio Cesar Ignacio Espinoza[10], Shruti Malviya[7], Shinichi Sunagawa[13], Céline Dimier[8], Stefanie Kandels-Lewis[13,14], Marc Picheral[1], Julie Poulain[15], Sarah Searson[1,2], *Tara* Oceans coordinators, Lars Stemmann[1], Fabrice Not[8], Pascal Hingamp[16], Sabrina Speich[17], Mick Follows[18], Lee Karp-Boss[19], Emmanuel Boss[19], Hiroyuki Ogata[20], Stephane Pesant[21,22], Jean Weissenbach[15,23,24]**,** Patrick Wincker[15,23,24], Silvia G. Acinas[25], Peer Bork[13,26], Colomban de Vargas[8], Daniele Iudicone[27], Matthew B. Sullivan[10,11], Jeroen Raes[3,4,5], Eric Karsenti[7,14], Chris Bowler[7], Gabriel Gorsky[1]

[*] These authors contributed equally to this work

Affiliations:

[1.] Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d'oceanographie de Villefranche (LOV), Observatoire Océanologique, Villefranche-sur-Mer, France

[2.] Department of Oceanography, University of Hawaii, Honolulu, Hawaii, USA

[3.] Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium.

[4.] Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium.

[5.] Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium.

[6.] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institut de Biologie Paris-Seine (IBPS), Evolution Paris Seine, F-75005, Paris, France.

[7.] Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France.

[8.] Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, Roscoff, France

[9.] LINA UMR 6241, Université de Nantes, EMN, CNRS, 44322 Nantes, France.

[10.] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ, 85721, USA.

[11.] Current affiliation: Department of Microbiology, The Ohio State University, Columbus OH 43210, USA

[12] Current affiliation: Aix-Marseille Univ., Mediterranean Institute of Oceanography (MIO), 13288, Marseille, Cedex 09, France ; Université du Sud Toulon-Var, MIO, 83957, La Garde cedex, France ; CNRS/INSU, MIO UMR 7294; IRD, MIO UMR235.

[13.] Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, 69117 Heidelberg, Germany,

[14.] Directors' Research European Molecular Biology Laboratory Meyerhofstr. 1 69117 Heidelberg Germany,

[15.] CEA - Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry France.

[16.] Aix Marseille Université CNRS IGS UMR 7256 13288 Marseille France

[17.] Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond 75231 Paris Cedex 05 France.

[18.] Dept of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA.

[19.] School of Marine Sciences, University of Maine, Orono, USA.

[20.] Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, 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.] CNRS, UMR 8030, CP5706, Evry France.

[24.] Université d'Evry, UMR 8030, CP5706, Evry France.

[25.] Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC Pg. Marítim de la Barceloneta 37-49 Barcelona E08003 Spain.

[26.] Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany,

[27.] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121, Naples, Italy.

The biological carbon pump is the process by which CO_2 is transformed to organic carbon *via* photosynthesis, exported through sinking particles, and finally sequestered in the deep ocean or sediment. While the intensity of the pump correlates with plankton community composition, the underlying ecosystem structure and interactions driving the process remain largely uncharacterised. Here we use environmental and metagenomic data gathered during the *Tara* Oceans expedition to improve our understanding of carbon export in the oligotrophic ocean. We show that specific euphotic plankton communities correlate with carbon export and highlight unexpected and overlooked taxa such as Radiolaria, alveolate parasites, as well as *Synechococcus* and their phages, as lineages most strongly associated with carbon export in the subtropical, nutrient-depleted, oligotrophic ocean. Additionally, we show that the relative abundance of just a few bacterial and viral genes can predict most of the variability in carbon export in these regions.

Marine planktonic photosynthetic organisms are responsible for approximately fifty percent of Earth's primary production and they fuel the global ocean biological carbon pump[1]. The intensity of the pump is correlated to plankton community composition[2,3], and controlled by the relative rates of primary production and carbon remineralisation[4]. About 10% of this newly produced organic carbon in the surface ocean is exported through gravitational sinking of particles. Finally, after multiple transformations, only a fraction of the exported material will reach the deep ocean where it is sequestered over thousand-year timescales of the ocean's overturning circulation[5].

Like most biological systems, marine ecosystems in the sunlit upper layer of the ocean (denoted the euphotic zone) are complex[6,7], characterised by a wide range of biotic and abiotic interactions[8-10] and in constant balance between carbon production, transfer to higher trophic levels, remineralisation, and export to the deep layers[11]. The marine ecosystem structure and its taxonomic and functional composition likely evolved to comply with this loss of energy by modifying organism turnover times and by the establishment of complex feedbacks between them[6] and the substrates they can exploit for metabolism[12]. Decades of groundbreaking research have focused on identifying independently the key players involved in the biological carbon pump. Among autotrophs, diatoms are commonly attributed to being important in carbon flux because of their large size and fast sinking rates[13-15] while small autotrophic picoplankton may contribute directly as a result of subduction of surface water resulting from sub-mesoscale dynamic features[16] or indirectly by aggregating with larger settling particles or through their consumption by organisms at higher trophic levels[17]. Among heterotrophs, zooplankton such as copepods impact carbon flux *via* production of fast-sinking fecal pellets while migrating hundreds of meters in the water-column[18,19]. These observations, focusing on just a few components of the marine ecosystem, highlight that carbon export results from multiple biotic interactions and that a better understanding of the mechanisms involved in its regulation will likely require an analysis of the entire planktonic ecosystem.

Advanced sequencing technologies now offer the opportunity to simultaneously survey whole planktonic communities and associated molecular functions in unprecedented detail. Such a holistic approach may allow the identification of community- or gene-based biomarkers that could be used to monitor and predict ecosystem functions, e.g., related to the biogeochemistry of the ocean[20-22]. Here, we leverage global-scale ocean genomics datasets[10,23-25] and associated environmental data to assess the coupling between ecosystem structure, functional repertoire, and the carbon export component of the biological carbon pump.

Carbon export and plankton community composition

The *Tara* Oceans global circumnavigation crossed diverse ocean ecosystems and sampled plankton at an unprecedented scale[20,26] (see Methods). Hydrographic data were measured *in situ* or in seawater samples at all stations, as well as nutrients, oxygen and photosynthetic pigments (see Methods). Net Primary Production (NPP) was derived from satellite measurements (see Methods). In addition, particle size distributions (100 μm to a few mm) and concentrations were measured using an Underwater Vision Profiler (UVP) from which carbon export, corresponding to the carbon flux (Fig. 1) at 150 m, was calculated to range from 0.014 to 18.3 $mg.m^{-2}.d^{-1}$ using previously validated methods (see Methods). The approach allowed us to assemble the largest homogeneous carbon flux dataset during a single expedition, corresponding to more than 600 profiles over 150 stations. This dataset is of similar magnitude to the body of historical data available in the literature that includes the 134 deep sediment trap-based carbon flux time-series[27] from the JGOFS program and the 419 thorium-derived particulate organic carbon (POC) export measurements[28].

From 68 globally distributed sites, a total of 7.2 Tb of metagenomics data, representing *circa* 40 million non-redundant genes, around 35,000 Operational Taxonomic Units (OTUs) of prokaryotes (Bacteria and Archaea) and numerous mainly uncharacterized viruses and picoeukaryotes, have been described recently[23,25]. In addition, a set of 2.3 million eukaryotic 18S rDNA ribotypes was generated from a subset of 47 sampling sites corresponding to approximately 130,000 OTUs[24]. Finally, 5,476 viral “populations” were identified at 43 sites from viral metagenomic contigs, only 39 (<0.1%) of which had been previously observed[25] (see Methods). These genomics data combined across all domains of life together with carbon flux estimates and other environmental parameters were used to explore the relationships between marine biogeochemistry and euphotic plankton communities (see

Methods) in the oligotrophic open ocean. Our study did not include high latitude areas due to the current lack of available molecular data.

Using a method for regression-based modeling of high dimensional data in biology (specifically a sparse Partial Least Square analysis - sPLS[29], Extended data Fig. 1), we detected several plankton lineages for which relative sequence abundance correlated with carbon export and other environmental parameters, most notably with NPP, as expected (Fig. 2 and see Supplementary Information SI1). These included diatoms, dinoflagellates and metazoa (zooplankton), lineages classically identified as key contributors to carbon export.

Plankton community networks associated with carbon export

While the analysis presented in Fig. 2 supports previous findings about key organisms involved in carbon export from the euphotic zone[14,15,17-19], it is not able to capture how the intrinsic structure of the planktonic community relates to this biogeochemical process. Conversely, although other recent holistic approaches[10,30,31] used species co-occurrence networks to reveal potential biotic interactions, they do not provide a robust description of sub-communities driven by abiotic interactions. To overcome these issues, we applied a systems biology approach known as Weighted Gene Correlation Network Analysis (WGCNA[32,33]) to detect significant associations between the *Tara* Oceans genomics data and carbon export. This method delineates communities in the euphotic zone that are the most associated with carbon export rather than predicting organisms associated with sinking particles.

In brief, the WGCNA approach builds a network in which nodes are features (in this case plankton lineages or gene functions) and links are evaluated by the robustness of co-occurrence scores. WGCNA then clusters the network into modules (hereafter denoted subnetworks) that can be examined to find strong and significant subnetwork-trait relationships. We then filtered each subnetwork using a Partial Least Square (PLS) analysis that emphasizes key nodes (based on the Variable Importance in Projection (VIP) scores; see Methods and Extended data Fig. 1). These particular nodes are mandatory to summarize a subnetwork (or community) related to carbon export. In particular, they are of interest for evaluating (i) subnetwork robustness and (ii) predictive power for a given trait (see Methods and Extended data Fig. 1).

We applied WGCNA to the relative abundance tables of eukaryotic, prokaryotic and viral lineages[23-25] and identified unique subnetworks significantly associated with carbon export within each dataset (see Methods and Supplementary Information SI1, SI2, SI3). The eukaryotic subnetwork (subnetwork-trait relationship to carbon export, Pearson cor. = 0.81, $p = 5e^{-15}$) contained 49 lineages (Extended data Fig. 2a and Supplementary Information SI2) among which twenty percent represented photosynthetic organisms (Fig. 3a and Supplementary Information SI2). Surprisingly, this small subnetwork's structure correlates very strongly to carbon export (Pearson cor. = 0.87, $p = 5e^{-16}$, Extended data Fig. 2d) and it predicts as much as 69% (Leave-One-Out Cross-Validated (LOOCV), $R^2 = 0.69$) of the variability in carbon export (Extended data Fig. 3a). Only ~6% of the subnetwork nodes correspond to diatoms and they show lower VIP scores than dinoflagellates (Supplementary Information SI2). This is likely because our samples are not from silicate replete conditions where diatoms were blooming (see Methods). Furthermore, our analysis did not incorporate data from high latitudes, where diatoms are known to be particularly important for carbon export, so this result suggests that dinoflagellates have a heretofore unrecognized role in carbon export processes in subtropical oligotrophic 'type' ecosystems, one of the largest biome on Earth. More precisely four of the five highest VIP scoring eukaryotic lineages that correlated with carbon flux were heterotrophs such as Metazoa (copepods), non-photosynthetic Dinophyceae, and Rhizaria (Fig. 3a and Supplementary Information SI2).

These results corroborate recent metagenomics analysis of microbial communities from sediment traps in the oligotrophic North Pacific subtropical gyre[34]. Consistently, *in situ* imaging surveys have revealed Rhizarian lineages, made up of large fragile organisms such as the Collodaria, to represent an until now under-appreciated component of global plankton biomass[35], which here also appear to be of relevance for carbon export. Another 14% of lineages from the subnetwork correspond to parasitic organisms, a largely under-explored component of planktonic ecosystems.

The prokaryotic subnetwork that associated most significantly with carbon export (subnetwork-trait relationship to carbon export, Pearson cor. = 0.32, p = 9e^{-03}) contained 109 OTUs (Extended data Fig. 2b and Supplementary Information SI3), its structure correlated well to carbon export (Pearson cor. = 0.47, p = 5e^{-06}, Extended data Fig. 2e) and it could predict as much as 60% of the carbon export (LOOCV, R^2 = 0.60) (Extended data Fig. 3b). By far the highest VIP score within this community was assigned to *Synechococcus*, followed by *Cobetia*, *Pseudoalteromonas* and *Idiomarina,* as well as *Vibrio* and *Arcobacter* (Fig. 3b and Supplementary Information SI3). Noteworthy, *Prochlorococcus* genera and SAR11 clade fall out of this community, while the significance of *Synechococcus* for carbon export could be validated using absolute cell counts estimated by flow cytometry (Pearson cor. = 0.64, p = 4e^{-10}, Extended data Fig. 4b). Moreover, *Prochlorococcus* cell counts did not correlate with carbon export (Pearson cor. = -0.13, p = 0.27, Extended data Fig. 4a) whereas the *Synechococcus* to *Prochlorococcus* cell count ratio correlated positively and significantly (Pearson cor. = 0.54, p = 4e^{-07}, Extended data Fig. 4c), suggesting the relevance of *Synechococcus*, rather than *Prochlorococcus*, to carbon export. Interestingly, *Pseudoalteromonas*, *Idiomarina*, *Vibrio* and *Arcobacter* (of which several species are known to be associated with eukaryotes[36]) have also been observed in live and poisoned sediment traps[34] and these genera display very high VIP scores in our subnetwork associated with carbon export. Additional genera reported as being enriched in poisoned traps (also known as being associated with eukaryotes) include *Enterovibrio* and *Campylobacter*, and are present as well in our carbon export subnetwork.

Interestingly, the viral subnetwork (*n*=277) most related to carbon export (Pearson cor. = 0.93, $p = 2e^{-15}$, Extended data Fig. 2c) contained particularly high VIP scores for two *Synechococcus* phages (Fig. 3c and Supplementary Information SI4), which represented a 16-fold enrichment (Fisher's exact test $p = 6.4e^{-09}$). Its structure also correlated with carbon export (Pearson cor. = 0.88, $p = 6e^{-93}$, Extended data Fig. 2f) and it could predict up to 89% of the variability of carbon export (LOOCV, $R^2 = 0.89$) (Extended data Fig. 3c). The significance of these convergent results is reinforced by the fact that sequences from these datasets are derived from organisms collected on independent size filters (see Methods), and further implicates the importance of top-down processes in carbon export.

With the aim of integrating eukaryotic, prokaryotic, and viral carbon export communities, we synthesized their respective subnetworks using, as a backbone, a single global co-occurrence network established previously[10]. The resulting network focused on key lineages and their predicted co-occurrences (Fig. 4). Lineages with high VIP values (such as *Synechococcus*) are revealed here as hubs of the co-occurrence network[10], illustrating the potentially strategic key roles within the integrated network of lineages under-appreciated by conventional methods to study carbon export in the ocean. Associations between the hub lineages are mostly mutually exclusive which may explain the relatively weak correlation of some of these lineages with carbon export when using standard correlation analyses as shown in Fig. 2.

Gene functions associated with carbon export

Given the potential importance of prokaryotic processes influencing the biological carbon pump[22], we used the same analytical approaches to examine the prokaryotic genomic functions associated with carbon export in the annotated Ocean Microbial Reference Gene Catalogue from *Tara* Oceans[23]. We built a global co-occurrence network for functions (i.e., Orthologous Groups of genes or OGs) from the euphotic zone and identified two subnetworks of functions that are significantly associated with carbon export (Fig. 5a, Extended data Fig. 5a, light and dark green subnetworks; FNET1 and FNET2, respectively, and Extended data Fig. 5c).

The majority of functions in FNET1 and FNET2 correlate well with carbon export (FNET1: mean Pearson cor. = 0.45, s.d. 0.09 and FNET2: mean Pearson cor. = 0.34, s.d. 0.10). Interestingly, FNET2 functions (*n*=220) encode mostly (83%) core functions (i.e., functions observed in all euphotic samples, see Methods) while the majority of FNET1 functions (*n*=441) are non-core (85%) (see Supplementary Information SI5, SI6), highlighting both essential and adaptive ecological functions associated with carbon export. Top VIP scoring functions in the FNET1 subnetwork are membrane proteins such as ABC-type sugar transporters (Fig. 5a). This subnetwork also contains many functions specific to the *Synechococcus* accessory photosynthetic apparatus (e.g., relating to phycobilisomes, phycocyanin and phycoerythrin; see Supplementary Information SI5), which is consistent with the major role of this genus for carbon export inferred from the prokaryotic subnetwork (Fig. 3b). In addition, functions related to carbohydrates, inorganic ion transport and metabolism, as well as transcription, are also well represented (Fig. 5b), suggesting overall a subnetwork of functions dedicated to photosynthesis and growth.

The FNET2 subnetwork contains several functions encoded by genes taxonomically assigned to *Candidatus pelagibacter* and *Prochlorococcus*, known as occupying similar oceanic regions as *Synechococcus*, but overall most of its relative abundance (74%) is taxonomically unclassified (Extended data Fig. 6). Top VIP scoring functions in FNET2 are also membrane proteins and ABC-type sugar transporters, as well as functions involved in carbohydrate breakdown such as a chitinase (Fig. 5a). These features highlight the potential roles of bacteria in the formation and degradation of marine aggregates[37]. Strikingly, 77% and 58%, of OGs with a VIP score > 1 in FNET1 and FNET2, respectively, are functionally uncharacterized[38,39] (Fig. 5b), pointing to the strong need for future molecular work to explore these functions (see Supplementary Information SI5, SI6).

The relevance of the identified bacterial functions to predict carbon export was also confirmed by PLS regression (Extended data Fig. 6b and 6c). As proposed for plankton communities, the functional subnetworks predict 41% and 48% of carbon export variability (LOOCV, $R^2 = 0.41$ and 0.48 for FNET1 and FNET2, respectively) with a minimal number of functions (Fig. 5b, 123 and 54 functions with a VIP score > 1 for FNET1 and FNET2, respectively). Finally, higher predictive power was obtained using subnetworks of viral protein clusters (Extended data Fig. 5b, 5d and 7a), predicting 55% and 89% of carbon export variability (LOOCV $R^2 = 0.55$ and 0.89 for VNET1 and VNET2, respectively; Extended data Fig. 7b, Supplementary Information, SI7, SI8), suggesting again the key role, of not only bacteria, but also their phages in biological processes sustaining carbon export at a global level.

Discussion

In this report we have revealed the potential contribution of under-appreciated components of plankton communities, as well as confirmed the importance of prokaryotes and viruses, in the carbon export component of the biological carbon pump in the nutrient-depleted oligotrophic ocean. Carbon export was estimated from particle size distribution at 150 m measured with the UVP, and we assumed similar particle composition across all size classes.

Furthermore, because of instrument and method limitations, particles smaller than 250 μm were not used for these estimations (see Methods). These export estimates evaluate how much carbon leaves the euphotic zone, but they are not necessarily related to sequestration, which occurs deeper in the water column and over longer timescales. Overall, the use of the UVP was the only realistic method to evaluate carbon flux over the 3 years expedition because deployment of sediment traps at all stations would have been impossible. While our findings are consistent with the numerous previous studies that have highlighted the central role of copepods and diatoms in the biological carbon pump[14,15,17-19], they place them in an ecosystem context and generate hypotheses as to the processes that determine the intensity of export, such as parasitism and predation. For example, while viruses are commonly assumed to lyse cells and maintain fixed organic carbon in surface waters, thereby reducing the intensity of the biological carbon pump[40], there are hints that viral lysis may increase carbon export through the production of colloidal particles and aggregate formation[41]. Our current study suggests that these latter roles may be more ubiquitous than currently appreciated. The importance of aggregation and cell stickiness as inferred from gene network analysis, should be further explored mechanistically to investigate the biological significance of these findings.

The future evolution of the oceanic carbon sink remains uncertain because of poorly constrained processes, particularly those associated with the biological pump. With current trends in climate change, the size and biodiversity of phytoplankton are predicted to decrease globally[42,43]. Furthermore, in spite of the potential importance of viruses revealed in this study, they have largely been ignored because of limitations in sampling technologies. Consequently, as oligotrophic gyres expand and global mean NPP decreases[44], the field is currently unable to predict the consequences for carbon export from the ocean's euphotic zone. By pinpointing key species that appear to be strongly associated with carbon export in these areas, as well as their co-occurences within plankton communities and key microbial functions, the integrated datasets combined with advanced computational techniques used in this study could provide a framework to address this critical bottleneck.

One of the grand challenges in the life sciences is to link genes to ecosystems[45], based on the posit that genes can have predictable ecological footprints at community and ecosystem levels[46-48]. The extensive data sets from *Tara* Oceans have allowed us to predict as much as 89% of the variability in carbon export from the oligotrophic surface ocean with just a small number of genes, largely with unknown functions, encoded by prokaryotes and viruses. These findings can be used as a basis to include biological complexity and guide experimental work designed to inform modeling of the global carbon cycle and to understand how it influences and is influenced by changes in climate. Such statistical analyses scaling from gene-to-ecosystems may open the way to the development of a new conceptual and methodological framework to better understand the mechanisms underpinning key ecological processes.

References and Notes

1 Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237-240, doi:10.1126/Science.281.5374.237 (1998).

2 Boyd, P. W. & Newton, P. Evidence of the potential Influence of planktonic community structure on the interannual variability of particulate organic-carbon flux. *Deep-Sea Res. I.* **42**, 619-639 (1995).

3 Guidi, L. *et al.* Effects of phytoplankton community on production, size, and export of large aggregates: A world-ocean analysis. *Limnol. Oceanogr.* **54**, 1951-1963 (2009).

4 Kwon, E. Y., Primeau, F. & Sarmiento, J. L. The impact of remineralization depth on the air-sea carbon balance. *Nat Geosci* **2**, 630-635 (2009).

5 IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2013).

6 Kitano, H. Biological robustness. *Nat Rev Genet* **5**, 826-837, doi:10.1038/Nrg1471 (2004).

7 Suweis, S., Simini, F., Banavar, J. R. & Maritan, A. Emergence of structural and dynamical properties of ecological mutualistic networks. *Nature* **500**, 449-452, doi:10.1038/Nature12438 (2013).

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

9 Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459**, 193-199, doi:10.1038/Nature08058 (2009).

10 Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome. *Science* **348**, doi:10.1126/science.1262073 (2015).

11 Giering, S. L. C. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**, 480-483 (2014).

12 Azam, F. Microbial control of oceanic carbon flux: The plot thickens. *Science* **280**, 694-696 (1998).

13 Agusti, S. *et al.* Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the biological pump. *Nat Commun* **6**, doi:10.1038/Ncomms8608 (2015).

14 Sancetta, C., Villareal, T. & Falkowski, P. Massive Fluxes of Rhizosolenid Diatoms - a Common Occurrence. *Limnol. Oceanogr.* **36**, 1452-1457 (1991).

15 Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic north Pacific gyre at station ALOHA. *Mar. Ecol. Prog. Ser.* **182**, 55-67, doi:10.3354/meps182055 (1999).

16 Omand, M. M. *et al.* Eddy-driven subduction exports particulate organic carbon from the spring bloom. *Science* **348**, 222-225, doi:10.1126/science.1260062 (2015).

17 Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean. *Science* **315**, 838-840 (2007).

18 Steinberg, D. K. *et al.* Bacterial vs. zooplankton control of sinking particle flux in the ocean's twilight zone. *Limnol. Oceanogr.* **53**, 1327-1338 (2008).

19 Turner, J. T. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Prog. Oceanogr.* **130**, 205-248, doi:10.1016/j.pocean.2014.08.005 (2015).

20 Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *Plos Biol.* **9**, doi:10.1371/journal.pbio.1001177 (2011).

21 Strom, S. L. Microbial ecology of ocean biogeochemistry: A community perspective. *Science* **320**, 1043-1045, doi:10.1126/Science.1153527 (2008).

22 Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**, 1257594, doi:10.1126/Science.1257594 (2015).

23 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, doi:10.1126/science.1261359 (2015).

24 de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, doi:10.1126/science.1261605 (2015).

25 Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**, doi:10.1126/science.1261498 (2015).

26 Bork, P. *et al.* Tara Oceans studies plankton at PLANETARY SCALE. *Science* **348**, 873-873, doi:10.1126/science.aac5605 (2015).

27 Honjo, S., Manganini, S. J., Krishfield, R. A. & Francois, R. Particulate organic carbon fluxes to the ocean interior and factors controlling the biological pump: A synthesis of global sediment trap programs since 1983. *Prog. Oceanogr.* **76**, 217-285, doi:10.1016/j.pocean.2007.11.003 (2008).

28 Henson, S. A., Sanders, R. & Madsen, E. Global patterns in efficiency of particulate organic carbon export and transfer to the deep ocean. *Global. Biogeochem. Cy.* **26**, doi:10.1029/2011GB004099 (2012).

29 Lê Cao, K. A., Rossouw, D., Robert-Granié, C. & Besse, P. A Sparse PLS for Variable Selection when Integrating Omics Data. *Stat Appl Genet Mol* **7**, doi:10.2202/1544-6115.1390 (2008).

30 Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes from environmental and whole-genome sequence data. *Genome Res.* **20**, 947-959, doi:10.1101/Gr.104521.109 (2010).

31 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538-550, doi:10.1038/Nrmicro2832 (2012).

32 Aylward, F. O. *et al.* Microbial community transcriptional networks are conserved in three domains at ocean basin scales. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1502883112 (2015).

33 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics* **9** (2008).

34 Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front Microbiol* **6**, Artn 469, doi:10.3389/Fmicb.2015.00/169 (2015).

35 Biard, T. *et al. In situ* imaging reveals the biomass of large protists in the global ocean. *Nature* (submitted).

36 Thomas, T. *et al.* Analysis of the Pseudoalteromonas tunicata Genome Reveals Properties of a Surface-Associated Life Style in the Marine Environment. *PLoS ONE* **3**, doi:10.1371/journal.pone.0003252 (2008).

37 Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* **5**, 782-791, doi:10.1038/nrmicro1747 (2007).

38 Shi, Y. M., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266-U154, doi:10.1038/nature08055 (2009).

39 Yooseph, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *Plos Biol.* **5**, 432-466, doi:10.1371/journal.pbio.0050016 (2007).

40 Suttle, C. A. Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801-812, doi:10.1038/Nrmicro1750 (2007).

41 Weinbauer, M. G. Ecology of prokaryotic viruses. *Fems Microbiol Rev* **28**, 127-181, doi:10.1016/j.femsre.2003.08.001 (2004).

42 Finkel, Z. V. *et al.* Phytoplankton in a changing world: cell size and elemental stoichiometry. *J. Plankton Res.* **32**, 119-137 (2010).

43 Sommer, U. & Lewandowska, A. Climate change and the phytoplankton spring bloom: warming and overwintering zooplankton have similar effects on phytoplankton. *Glob. Change Biol.* **17**, 154-162, doi:10.1111/J.1365-2486.2010.02182.X (2011).

44 Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444**, 752-755 (2006).

45 DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496-503, doi:10.1126/Science.1120250 (2006).

46 Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics. *P. Natl. Acad. Sci. USA* **106**, 1374-1379, doi:10.1073/Pnas.0808022106 (2009).

47 Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300-1302, doi:10.1126/Science.277.5330.1300 (1997).

48 Wymore, A. S. *et al.* Genes to ecosystems: exploring the frontiers of ecology with one of the smallest biological units. *New Phytol* **191**, 19-36, doi:10.1111/J.1469-8137.2011.03730.X (2011).

Figure Legends:

Figure 1 | Global view of carbon fluxes along the *Tara* Oceans circumnavigation route. Carbon flux in mg.m^{-2}.d^{-1} estimated from particles size distribution and abundance measured with the Underwater Vision Profiler 5 (UVP5).

Figure 2 | Eukaryotic community associated to carbon export seen using standard methods for regression-based modeling of high dimensional data. Eukaryotic lineages associated to carbon export as revealed by sPLS analysis. Correlations between lineages and environmental parameters are depicted as a clustered heatmap and lineages with a correlation to carbon export higher than 0.2 are highlighted.

Figure 3 | Ecological networks reveal key taxa lineages associated with carbon export at global scale. The relative abundances of taxa in selected subnetworks were used to estimate carbon export and to identify key lineages associated with the process. **a,** The selected eukaryotic subnetwork (*n*=49, see Supplementary Information SI2) can predict carbon export with high accuracy (PLS regression, LOOCV, R^2=0.69, see Extended data Fig. 3a). Lineages with the highest VIP score (dots size is proportional to the VIP score in the scatter plot) in the PLS are depicted as red dots corresponding to three Rhizaria (Collodaria, *Collozoum inerme* and *Sticholonche* sp.), one copepod (*Oithona* sp.), one siphonophore (*Lilyopsis*), three Dinophyceae and one ciliate (*Spirotontonia turbinata*). **b,** The selected prokaryotic subnetwork (*n*=109, see Supplementary Information SI3) can predict carbon export with good accuracy (PLS regression, LOOCV, R^2=0.60, see Extended data Fig. 3b). **c,** The selected viral population subnetwork (*n*=277, see Supplementary Information SI4) can predict carbon export with high accuracy (PLS regression, LOOCV, R^2=0.89, see Extended data Fig. 3c). Two viral populations with a high VIP score (red dots) are predicted as *Synechococcus* phages (see Supplementary Information SI4).

Figure 4 | Plankton community network built from eukaryotic, prokaryotic and viral subnetworks related to carbon export. Major lineages were selected within the three subnetworks (VIP > 1). Co-occurrences between all lineages of interest were extracted from a previously established global co-occurrence network (see methods). Only lineages discussed within the study are pinpointed. The resulting graph is composed of 329 nodes, 467 edges, with a diameter of 7, and average weighted degree of 4.6.

Figure 5 | Bacterial functional networks reveal key functions associated with carbon export at global scale. A bacterial functional network was built based on Orthologous Group/Gene (OG) relative abundances using the WGCNA methodology (see Methods) and correlated to classical oceanographic parameters. **a,** Two functional subnetworks (light and dark green, FNET1 (*n*=220) and FNET2 (*n*=441), respectively) are significantly associated with carbon export (FNET1: Pearson cor. 0.42, p = 4e^{-09} and FNET2: 0.54, p = 7e^{-06}, see Extended data Fig. 5a). The highest VIP score functions from top to bottom correspond to red dots from right to left. **b,** Higher functional categories are depicted for functions with a VIP score >1 (PLS regression, LOOCV, FNET1 R^2=0.41 and FNET2 R^2=0.48, see Extended data Fig. 6) in both functional subnetworks,

Methods

Environmental data collection

From 2009-2013, environmental data (Supplementary Information SI9) were collected across all major oceanic provinces in the context of the *Tara* Oceans expeditions[20]. Sampling stations were selected to represent distinct marine ecosystems at a global scale[49]. Note that Southern Ocean stations were not examined herein because they were ranked as outliers due to their exceptional environmental characteristics and biota[23,24]. Environmental data were obtained from vertical profiles of a sampling package[50,51]. It consisted of conductivity and temperature sensors, chlorophyll and CDOM fluorometers, light transmissometer (Wetlabs C-star 25cm), a backscatter sensor (WetLabs ECO BB), a nitrate sensor (SATLANTIC ISUS) and a Hydroptic Underwater Vision Profiler (UVP; Hydroptics[52]. Nitrate and fluorescence to chlorophyll concentrations as well as salinity were calibrated from water samples collected with Niskin bottle[50]. Net Primary Production (NPP) data were extracted from 8 day composites of the Vertically Generalized Production Model (VGPM[53]) at the week of sampling[54]. Carbon fluxes and carbon export, corresponding to the carbon flux at 150 m, were estimated based on particle concentration and size distributions obtained from the UVP[51] and details are presented below.

From particle size distribution to carbon export estimation

Previous research has shown that the distribution of particle size follows a power law over the μm to the mm size range[3,55,56]. This *Junge*-type distribution translates into the following mathematical equation, whose parameters can be retrieved from UVP images:

$$n(d) = ad^k \qquad \text{(eq. 1)}$$

where *d* is the particle diameter, and exponent *k* is defined as the slope of the number spectrum when equation (2) is log transformed. This slope is commonly used as a descriptor of the shape of the aggregate size distribution.

The carbon-based particle size approach relies on the assumption that the total carbon flux of particles (*F*) corresponds to the flux spectrum integrated over all particle sizes:

$$F = \int_0^\infty n(d).\, m(d).\, w(d) dd \qquad \text{(eq. 2)}$$

where *n*(*d*) is the particle size spectrum, i.e., equation (1), and *m*(*d*) is the mass (here carbon content) of a spherical particle described as:

$$m(d) = \alpha d^3 \qquad \text{(eq. 3)}$$

where $\alpha = \pi\rho/6$, ρ is the average density of the particle, and *w*(*d*) is the settling rate calculated using Stokes Law:

$$w(d) = \beta d^2 \qquad \text{(eq. 4)}$$

where $\beta = g(\rho - \rho_0)(18\nu\rho_0)^{-1}$, g is the gravitational acceleration, $\rho_0$the fluid density, and ν the kinematic viscosity.

In addition, mass and settling rates of particles, *m(d)* and *w(d)*, respectively, are often described as power law functions of their diameter obtained by fitting observed data, $m(d).\, w(d) = Ad^B$. The particles carbon flux can then be estimated using an approximation of Eq. 2 over a finite number (x) of small logarithmic intervals for diameter d spanning from 250 μm to 1.5 mm (particles <250 μm and >1.5 mm are not considered, consistent with the method presented by *Guidi et al., [2008]*[57]) such as

$$F = \sum_{i=1}^{x} n_i A d_i^B \vartriangle d_i \qquad \text{(eq. 5)}$$

where A=12.5±3.40 and B=3.81 ± 0.70 have been estimated using a global dataset that compared particle fluxes in sediment traps and particle size distributions from the UVP images.

Genomic data collection

For the sake of consistency between all available datasets from the *Tara* Oceans expeditions, we considered subsets of the data recently published in Science[23-25]. In brief, one sample corresponds to data collected at one depth (surface (SRF) or Deep Cholorophyll Maximum (DCM) determined from the profile of chlorophyll fluorometer) and at one station. To study the eukaryotic community in our current manuscript, we selected stations at which we had environmental data and carbon export estimated at 150 m with the UVP and all size fractions. Consequently a subset of 33 stations (corresponding to 56 samples) has been created compared to the 47 stations analyzed in *de Vargas et al.* [2015]. A similar procedure has been applied to the prokaryotic and viral datasets, reducing the *Sunagawa et al.* [2015] prokaryotic dataset to a subset of 104 samples from 62 stations and the *Brum et al.* [2015] viral dataset into a subset of 37 samples from 22 stations (See Supplementary Information SI10). In addition a detailed table is provided summarizing which samples (depth and station) are available for each domain (Supplementary Information SI11).

Eukaryotic taxa profiling

Photic-zone eukaryotic plankton diversity has been investigated through millions of environmental Illumina reads. Sequences of the 18S ribosomal RNA gene V9 region were obtained by PCR amplification and a stringent quality-check pipeline has been applied to remove potential chimera or rare sequences (details on data cleaning in *de Vargas et al.* [2015][24]). For 47 stations, and if possible at two depths (SRF and DCM), eukaryotic communities were sampled in the *piconano*- (0.8-5 µm), *micro*- (20-180 µm) and *meso*-plankton (180-2000 µm) fractions (a detailed list of these samples is given in Supplementary Information SI12). In the framework of the carbon export study, sequences from all size fractions were pooled in order to get the most accurate and statistically reliable dataset of the eukaryotic community. The 2.3 million eukaryotic ribotypes were assigned to known eukaryotic taxonomic entities by global alignment to a curated database[24]. To get the most accurate vision of the eukaryotic community, sequences showing less than 97% identity with reference sequences were excluded. The final eukaryotic relative abundance matrix used in our analyses included 1,750 lineages (taxonomic assignation has been performed using a last common ancestor methodology, and had thus been performed down to species level when possible) in 56 samples from 33 stations. Pooled abundance (number of V9 sequences) of each lineage has been normalized by the total sum of sequences in each sample.

Prokaryotic taxa profiling

To investigate the prokaryotic lineages, communities were sampled in the pico-plankton. Both filter sizes have been used along the *Tara* Oceans transect: up to station #52, prokaryotic fractions correspond to a 0.22-1.6 μm size fraction, and from station #56, prokaryotic fractions correspond to a

0.22-3 μm size fraction. Prokaryotic taxonomic profiling was performed using 16S rRNA gene tags directly identified in Illumina-sequenced metagenomes ($_{\mathrm{mi}}$tags) as described in *Logares et al.,* [2014][58]. 16S $_{\mathrm{mi}}$tags were mapped to cluster centroids of taxonomically annotated 16S reference sequences from the SILVA database[59] (release 115: SSU Ref NR 99) that had been clustered at 97% sequence identity using USEARCH v6.0.307[60]. 16S $_{\mathrm{mi}}$tag counts were normalized by the total reads count in each sample (further details in *Sunagawa et al.* [2015][23]). The photic-zone prokaryotic relative abundance matrix used in our analyses included 3,253,962 $_{\mathrm{mi}}$tags corresponding to 1,328 genera in 104 samples from 62 stations.

Prokaryotic functional profiling

For each prokaryotic sample, gene relative abundance profiles were generated by mapping reads to the OM-RGC using the MOCAT pipeline[61]. The relative abundance of each reference gene was calculated as gene length-normalized base counts. And functional abundances were calculated as the sum of the relative abundances of these reference genes, annotated to OG functional groups. In our analyses, we used the subset of the OM-RGC that was annotated to Bacteria or Archaea (24.4 M genes). Using a rarefied (to 33 M inserts) gene count table, an OG was considered to be part of the ocean microbial core if at least one insert from each sample was mapped to a gene annotated to that OG. For further details on the prokaryotic profiling please refer to *Sunagawa et al.* [2015][23]. The final prokaryotic functional relative abundance matrix used in our analyses included 37,832 OGs or functions in 104 samples from 62 stations. Genes from functions of FNET1 and FNET2 subnetworks were taxonomically annotated using a modified dual BLAST-based last common ancestor (2bLCA) approach[62]. We used RAPsearch2[63] rather than BLAST to efficiently process the large data volume and a database of non-redundant protein sequences from UniProt (version: UniRef_2013_07) and eukaryotic transcriptome data not represented in UniRef (see Supplementary Information SI5, SI6, for full annotations).

Enumeration of prokaryotes by flow cytometry

For prokaryote enumeration by flow cytometry, three aliquots of 1 ml of seawater (pre-filtered by 200-µm mesh) were collected from both SRF and DCM. The samples were fixed immediately using cold 25% glutaraldehyde (final concentration 0.125%), left in the dark for 10 min at room temperature, flash-frozen and kept in liquid nitrogen on board and then stored at -80°C on land. Two subsamples were taken to separate counts of heterotrophic prokaryotes (not shown herein) and phototrophic picoplankton. For heterotrophic prokaryote determination, 400 µl of sample was added to a diluted SYTO-13 (Molecular Probes Inc., Eugene, OR, USA) stock (10:1) at 2.5 μmol l^{-1} final concentration, left for about 10 min in the dark to complete the staining and run in the flow cytometer. We used a FacsCalibur (Becton & Dickinson) flow cytometer equipped with a 15 mW Argon-ion laser (488 nm emission). At least 30,000 events were acquired for each subsample (usually 100,000 events). Fluorescent beads (1 µm, Fluoresbrite carboxylate microspheres, Polysciences Inc., Warrington, PA) were added at a known density as internal standards. The bead standard concentration was determined by epifluorescence microscopy. For phototrophic picoplankton, we used the same procedure as for heterotrophic prokaryote, but without addition of SYTO-13. Data analysis was performed with FlowJo software (Tree Star, Inc.).

Profiling of viral populations

In order to associate viruses to carbon export we used viral populations as defined in *Brum et al.* [2015][25] using a set of 43 *Tara* Oceans viromes. Briefly, viral populations were defined as large contigs (>10 predicted genes and >10 kb) identified as most likely originating from bacterial or archaeal viruses. These 6,322 contigs remained and were then clustered into populations if they shared more than 80% of their genes at >95% nucleotide identity. This resulted in 5,477 'populations' from the 6,322 contigs, where as many as 12 contigs were included per population. For each population, the longest contig was chosen as the 'seed' representative sequence. The relative abundance of each population was computed by mapping all quality-controlled reads to the set of 5,477 non-redundant populations (considering only mapping quality scores greater than 1) with Bowtie2[64] and if more than 75% of the reference sequence was covered by virome reads. The relative abundance of a population in a sample was computed as the number of base pairs recruited to the contig normalized to the total number of base pairs available in the virome and the contig length if more than 75% of the reference sequence was covered by virome reads, and set to 0 otherwise (see *Brum et al.* [2015][25] for further details). The final viral population abundance matrix used in our analyses included 5,291 viral population contigs in 37 samples from 22 stations.

Viral host predictions

The longest contig in a population was defined as the seed sequence and considered the best estimate of that population's origin. These seed sequences were used to assess taxonomic affiliation of each viral population. Cases where >50% of the genes were affiliated to a specific reference genome from RefSeq Virus (based on a BLASTp comparison with thresholds of 50 for bit score and 10^{-5} for e-value) with an identity percentage of at least 75% (at the protein sequence level) were considered as confident affiliations to the corresponding reference virus. The viral population host group was then estimated based on these confident affiliations (see Supplementary Information SI13 for host affiliation of viral population contigs associated to carbon export).

Viral protein clusters

Viral protein clusters (PCs) correspond to ORFs initially mapped to existing clusters (POV, GOS and phage genomes). The remaining, unmapped ORFs were self-clustered, using cd-hit as described in *Brum et al.* [2015][25]. Only PCs with more than two ORFs were considered bona fide and were used for subsequent analyses. To compute PC relative abundance for statistical analyses, reads were mapped back to predicted ORFs in the contigs dataset using Mosaik as described in *Brum et al.* [2015][25]. Read counts to PCs were normalized by sequencing depth of each virome. Importantly, we restricted our analyses to 4,294 PCs associated to the 277 viral population contigs significantly associated to carbon export in 37 samples from 22 stations.

Sparse Partial Least Squares analysis

In order to directly associate eukaryotic lineages to carbon export and other environmental traits (Fig. 2), we used sparse Partial Least Square (sPLS[65] as implemented in the R package *mixOmics*[29]. We applied the sPLS in regression mode, which will model a causal relationship between the lineages and the environmental traits, *i.e.* PLS will predict environmental traits (*e.g.* carbon export) from lineage abundances. This approach enabled us to identify high correlations (see Supplementary Information SI1) between certain lineages and carbon export but without taking into account the global structure of the planktonic community.

Co-occurrence network model analysis

Weighted correlation network analysis (WGCNA) was performed to delineate feature (lineages, viral populations, PCs or functions) subnetworks based on their relative abundance[66,67]. A signed adjacency measure for each pair of features was calculated by raising the absolute value of their Pearson correlation coefficient to the power of a parameter p. The default value p=6 was used for each global network, except for the Prokaryotic functional network where p had to be lowered to 4 in order to optimize the scale-free topology network fit. Indeed, this power allows the weighted correlation network to show a scale free topology where key nodes are highly connected with others. The obtained adjacency matrix was then used to calculate the topological overlap measure (TOM), which for each pair of features, taking into account their weighted pairwise correlation (direct relationships) and their weighted correlations with other features in the network (indirect relationships). For identifying subnetworks a hierarchical clustering was performed using a distance based on the TOM measure. This resulted in the definition of several subnetworks, each represented by its first principal component.

These characteristic components play a key role in weighted correlation network analysis. On the one hand, the closeness of each feature to its cluster, referred to as the subnetwork membership, is measured by correlating its relative abundance with the first principal component of the subnetwork. On the other hand, association between the subnetworks and a given trait is measured by the pairwise Pearson correlation coefficients between the considered environmental trait and their respective principal components. A similar protocol has been performed on the eukaryotic relative abundance matrix, the prokaryotic relative abundance matrix, the prokaryotic functions relative abundance matrix and the viral population and PC relative abundance matrices. All procedures were applied on Hellinger-transformed log-scaled abundances. Noteworthy, the protocol is not sensitive to copy number variation as observed across different eukaryotic species, because the association between two species relies on a correlation score between relative abundance measurements. Computations were carried out using the R package *WGCNA*[33].

Given the nature of the eukaryotic dataset (three distinct size fractions), the sampling process may lead to the loss of size fractions. In particular, samples #1, #3, #17, #37, #39, #43, #48, #53, #54, #55, #66 are eventually biases by such a loss (Supplementary Information SI12). A complementary WGCNA analysis was performed with addition of these samples to evaluate the robustness of our protocol to missing size fractions. The composition of the eukaryotic subnetwork built with an extended dataset (*i.e.,* 67 samples from 37 stations for which size fractions were missing in 11 samples) was compared to the subnetwork as presented above (*i.e.,* 56 samples from 33 stations). Both subnetworks shown an overlap of 75% of lineage, whereas four of the top five VIP lineages with the extended dataset (see Extended data Fig. 8 for details) can be found in the top six VIP lineages of the above subnetwork (Supplementary Information SI2), emphasizing highly similar results and a small sensitivity to size fraction loss.

Extraction of subnetworks related to carbon export

For each subnetwork (called modules within WGCNA) extracted from each global network, pairwise Pearson correlation coefficients between the subnetwork principal components and the carbon export estimation was computed, as well as corresponding p-values corrected for multiple testing using the Benjamini & Hochberg FDR procedure. The subnetworks showing the highest correlation scores are of interest and were investigated. One subnetwork (49 nodes) was significant within the eukaryotic network; one subnetwork (109 nodes) was significant for the prokaryotic network; one subnetwork (277 nodes) was significant within the virus network; two subnetworks (441 and 220 nodes) were significant within the prokaryotic functional network, and two subnetworks (1,879 and 2,147 nodes) were significant within the viral PCs network.

Partial Least Squares regression

In addition to the network analyses, we asked whether the identified subnetworks can be used as predictors for the carbon export estimations. To answer this question, we used Partial least squares (PLS) regression, which is a dimensionality-reduction method that aims at determining predictor combinations with maximum covariance with the response variable. The identified combinations, called latent variables, are used to predict the response variable. The predictive power of the model is assessed by correlating the predicted vector with the measured values. The significance of the prediction power was evaluated by permuting the data 10,000 times. For each permutation, a PLS model was built to predict the randomized response variable and a Pearson correlation was calculated between the permuted response variable and in Leave-One-Out Cross-Validation (LOOCV) predicted values. The 10,000 random correlations are compared to the performance of the PLS model that were used to predict the true response variable. In addition, the predictors were ranked according to their value importance in projection (VIP)[68]. The VIP measure of a predictor estimates its contribution in the PLS regression. The predictors having high VIP values are assumed important for the PLS prediction of the response variable. The VIP values of the prokaryotic functional subnetworks are provided in Supplementary Information SI5, SI6. For the sake of illustration, only lineages or functions with VIP > 1[68] are discussed and pictured in Figure 4 and 5. Our computations were carried out using the R package pls[69]. All programs are available under GPL Licence.

Subnetwork representations

Nodes of the subnetworks represent either lineages (eukaryotic, prokaryotic or viral) or functions (prokaryotic or viral). Subnetworks related to the carbon export have been represented in two distinct formats. Scatter plots represent each nodes based on their Pearson correlation to the carbon export and their respective node centrality within the subnetwork. The latter has been recomputed using significant Spearman correlations above 0.3 (>0.9 for viral PCs) as edges, this is done for visualization purposes since WGCNA subnetworks (based on the Topology Overlap Measure (TOM) between nodes) are hyper-connected. Size representation of nodes are proportional to the VIP score after PLS. The hiveplots depict the same subnetworks by focusing on two main features: x-axis and y-axis depict nodes of subnetworks ranked by their VIP scores and Pearson correlation to the carbon export, respectively.

References and Notes (Methods)

49 Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* **2**, 150023, doi:10.1038/sdata.2015.23 (2015).

50 Picheral, M. *et al.* Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the Tara Oceans expedition 2009-2013. doi:10.1594/PANGAEA.836319 (2014).

51 Picheral, M. *et al.* Vertical profiles of environmental parameters measured from physical, optical and imaging sensors during Tara Oceans expedition 2009-2013. doi:10.1594/PANGAEA.836321 (2014).

52 Picheral, M. *et al.* The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnol. Oceanogr. Meth.* **8**, 462–473, doi:10:4319/lom.2010.8.462 (2010).

53 Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **42**, 1-20 (1997).

54 Chaffron, S. *et al.* Contextual environmental data of selected samples from the Tara Oceans Expedition (2009-2013). doi:10.1594/PANGAEA.840718 (2014).

55 McCave, I. N. Size spectra and aggregation of suspended particles in the deep ocean. *Deep-Sea Res. I.* **31**, 329-352 (1984).

56 Sheldon, R. W., Prakash, A. & Sutcliff, W. H. Size distribution of particles in ocean. *Limnol. Oceanogr.* **17**, 327-340 (1972).

57 Guidi, L. *et al.* Relationship between particle size distribution and flux in the mesopelagic zone. *Deep-Sea Res. I.* **55**, 1364-1374, doi:10.1016/j.dsr.2008.05.014 (2008).

58 Logares, R. *et al.* Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* **16**, 2659-2671, doi:Doi 10.1111/1462-2920.12250 (2014).

59 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-D596, doi:10.1093/Nar/Gks1219 (2013).

60 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461, doi:10.1093/Bioinformatics/Btq461 (2010).

61 Kultima, J. R. *et al.* MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS ONE* **7**, ARTN e47656, doi:10.1371/journal.pone.0047656 (2012).

62 Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678-1695, doi:10.1038/Ismej.2013.59 (2013).

63 Zhao, Y. A., Tang, H. X. & Ye, Y. Z. RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* **28**, 125-126, doi:10.1093/Bioinformatics/Btr595 (2012).

64 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-U354, doi:10.1038/Nmeth.1923 (2012).

65 Shen, H. P. & Huang, J. H. Z. Sparse principal component analysis via regularized low rank matrix approximation. *J Multivariate Anal* **99**, 1015-1034, doi:10.1016/J.Jmva.2007.06.007 (2008).

66 Langfelder, P. & Horvath, S. Eigengene networks for studying the relationships between co-expression modules. *Bmc Syst Biol* **1**, Artn 54, doi:10.1186/1752-0509-1-54 (2007).

67 Li, A. & Horvath, S. Network neighborhood analysis with the multi-node topological overlap measure. *Bioinformatics* **23**, 222-231, doi:10.1093/Bioinformatics/Btl581 (2007).

68 Chong, I. G. & Jun, C. H. Performance of some variable selection methods when multicollinearity is present. *Chemometr. Intell. Lab.* **78**, 103-112, doi:10.1016/J.Chemolab.2004.12.011 (2005).

69 Mevik, B. H. & Wehrens, R. The pls package: Principal component and partial least squares regression in R. *J Stat Softw* **18**, 1-23 (2007).

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, Rega Institute, KU Leuven, 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, SAMOSA, ANR-13-ADAP-0010), European Union FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376), ERC Advanced Grant Award to CB (Diatomite: 294823), Gordon and Betty Moore Foundation grant (#3790 and #2631) and the UA Technology and Research Initiative Fund and the Water, Environmental, and Energy Solutions Initiative 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 to SGA, JSPS KAKENHI Grant Number 26430184 to HO, and FWO, BIO5, Biosphere 2 to MBS. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the Tara schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. 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). The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. This article is contribution number ZZZ of *Tara* Oceans.

Author Contributions

L.G., S.C., Lu.B. and D.E. designed the study and wrote the paper. C.D., M.P., J.P. and Sa.S. collected *Tara* Oceans samples. S.K-L managed the logistics of the *Tara* Oceans project. L.G. and M.P. analysed oceanographic data. S.C. and Lu.B. analysed taxonomic data. S.C., Lu.B., D.E. and S.R. performed the genomic and statistical analyses. A.L., Y.D., L.G., S.C., Lu.B. and D.E. produced and analysed the networks. E.K., C.B. and G.G. supervised the study. M.S., J.R., E.K., C.B. and G.G. provided constructive comments, revised and edited the manuscript. *Tara* Oceans coordinators provided a creative environment and constructive criticism throughout the study. All authors discussed the results and commented on the manuscript.

Author Information

Data described herein is available at EBI under the project identifiers PRJEB402, PRJEB6610 and PRJEB7988, PANGAEA[50,51,54], and a companion website (http://www.raeslab.org/companion/ocean-carbon-export.html). The data release policy regarding future public release of Tara Oceans data is described in *Pesant et al.,* [2015][49]. All authors approved the final manuscript. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to lguidi@obs-vlfr.fr, samuel.chaffron@vib-kuleuven.be, lucie.bittner@upmc.fr, damien.eveillard@univ-nantes.fr, Jeroen.Raes@vib-kuleuven.be, karsenti@embl.de, cbowler@biologie.ens.fr, gorsky@obs-vlfr.fr

Extended data legends:

Extended Data Figure 1: Overview of analytical methods used in the manuscript. **a,** Depiction of a standard pairwise analysis that considers a sequence relative abundance matrix for s samples (s x OTUs (Operational Taxonomic Units)) and its corresponding environmental matrix (s x p (parameters)). sPLS results emphasize OTU(s) that are the most correlated to environmental parameters. **b,** Depiction of a graph-based approach. Using only a relative abundance matrix (s x OTUs), WGCNA builds a graph where nodes are OTUs and edges represent significant co-occurrence. Co-occurrence scores between nodes are weights allocated to corresponding edges. These weights are magnified by a power-law function until the graph becomes scale-free. The graph is then decomposed within subnetworks (groups of OTUs) that are analyzed separately. One subnetwork (group of OTUs) is considered of interest when its topology is related to the trait of interest; in the current case carbon export. For each subnetwork (for instance the subnetwork related to carbon export), each OTU is spread within a feature space that plots each OTU based on its membership to the subnetwork (x-axis) and its correlation to the environmental trait of interest (i.e., carbon export). A good regression of all OTUs emphasizes the putative relation of the subnetwork topology and the carbon export trait (*i.e.* the more a given OTU defines the subnetwork topology, the more it is correlated to carbon export). **c,** Depiction of the machine learning (PLS) approach that was applied following subnetwork identification and selection. Greater VIP scores (*i.e.* larger circles) emphasized most important OTUs. VIP refers to Variable Importance in Projection and reflects the relative predictive power of a given OTU. OTUs with VIP score greater than one are considered as important in the predictive model and their selection do not alter the overall predictive power.

Extended Data Figure 2: Domain-specific ecological subnetworks associated to environmental parameters and species subnetwork structures correlate to carbon export. **a,b,c,** Global ecological networks were built for the 3 domains of life using the WGCNA methodology (see methods) and correlated to classical oceanographic parameters as well as carbon export (estimated at 150 m from particles size distribution and abundance). Each domain-specific global network is decomposed into smaller coherent subnetworks (depicted by distinct colours on the y-axis) and their eigen vector is correlated to all environmental parameters. Similar to a correlation at the network scale, this approach directly links subnetworks to environmental parameters (*i.e.* the more the taxa contribute to the subnetwork structure, the more their abundance are correlated to the parameter). The measure allows to identify subnetworks for which the overall structure is related to the carbon export. **a,** A single eukaryotic subnetwork (n=58, N=1'870) is strongly associated to carbon export (Pearson cor. 0.81, p = 5e^{-15}). **b,** A single prokaryotic subnetwork (n=109, N=1'527) is moderately associated to carbon export (Pearson cor. 0.32, p = 9^{e-03}). **c,** A single viral subnetwork (n=277, N=5'476) is strongly associated to carbon export (Pearson cor. 0.93, p = 2^{e-15}). **d,e,f,** The WGCNA approach directly links subnetworks to environmental parameters, *i.e.* the more the features contribute to the subnetwork structure (topology), the more their abundance are correlated to the parameter. This measure allows to identify subnetworks for which the overall structure, summarized as the eigen vector of the subnetwork, is related to the carbon export. **d,** The eukaryotic subnetwork structure correlates to carbon export (Pearson cor. = 0.87, p = 5^{e-16}). **e,** The prokaryotic subnetwork structure correlates to carbon export (Pearson cor. = 0.47, p = 5^{e-06}). **f,** The viral population subnetwork structure correlates to carbon export (Pearson cor. = 0.88, p = 6^{e-93}).

Extended Data Figure 3: Species subnetworks predict carbon export. PLS regression was used to predict carbon export using lineage abundances in selected subnetworks. LOOCV was performed and VIP scores computed for each lineage. **a,** The eukaryotic subnetwork predicts carbon export with a R^2 of 0.69. **b,** The prokaryotic subnetwork predicts carbon export with a R^2 of 0.60. **c,** The viral population subnetwork predicts carbon export with a R^2 of 0.89.

Extended Data Figure 4: *Synechococcus* (rather than *Prochlorococcus*) absolute cell counts correlate well to carbon export. **a,** *Prochlorococcus* cell counts estimated by flow cytometry do not correlate to carbon export (mean carbon flux at 150m, Pearson cor. = -0.13, p = 0.27). **b,** *Synechococcus* cell counts estimated by flow cytometry correlate significantly to carbon export

(Pearson cor. = 0.64, p = 4.0^{e-10}). **c,** *Synechococcus / Prochlorococcus* cell counts ratio correlates significantly to carbon export (Pearson cor. = 0.54, p = 4.0^{e-07}).

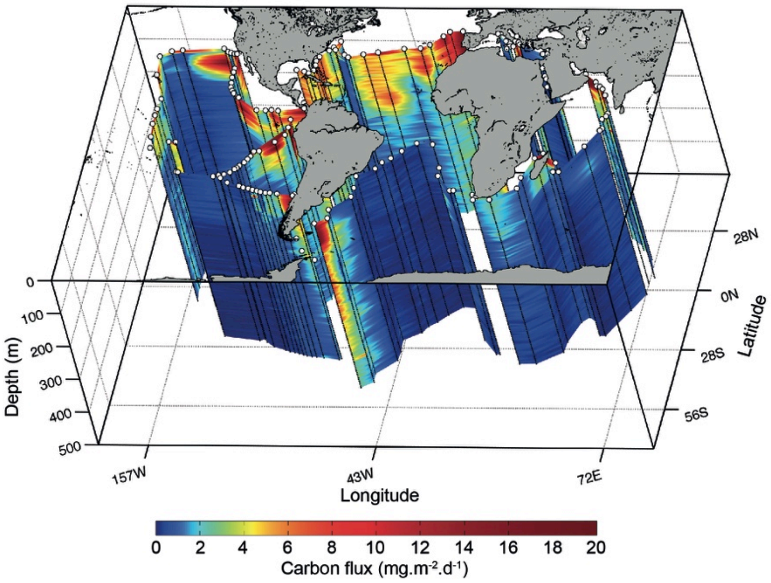
Extended Data Figure 5: Function and gene subnetworks associated to environmental parameters and their structure correlate to carbon export. **a,b,** Global ecological networks were built for the prokaryotic functions and viral PCs using the WGCNA methodology (see methods) and correlated to classical oceanographic parameters as well as carbon export. Each global network is decomposed into smaller coherent subnetworks (depicted by distinct colours on the y-axis) and their eigen vector is correlated to all environmental parameters. Similar to a correlation at the network scale, this approach directly links subnetworks to environmental parameters (*i.e.* the more the taxa contribute to the subnetwork structure, the more their abundance are correlated to the parameter). The measure allows to identify subnetworks for which the overall structure is related to the carbon export. **a,** Two bacterial functional subnetworks (n=441 and n=220, N=37'832) are associated to carbon export (Pearson cor. 0.54, p = 1^{e-07} and 0.42, p = 1^{e-04}). **b,** Two viral PCs subnetworks (n=1'879 and n=2'147, N=4'678) are strongly associated to carbon export (Pearson cor. 0.75, p = 3^{e-07} and 0.91, p = 3^{e-14}). **c,d** The WGCNA approach directly links subnetworks to environmental parameters, *i.e.* the more the features contribute to the subnetwork structure (topology), the more their abundance are correlated to the parameter. This measure allows to identify subnetworks for which the overall structure, summarized as the eigen vector of the subnetwork, is related to the carbon export. **c,** The bacterial function subnetwork structures correlates to carbon export (FNET1 Pearson cor. = 0.68, p = 3^{e-61}, and FNET2 Pearson cor. = 0.47, p = 6^{e-13}). **d,** The viral PC subnetwork structures correlates to carbon export (VNET1 Pearson cor. = 0.91, p < 1^{e-200}, and VNET2 Pearson cor. = 0.96, p < 1^{e-200}).

Extended Data Figure 6: Cumulative abnundance of genus-level taxonomic annotations of genes encoding functions from FNET1 and FNET2 subnetworks and Bacterial function subnetworks predict carbon export. **a,** Genes contributing to the relative abundance of FNET1 and FNET2 subnetwork functions were taxonomically annotated by homolgy searches against a non-redundant gene reference database using a last common ancestor (LCA) approach (see methods). **b,c,** PLS regression was used to predict carbon export using abundances of functions (OGs) in selected subnetworks. LOOCV was performed and VIP scores computed for each function. **b,** Light green subnetwork (FNET1) functions predict carbon export with a R^2 of 0.41. **c,** Dark green subnetwork (FNET2) functions predict carbon export with a R^2 of 0.48.

Extended Data Figure 7: Viral protein cluster networks reveal potential marker genes for carbon export prediction at global scale. **a,** A viral protein cluster (PC) network was built using abundances of PCs predicted from viral population contigs associated to carbon export (Fig. 3b) using the WGCNA methodology (see methods) and correlated to classical oceanographic parameters. Two viral PC subnetworks (light and dark orange, VNET1 and VNET2, left and right panel respectively) are strongly associated to carbon export (VNET1: Pearson cor. 0.75, p = 3^{e-07} and VNET2: 0.91, p = 3^{e-14}, Extended data figure 5b). Size of dots is proportional to the VIP score computed for the PLS regression. **b,** Viral PC subnetworks predict carbon export. PLS regression was used to predict carbon export using abundances of viral protein clusters (PCs) in selected subnetworks. LOOCV was performed and VIP scores computed for each PC. Light orange subnetwork (VNET1, left panel) PCs predict carbon export with a R^2 of 0.55. Dark orange subnetwork (VNET2, right panel) PCs predict carbon export with a R^2 of 0.89.

Extended Data Figure 8: WGCNA and PLS regression analyses for the full Eukaryotic dataset. **a,** A single eukaryotic subnetwork (n=58, is strongly associated to carbon export (Pearson cor. 0.79, p = 3^{e-14}). **b,** The eukaryotic subnetwork structure correlates to carbon export (Pearson cor. = 0.94, p = 4^{e-27}). **c,** The eukaryotic subnetwork predicts carbon export with a R^2 of 0.76. **d**, Lineages with the highest VIP score (dots size is proportional to the VIP score in the scatter plot) in the PLS are depicted as red dots corresponding to two rhizarian (Collodaria), one copepod (*Euchaeta*), and three dinophyceae (*Noctiluca scintillans, Gonyaulax polygramma and Gonyaulax sp. (clade 4)*).

889



890

891 Figure 1

892

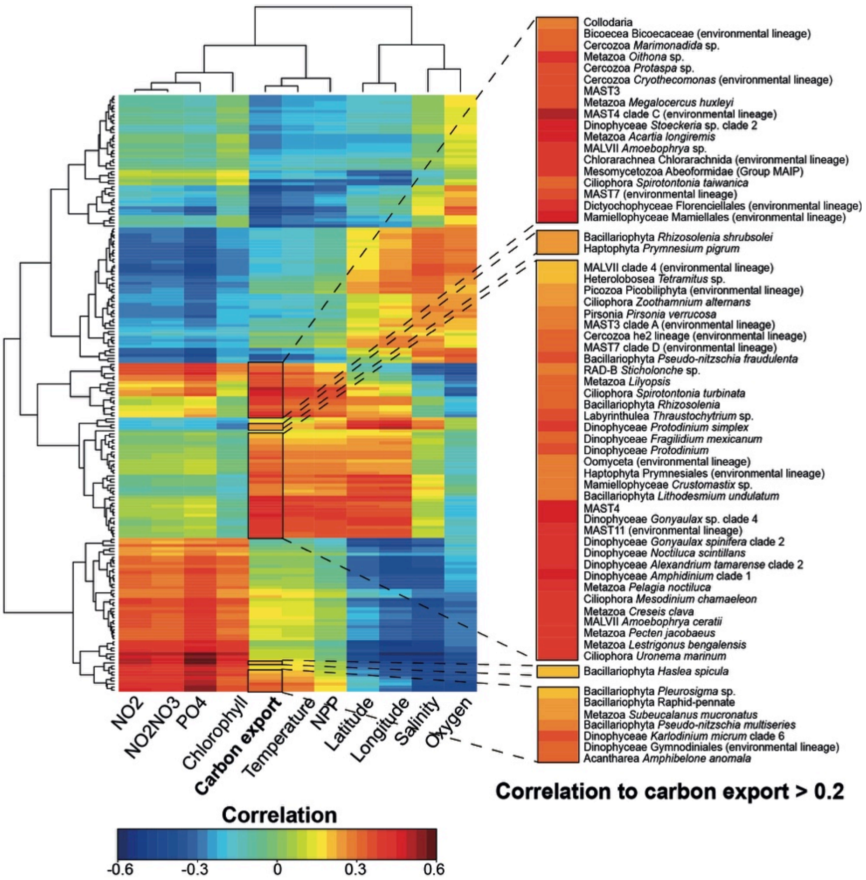

NO2
NO2NO3
PO4
Chlorophyll
Carbon export
Temperature
NPP
Latitude
Longitude
Salinity
Oxygen
Correlation
-0.6
-0.3
0
0.3
0.6
Collodaria
Bicoecea Bicoecaceae (environmental lineage)
Cercozoa *Marimonadida* sp.
Metazoa *Oithona* sp.
Cercozoa *Protaspa* sp.
Cercozoa *Cryothecomonas* (environmental lineage)
MAST3
Metazoa *Megalocercus huxleyi*
MAST4 clade C (environmental lineage)
Dinophyceae *Stoeckeria* sp. clade 2
Metazoa *Acartia longiremis*
MALVII *Amoebophrya* sp.
Chlorarachnea Chlorarachnida (environmental lineage)
Mesomycetozoa Abeoformidae (Group MAIP)
Ciliophora *Spirotontonia taiwanica*
MAST7 (environmental lineage)
Dictyochophyceae Florenciellales (environmental lineage)
Mamiellophyceae Mamiellales (environmental lineage)
Bacillariophyta *Rhizosolenia shrubsolei*
Haptophyta *Prymnesium pigrum*
MALVII clade 4 (environmental lineage)
Heterolobosea *Tetramitus* sp.
Picozoa Picobiliphyta (environmental lineage)
Ciliophora *Zoothamnium alternans*
Pirsonia *Pirsonia verrucosa*
MAST3 clade A (environmental lineage)
Cercozoa he2 lineage (environmental lineage)
MAST7 clade D (environmental lineage)
Bacillariophyta *Pseudo-nitzschia fraudulenta*
RAD-B *Sticholonche* sp.
Metazoa *Lilyopsis*
Ciliophora *Spirotontonia turbinata*
Bacillariophyta *Rhizosolenia*
Labyrinthulea *Thraustochytrium* sp.
Dinophyceae *Protodinium simplex*
Dinophyceae *Fragilidium mexicanum*
Dinophyceae *Protodinium*
Oomyceta (environmental lineage)
Haptophyta Prymnesiales (environmental lineage)
Mamiellophyceae *Crustomastix* sp.
Bacillariophyta *Lithodesmium undulatum*
MAST4
Dinophyceae *Gonyaulax* sp. clade 4
MAST11 (environmental lineage)
Dinophyceae *Gonyaulax spinifera* clade 2
Dinophyceae *Noctiluca scintillans*
Dinophyceae *Alexandrium tamarense* clade 2
Dinophyceae *Amphidinium* clade 1
Metazoa *Pelagia noctiluca*
Ciliophora *Mesodinium chamaeleon*
Metazoa *Creseis clava*
MALVII *Amoebophrya ceratii*
Metazoa *Pecten jacobaeus*
Metazoa *Lestrigonus bengalensis*
Ciliophora *Uronema marinum*
Bacillariophyta *Haslea spicula*
Bacillariophyta *Pleurosigma* sp.
Bacillariophyta Raphid-pennate
Metazoa *Subeucalanus mucronatus*
Bacillariophyta *Pseudo-nitzschia multiseries*
Dinophyceae *Karlodinium micrum* clade 6
Dinophyceae Gymnodiniales (environmental lineage)
Acantharea *Amphibelone anomala*
Correlation to carbon export > 0.2

Figure 2

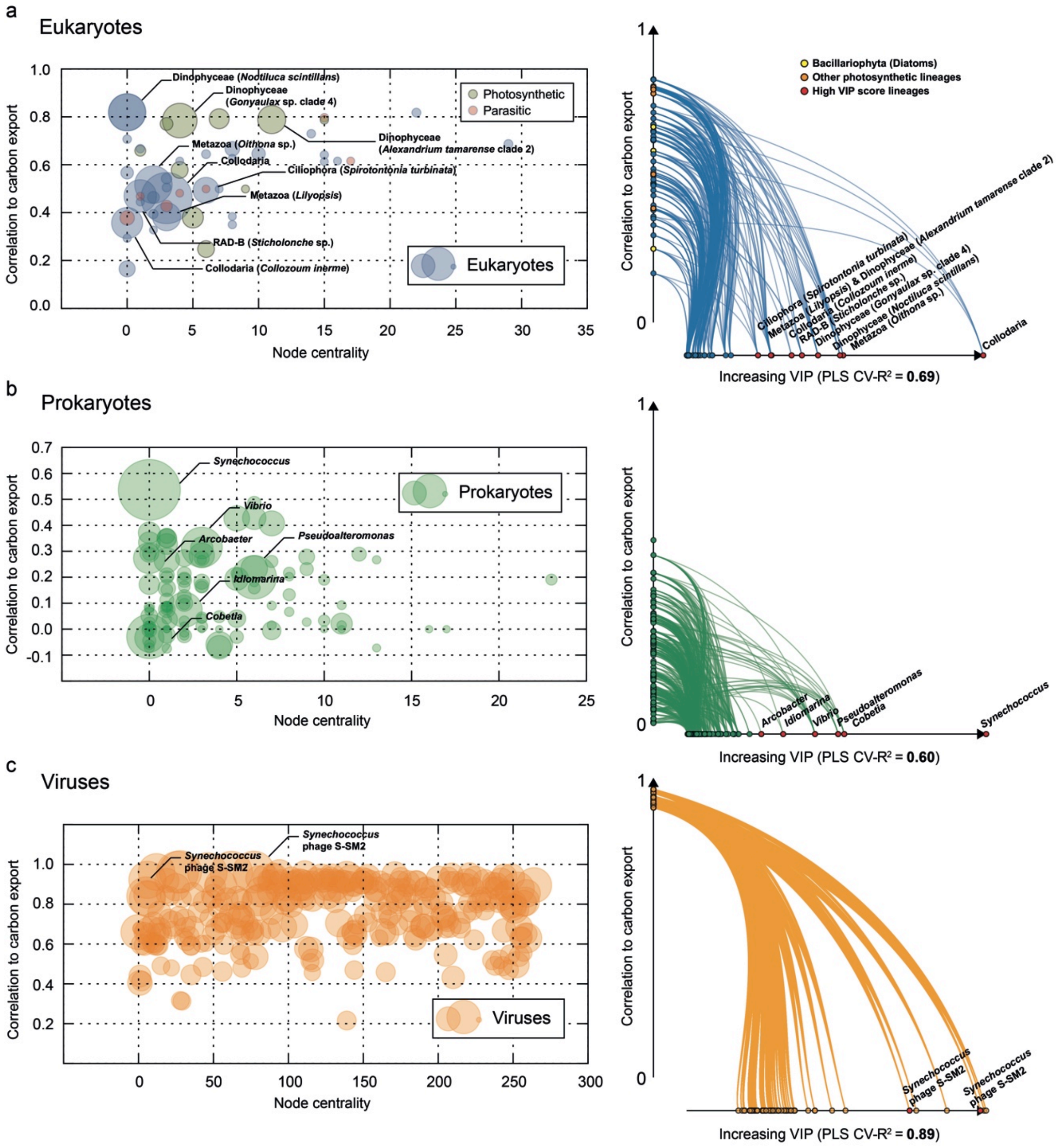

a
Eukaryotes
1.0
0.8
0.6
0.4
0.2
0.0
Correlation to carbon export
Node centrality
0
5
10
15
20
25
30
35
Dinophyceae (Noctiluca scintillans)
Dinophyceae (Gonyaulax sp. clade 4)
Dinophyceae (Alexandrium tamarense clade 2)
Metazoa (Oithona sp.)
Collodaria
Ciliophora (Spirotontonia turbinata)
Metazoa (Lilyopsis)
RAD-B (Sticholonche sp.)
Collodaria (Collozoum inerme)
Photosynthetic
Parasitic
Eukaryotes
1
0
Bacillariophyta (Diatoms)
Other photosynthetic lineages
High VIP score lineages
Ciliophora (Spirotontonia turbinata)
Metazoa (Lilyopsis) & Dinophyceae (Alexandrium tamarense clade 2)
Collodaria (Collozoum inerme)
RAD-B (Sticholonche sp.)
Dinophyceae (Gonyaulax sp. clade 4)
Dinophyceae (Noctiluca scintillans)
Metazoa (Oithona sp.)
Collodaria
Increasing VIP (PLS CV-R² = 0.69)
b
Prokaryotes
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
0
5
10
15
20
25
Synechococcus
Vibrio
Arcobacter
Pseudoalteromonas
Idiomarina
Cobetia
Prokaryotes
Arcobacter
Idiomarina
Vibrio
Pseudoalteromonas
Cobetia
Synechococcus
Increasing VIP (PLS CV-R² = 0.60)
c
Viruses
1.0
0.8
0.6
0.4
0.2
0
50
100
150
200
250
300
Synechococcus phage S-SM2
Synechococcus phage S-SM2
Viruses
Synechococcus phage S-SM2
Synechococcus phage S-SM2
Increasing VIP (PLS CV-R² = 0.89)

Figure 3

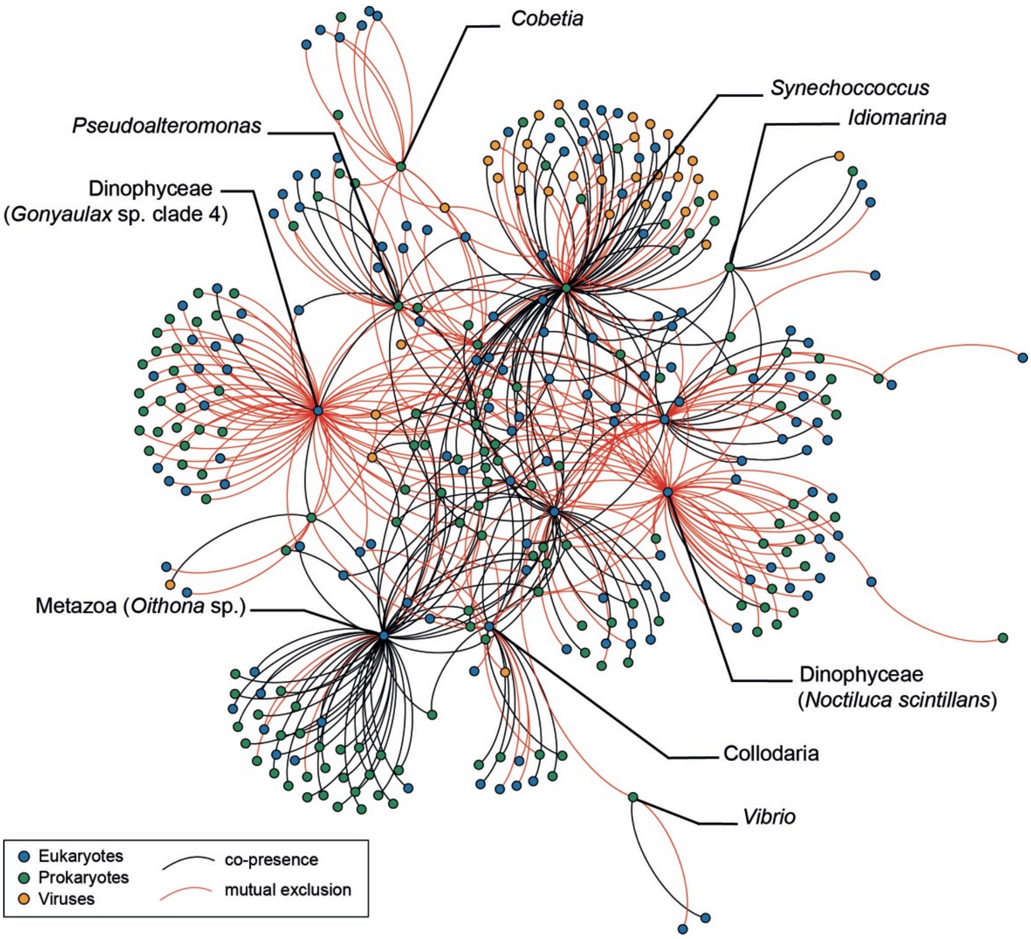

Cobetia
Synechoccoccus
Idiomarina
Pseudoalteromonas
Dinophyceae
(Gonyaulax sp. clade 4)
Metazoa (Oithona sp.)
Dinophyceae
(Noctiluca scintillans)
Collodaria
Vibrio
Eukaryotes
Prokaryotes
Viruses
co-presence
mutual exclusion

Figure 4

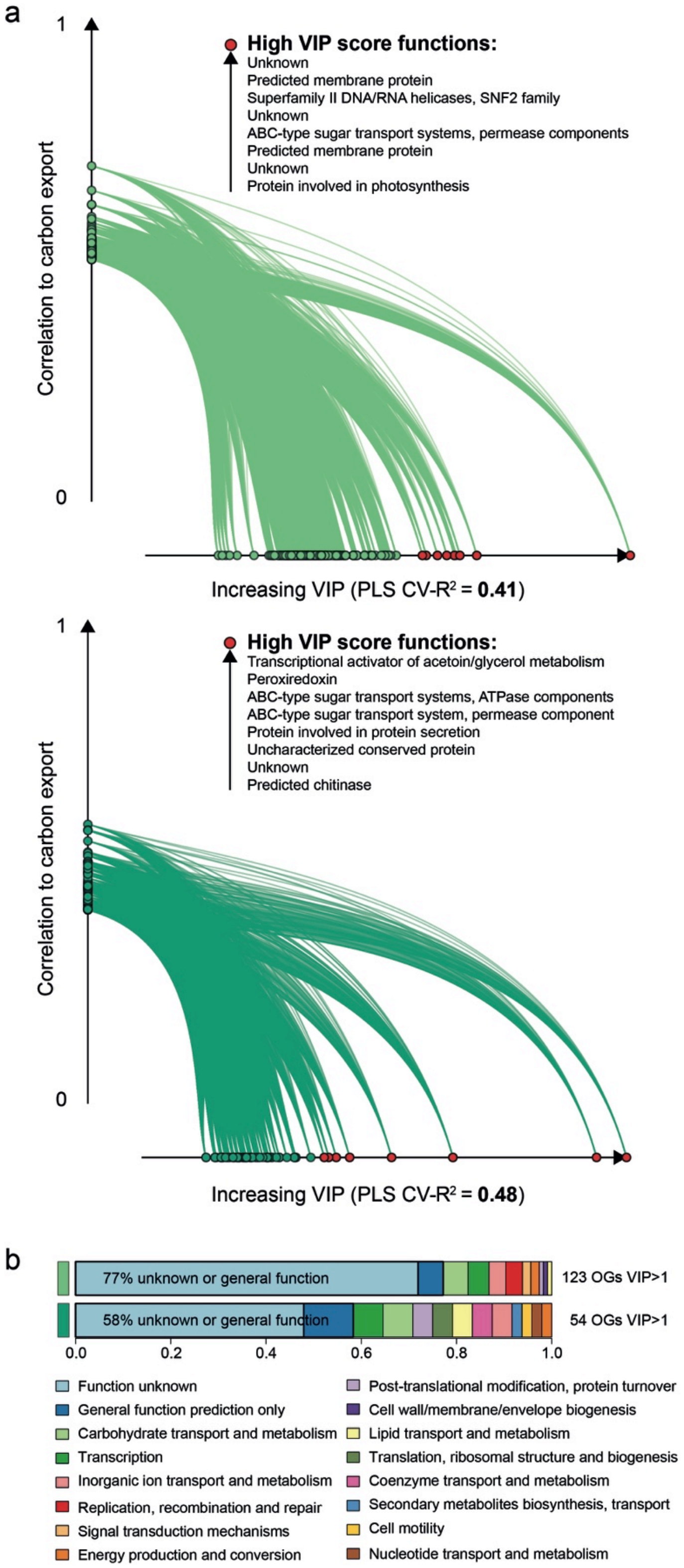
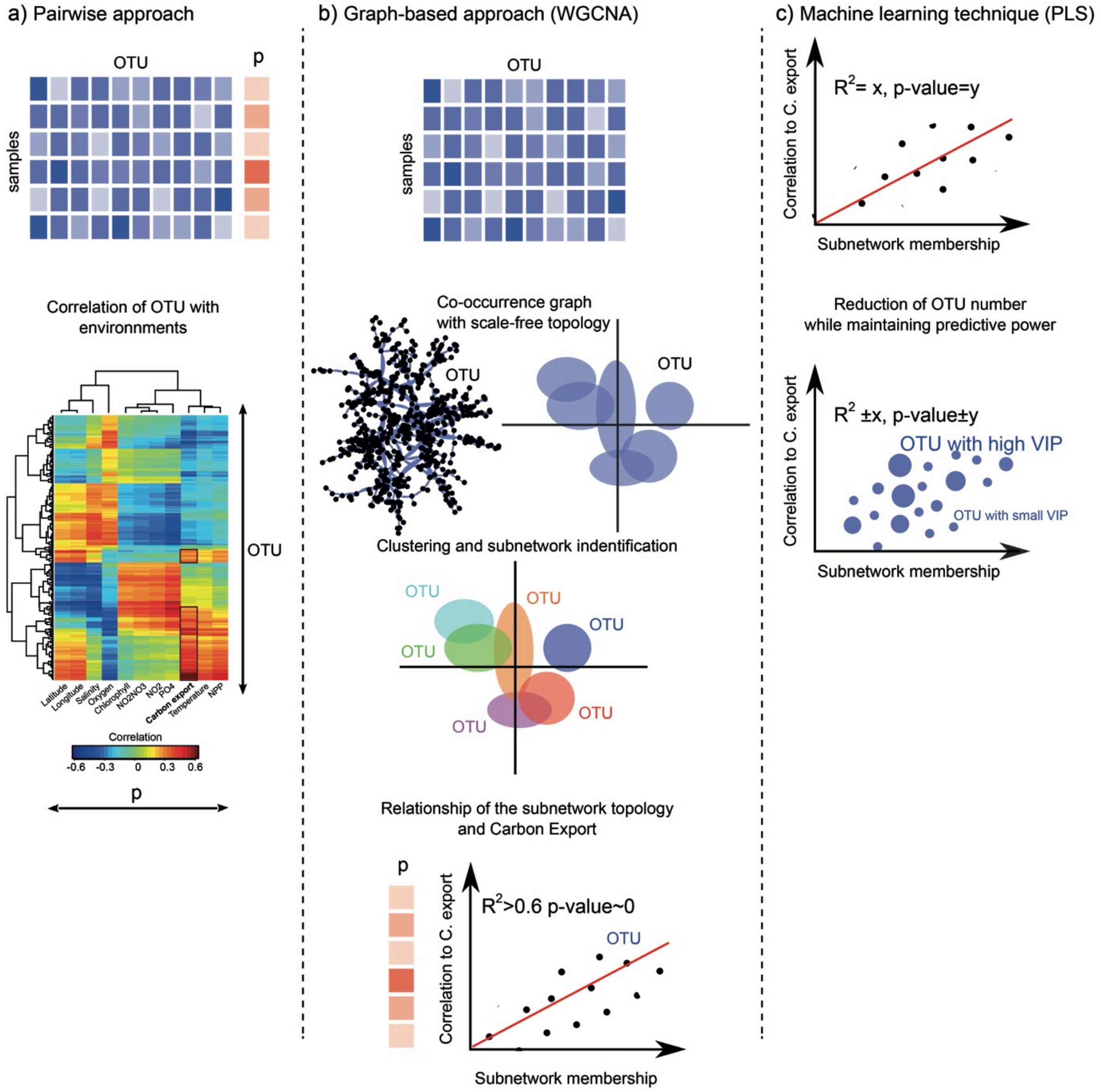
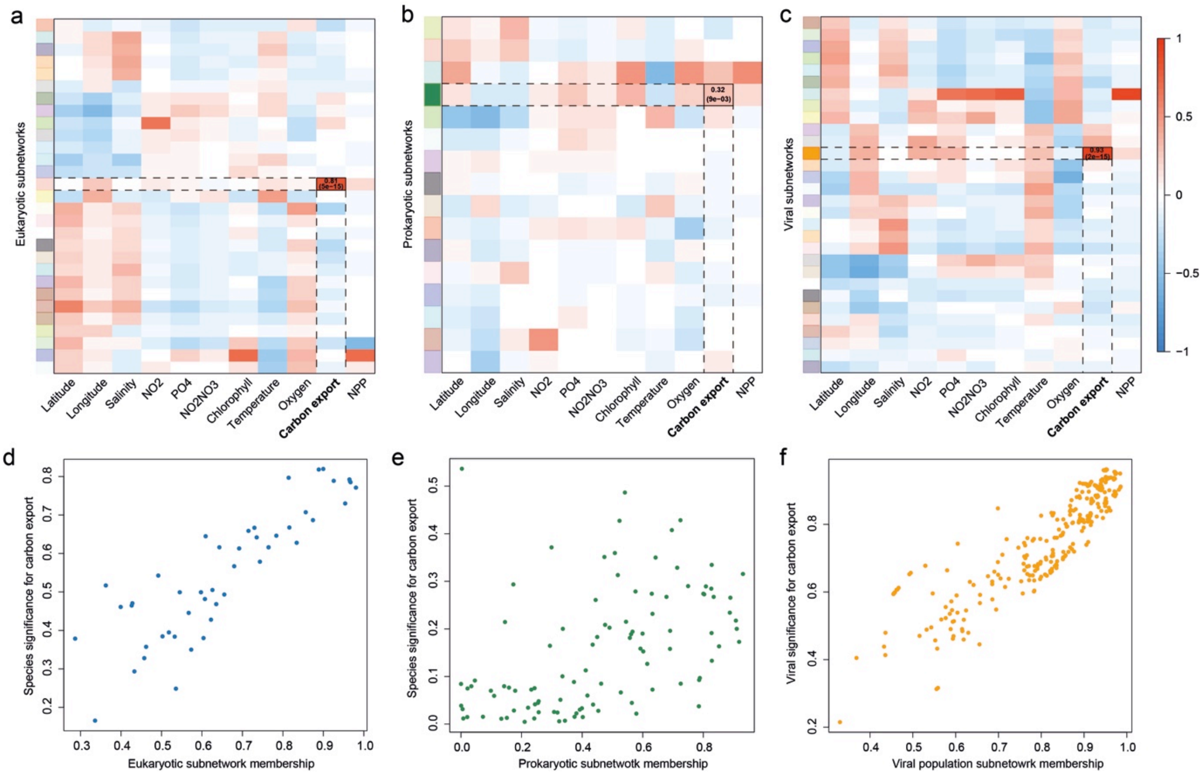

a
1
0
Correlation to carbon export
High VIP score functions:
Unknown
Predicted membrane protein
Superfamily II DNA/RNA helicases, SNF2 family
Unknown
ABC-type sugar transport systems, permease components
Predicted membrane protein
Unknown
Protein involved in photosynthesis
Increasing VIP (PLS CV-R² = 0.41)
1
0
Correlation to carbon export
High VIP score functions:
Transcriptional activator of acetoin/glycerol metabolism
Peroxiredoxin
ABC-type sugar transport systems, ATPase components
ABC-type sugar transport system, permease component
Protein involved in protein secretion
Uncharacterized conserved protein
Unknown
Predicted chitinase
Increasing VIP (PLS CV-R² = 0.48)
b
77% unknown or general function
123 OGs VIP>1
58% unknown or general function
54 OGs VIP>1
0.0
0.2
0.4
0.6
0.8
1.0
Function unknown
General function prediction only
Carbohydrate transport and metabolism
Transcription
Inorganic ion transport and metabolism
Replication, recombination and repair
Signal transduction mechanisms
Energy production and conversion
Post-translational modification, protein turnover
Cell wall/membrane/envelope biogenesis
Lipid transport and metabolism
Translation, ribosomal structure and biogenesis
Coenzyme transport and metabolism
Secondary metabolites biosynthesis, transport
Cell motility
Nucleotide transport and metabolism

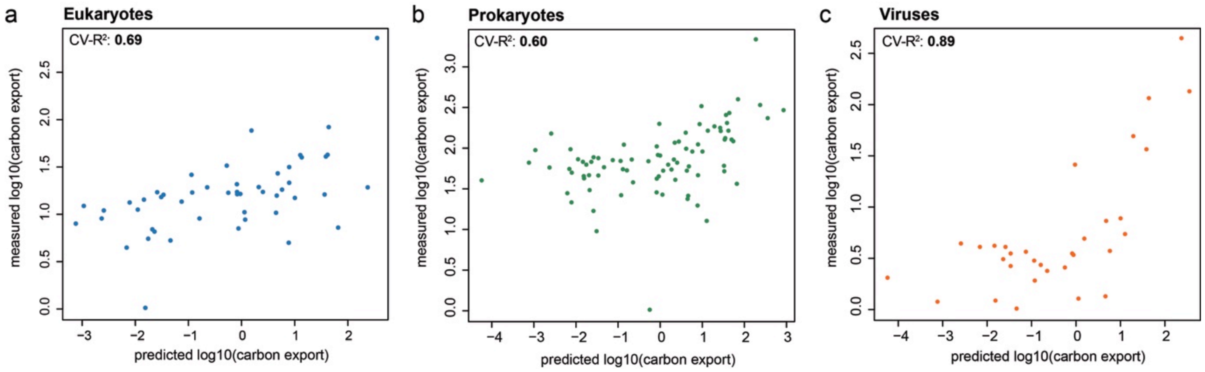
Figure 5



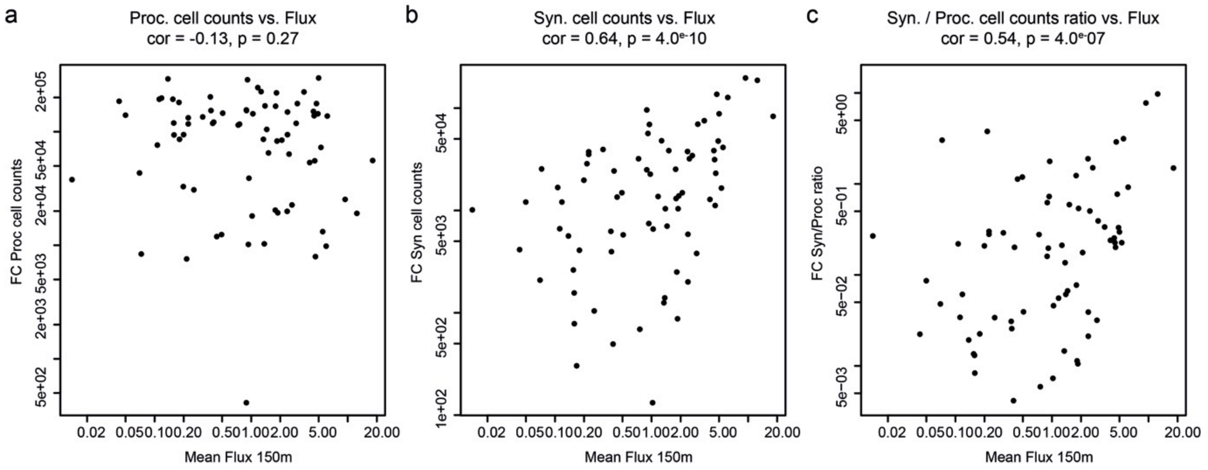
Extended Data Figure 1



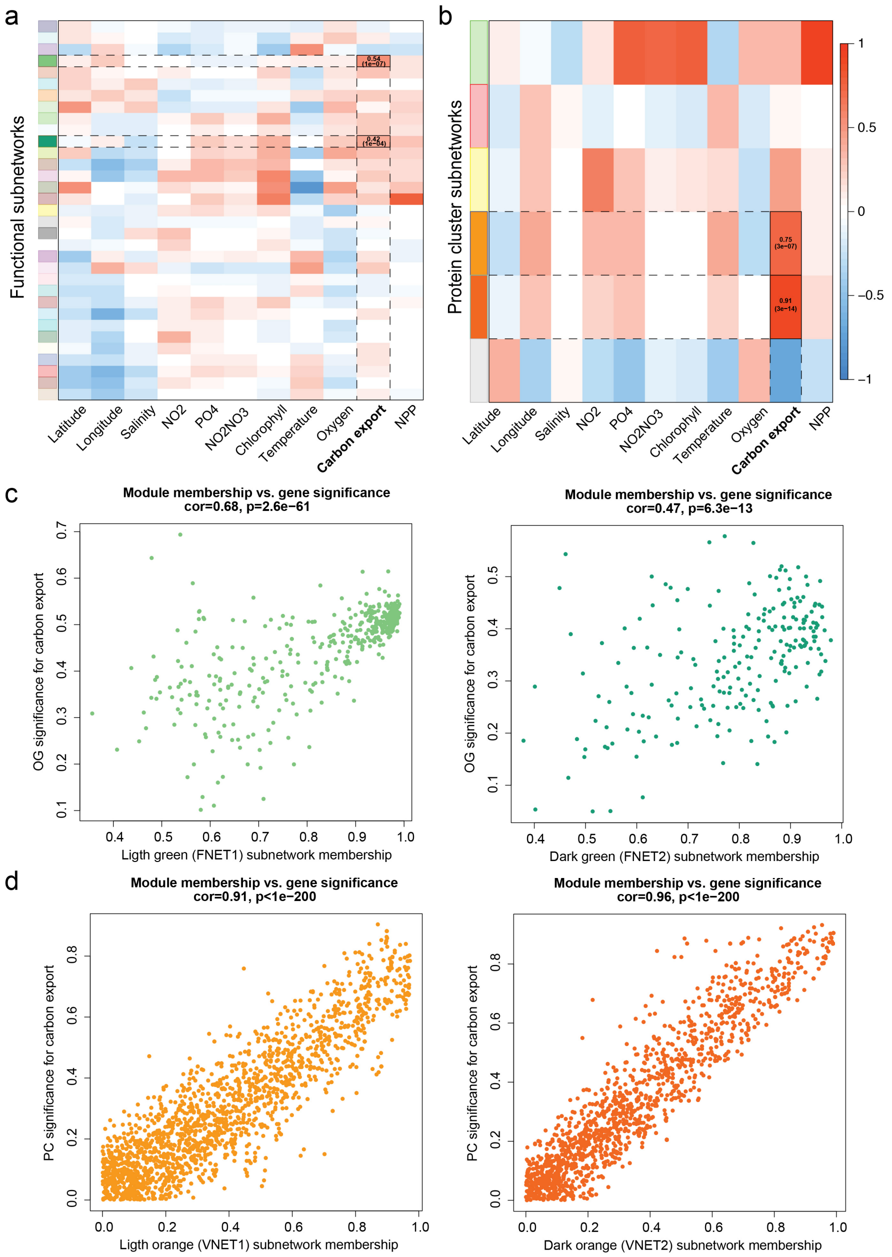
Extended Data Figure 2



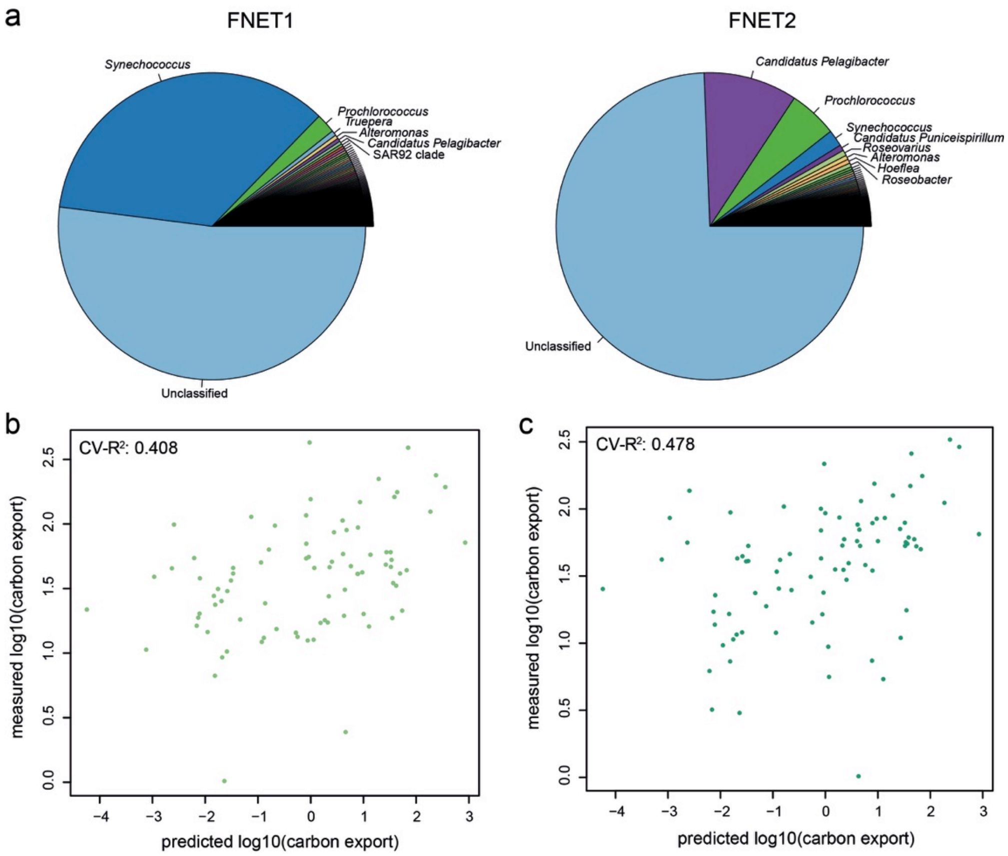
Extended Data Figure 3



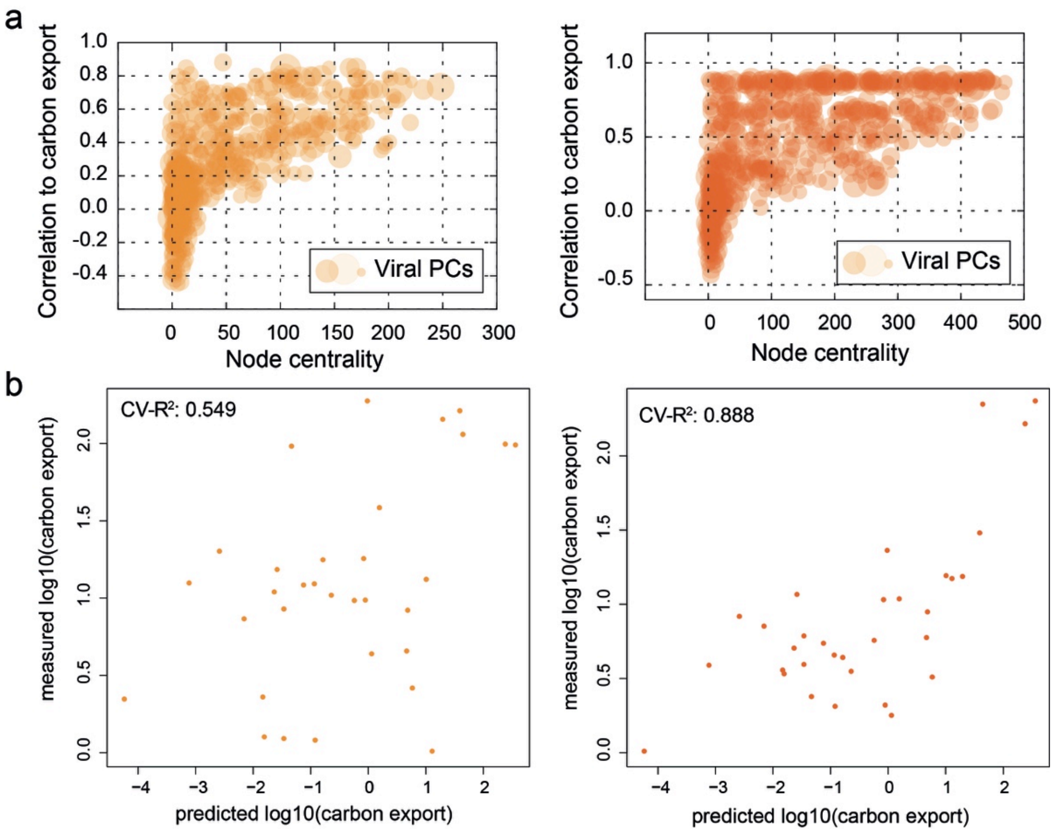
Extended Data Figure 4



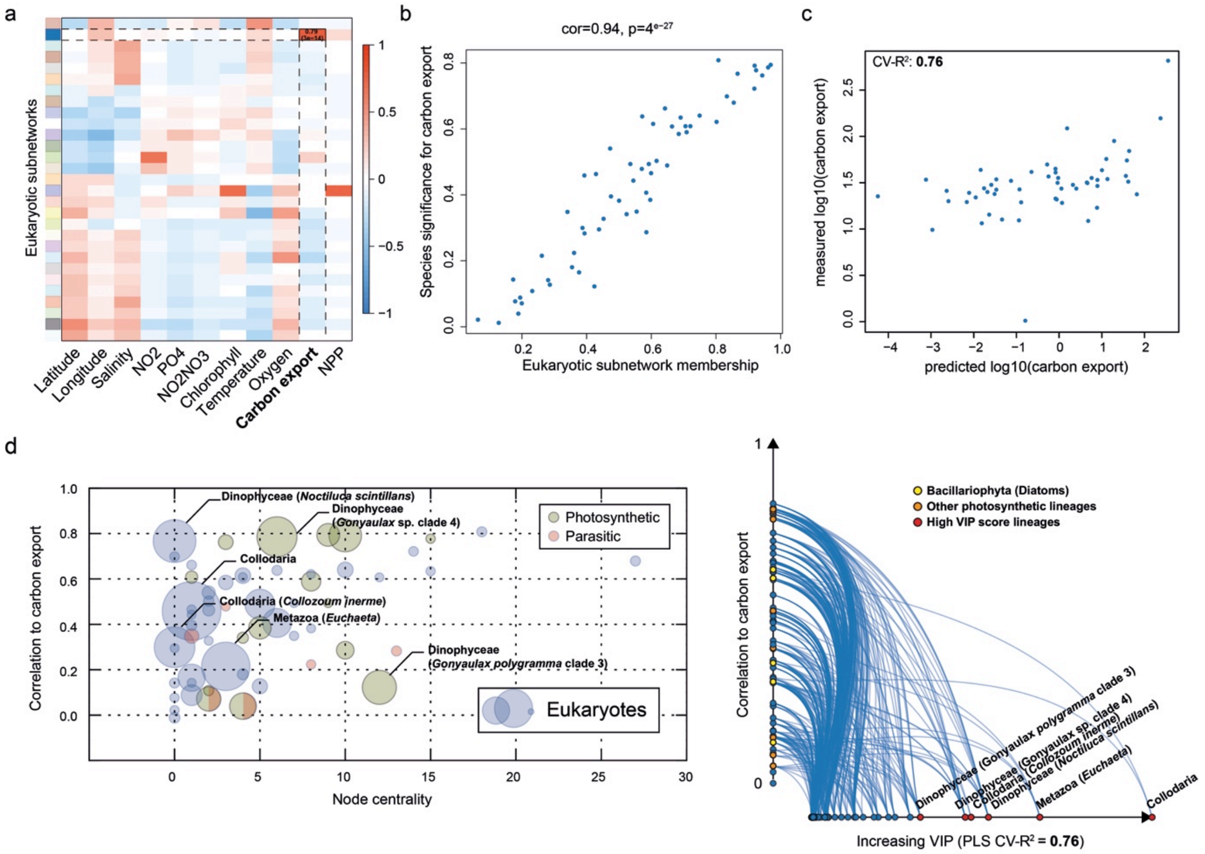
Extended Data Figure 5



Extended Data Figure 6



Extended Data Figure 7



Extended Data Figure 8