



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.

► To cite this version:

Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, et al.. Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, 2016, 532, pp.465-470. 10.1038/nature16942 . hal-01275276

HAL Id: hal-01275276

<https://hal.sorbonne-universite.fr/hal-01275276>

Submitted on 17 Feb 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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 Larhlimi⁹, Simon Roux^{10,11}, Youssef Darzi^{3,4}, Stephane Audic⁸, Léo Berline^{1,12}, Jennifer Brum^{10,11}, Luis Pedro Coelho¹³, Julio Cesar Ignacio Espinoza¹⁰, Shruti Malviya⁷, Shinichi Sunagawa¹³, Céline Dimier⁸, Stefanie Kandels-Lewis^{13,14}, Marc Picheral¹, Julie Poulain¹⁵, Sarah Searson^{1,2}, Tara Oceans coordinators, Lars Stemmann¹, Fabrice Not⁸, Pascal Hingamp¹⁶, Sabrina Speich¹⁷, Mick Follows¹⁸, Lee Karp-Boss¹⁹, Emmanuel Boss¹⁹, Hiroyuki Ogata²⁰, Stephane Pesant^{21,22}, Jean Weissenbach^{15,23,24}, Patrick Wincker^{15,23,24}, Silvia G. Acinas²⁵, Peer Bork^{13,26}, Colomban de Vargas⁸, Daniele Iudicone²⁷, Matthew B. Sullivan^{10,11}, Jeroen Raes^{3,4,5}, Eric Karsenti^{7,14}, Chris Bowler⁷, Gabriel Gorsky¹

* These authors contributed equally to this work

Affiliations:

- ¹ Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d’océanographie 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.

51 **The biological carbon pump is the process by which CO₂ is transformed to organic**
52 **carbon *via* photosynthesis, exported through sinking particles, and finally sequestered**
53 **in the deep ocean or sediment. While the intensity of the pump correlates with plankton**
54 **community composition, the underlying ecosystem structure and interactions driving**
55 **the process remain largely uncharacterised. Here we use environmental and**
56 **metagenomic data gathered during the *Tara* Oceans expedition to improve our**
57 **understanding of carbon export in the oligotrophic ocean. We show that specific**
58 **euphotic plankton communities correlate with carbon export and highlight unexpected**
59 **and overlooked taxa such as Radiolaria, alveolate parasites, as well as *Synechococcus***
60 **and their phages, as lineages most strongly associated with carbon export in the**
61 **subtropical, nutrient-depleted, oligotrophic ocean. Additionally, we show that the**
62 **relative abundance of just a few bacterial and viral genes can predict most of the**
63 **variability in carbon export in these regions.**

64 Marine planktonic photosynthetic organisms are responsible for approximately fifty percent
65 of Earth's primary production and they fuel the global ocean biological carbon pump¹. The
66 intensity of the pump is correlated to plankton community composition^{2,3}, and controlled by
67 the relative rates of primary production and carbon remineralisation⁴. About 10% of this
68 newly produced organic carbon in the surface ocean is exported through gravitational
69 sinking of particles. Finally, after multiple transformations, only a fraction of the exported
70 material will reach the deep ocean where it is sequestered over thousand-year timescales of
71 the ocean's overturning circulation⁵.

72 Like most biological systems, marine ecosystems in the sunlit upper layer of the ocean
73 (denoted the euphotic zone) are complex^{6,7}, characterised by a wide range of biotic and
74 abiotic interactions⁸⁻¹⁰ and in constant balance between carbon production, transfer to higher

75 trophic levels, remineralisation, and export to the deep layers¹¹. The marine ecosystem
76 structure and its taxonomic and functional composition likely evolved to comply with this
77 loss of energy by modifying organism turnover times and by the establishment of complex
78 feedbacks between them⁶ and the substrates they can exploit for metabolism¹². Decades of
79 groundbreaking research have focused on identifying independently the key players involved
80 in the biological carbon pump. Among autotrophs, diatoms are commonly attributed to being
81 important in carbon flux because of their large size and fast sinking rates¹³⁻¹⁵ while small
82 autotrophic picoplankton may contribute directly as a result of subduction of surface water
83 resulting from sub-mesoscale dynamic features¹⁶ or indirectly by aggregating with larger
84 settling particles or through their consumption by organisms at higher trophic levels¹⁷.
85 Among heterotrophs, zooplankton such as copepods impact carbon flux *via* production of
86 fast-sinking fecal pellets while migrating hundreds of meters in the water-column^{18,19}. These
87 observations, focusing on just a few components of the marine ecosystem, highlight that
88 carbon export results from multiple biotic interactions and that a better understanding of the
89 mechanisms involved in its regulation will likely require an analysis of the entire planktonic
90 ecosystem.

91 Advanced sequencing technologies now offer the opportunity to simultaneously survey
92 whole planktonic communities and associated molecular functions in unprecedented detail.
93 Such a holistic approach may allow the identification of community- or gene-based
94 biomarkers that could be used to monitor and predict ecosystem functions, e.g., related to the
95 biogeochemistry of the ocean²⁰⁻²². Here, we leverage global-scale ocean genomics
96 datasets^{10,23-25} and associated environmental data to assess the coupling between ecosystem
97 structure, functional repertoire, and the carbon export component of the biological carbon
98 pump.

99 **Carbon export and plankton community composition**

100 The *Tara* Oceans global circumnavigation crossed diverse ocean ecosystems and sampled
101 plankton at an unprecedented scale^{20,26} (see Methods). Hydrographic data were measured *in*
102 *situ* or in seawater samples at all stations, as well as nutrients, oxygen and photosynthetic
103 pigments (see Methods). Net Primary Production (NPP) was derived from satellite
104 measurements (see Methods). In addition, particle size distributions (100 μm to a few mm)
105 and concentrations were measured using an Underwater Vision Profiler (UVP) from which
106 carbon export, corresponding to the carbon flux (Fig. 1) at 150 m, was calculated to range
107 from 0.014 to 18.3 $\text{mg}\cdot\text{m}^{-2}\cdot\text{d}^{-1}$ using previously validated methods (see Methods). The
108 approach allowed us to assemble the largest homogeneous carbon flux dataset during a
109 single expedition, corresponding to more than 600 profiles over 150 stations. This dataset is
110 of similar magnitude to the body of historical data available in the literature that includes the
111 134 deep sediment trap-based carbon flux time-series²⁷ from the JGOFS program and the
112 419 thorium-derived particulate organic carbon (POC) export measurements²⁸.

113 From 68 globally distributed sites, a total of 7.2 Tb of metagenomics data, representing *circa*
114 40 million non-redundant genes, around 35,000 Operational Taxonomic Units (OTUs) of
115 prokaryotes (Bacteria and Archaea) and numerous mainly uncharacterized viruses and
116 picoeukaryotes, have been described recently^{23,25}. In addition, a set of 2.3 million eukaryotic
117 18S rDNA ribotypes was generated from a subset of 47 sampling sites corresponding to
118 approximately 130,000 OTUs²⁴. Finally, 5,476 viral “populations” were identified at 43 sites
119 from viral metagenomic contigs, only 39 (<0.1%) of which had been previously observed²⁵
120 (see Methods). These genomics data combined across all domains of life together with
121 carbon flux estimates and other environmental parameters were used to explore the
122 relationships between marine biogeochemistry and euphotic plankton communities (see

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

125 Using a method for regression-based modeling of high dimensional data in biology
126 (specifically a sparse Partial Least Square analysis - sPLS²⁹, Extended data Fig. 1), we
127 detected several plankton lineages for which relative sequence abundance correlated with
128 carbon export and other environmental parameters, most notably with NPP, as expected (Fig.
129 2 and see Supplementary Information SII). These included diatoms, dinoflagellates and
130 metazoa (zooplankton), lineages classically identified as key contributors to carbon export.

131 **Plankton community networks associated with carbon export**

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

143 In brief, the WGCNA approach builds a network in which nodes are features (in this case
144 plankton lineages or gene functions) and links are evaluated by the robustness of co-
145 occurrence scores. WGCNA then clusters the network into modules (hereafter denoted
146 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 lineages²³⁻²⁵ 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).

172 These results corroborate recent metagenomics analysis of microbial communities from
173 sediment traps in the oligotrophic North Pacific subtropical gyre³⁴. Consistently, *in situ*
174 imaging surveys have revealed Rhizarian lineages, made up of large fragile organisms such
175 as the Collodaria, to represent an until now under-appreciated component of global plankton
176 biomass³⁵, which here also appear to be of relevance for carbon export. Another 14% of
177 lineages from the subnetwork correspond to parasitic organisms, a largely under-explored
178 component of planktonic ecosystems.

179 The prokaryotic subnetwork that associated most significantly with carbon export
180 (subnetwork-trait relationship to carbon export, Pearson cor. = 0.32, $p = 9e^{-03}$) contained 109
181 OTUs (Extended data Fig. 2b and Supplementary Information SI3), its structure correlated
182 well to carbon export (Pearson cor. = 0.47, $p = 5e^{-06}$, Extended data Fig. 2e) and it could
183 predict as much as 60% of the carbon export (LOOCV, $R^2 = 0.60$) (Extended data Fig. 3b).
184 By far the highest VIP score within this community was assigned to *Synechococcus*,
185 followed by *Cobetia*, *Pseudoalteromonas* and *Idiomarina*, as well as *Vibrio* and *Arcobacter*
186 (Fig. 3b and Supplementary Information SI3). Noteworthy, *Prochlorococcus* genera and
187 SAR11 clade fall out of this community, while the significance of *Synechococcus* for carbon
188 export could be validated using absolute cell counts estimated by flow cytometry (Pearson
189 cor. = 0.64, $p = 4e^{-10}$, Extended data Fig. 4b). Moreover, *Prochlorococcus* cell counts did not
190 correlate with carbon export (Pearson cor. = -0.13, $p = 0.27$, Extended data Fig. 4a) whereas
191 the *Synechococcus* to *Prochlorococcus* cell count ratio correlated positively and significantly
192 (Pearson cor. = 0.54, $p = 4e^{-07}$, Extended data Fig. 4c), suggesting the relevance of
193 *Synechococcus*, rather than *Prochlorococcus*, to carbon export. Interestingly,
194 *Pseudoalteromonas*, *Idiomarina*, *Vibrio* and *Arcobacter* (of which several species are known
195 to be associated with eukaryotes³⁶) have also been observed in live and poisoned sediment
196 traps³⁴ and these genera display very high VIP scores in our subnetwork associated with

197 carbon export. Additional genera reported as being enriched in poisoned traps (also known
198 as being associated with eukaryotes) include *Enterovibrio* and *Campylobacter*, and are
199 present as well in our carbon export subnetwork.

200 Interestingly, the viral subnetwork ($n=277$) most related to carbon export (Pearson cor. =
201 0.93, $p = 2e^{-15}$, Extended data Fig. 2c) contained particularly high VIP scores for two
202 *Synechococcus* phages (Fig. 3c and Supplementary Information SI4), which represented a
203 16-fold enrichment (Fisher's exact test $p = 6.4e^{-09}$). Its structure also correlated with carbon
204 export (Pearson cor. = 0.88, $p = 6e^{-93}$, Extended data Fig. 2f) and it could predict up to 89%
205 of the variability of carbon export (LOOCV, $R^2 = 0.89$) (Extended data Fig. 3c). The
206 significance of these convergent results is reinforced by the fact that sequences from these
207 datasets are derived from organisms collected on independent size filters (see Methods), and
208 further implicates the importance of top-down processes in carbon export.

209 With the aim of integrating eukaryotic, prokaryotic, and viral carbon export communities, we
210 synthesized their respective subnetworks using, as a backbone, a single global co-occurrence
211 network established previously¹⁰. The resulting network focused on key lineages and their
212 predicted co-occurrences (Fig. 4). Lineages with high VIP values (such as *Synechococcus*)
213 are revealed here as hubs of the co-occurrence network¹⁰, illustrating the potentially strategic
214 key roles within the integrated network of lineages under-appreciated by conventional
215 methods to study carbon export in the ocean. Associations between the hub lineages are
216 mostly mutually exclusive which may explain the relatively weak correlation of some of
217 these lineages with carbon export when using standard correlation analyses as shown in Fig.
218 2.

219 **Gene functions associated with carbon export**

220 Given the potential importance of prokaryotic processes influencing the biological carbon

221 pump²², we used the same analytical approaches to examine the prokaryotic genomic
222 functions associated with carbon export in the annotated Ocean Microbial Reference Gene
223 Catalogue from *Tara Oceans*²³. We built a global co-occurrence network for functions (i.e.,
224 Orthologous Groups of genes or OGs) from the euphotic zone and identified two
225 subnetworks of functions that are significantly associated with carbon export (Fig. 5a,
226 Extended data Fig. 5a, light and dark green subnetworks; FNET1 and FNET2, respectively,
227 and Extended data Fig. 5c).

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

242 The FNET2 subnetwork contains several functions encoded by genes taxonomically
243 assigned to *Candidatus pelagibacter* and *Prochlorococcus*, known as occupying similar
244 oceanic regions as *Synechococcus*, but overall most of its relative abundance (74%) is

245 taxonomically unclassified (Extended data Fig. 6). Top VIP scoring functions in FNET2 are
246 also membrane proteins and ABC-type sugar transporters, as well as functions involved in
247 carbohydrate breakdown such as a chitinase (Fig. 5a). These features highlight the potential
248 roles of bacteria in the formation and degradation of marine aggregates³⁷. Strikingly, 77%
249 and 58%, of OGs with a VIP score > 1 in FNET1 and FNET2, respectively, are functionally
250 uncharacterized^{38,39} (Fig. 5b), pointing to the strong need for future molecular work to
251 explore these functions (see Supplementary Information SI5, SI6).

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

263 **Discussion**

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

269 Furthermore, because of instrument and method limitations, particles smaller than 250 μm
270 were not used for these estimations (see Methods). These export estimates evaluate how
271 much carbon leaves the euphotic zone, but they are not necessarily related to sequestration,
272 which occurs deeper in the water column and over longer timescales. Overall, the use of the
273 UVP was the only realistic method to evaluate carbon flux over the 3 years expedition
274 because deployment of sediment traps at all stations would have been impossible. While our
275 findings are consistent with the numerous previous studies that have highlighted the central
276 role of copepods and diatoms in the biological carbon pump^{14,15,17-19}, they place them in an
277 ecosystem context and generate hypotheses as to the processes that determine the intensity of
278 export, such as parasitism and predation. For example, while viruses are commonly assumed
279 to lyse cells and maintain fixed organic carbon in surface waters, thereby reducing the
280 intensity of the biological carbon pump⁴⁰, there are hints that viral lysis may increase carbon
281 export through the production of colloidal particles and aggregate formation⁴¹. Our current
282 study suggests that these latter roles may be more ubiquitous than currently appreciated. The
283 importance of aggregation and cell stickiness as inferred from gene network analysis, should
284 be further explored mechanistically to investigate the biological significance of these
285 findings.

286 The future evolution of the oceanic carbon sink remains uncertain because of poorly
287 constrained processes, particularly those associated with the biological pump. With current
288 trends in climate change, the size and biodiversity of phytoplankton are predicted to decrease
289 globally^{42,43}. Furthermore, in spite of the potential importance of viruses revealed in this
290 study, they have largely been ignored because of limitations in sampling technologies.
291 Consequently, as oligotrophic gyres expand and global mean NPP decreases⁴⁴, the field is
292 currently unable to predict the consequences for carbon export from the ocean's euphotic
293 zone. By pinpointing key species that appear to be strongly associated with carbon export in

294 these areas, as well as their co-occurrences within plankton communities and key microbial
295 functions, the integrated datasets combined with advanced computational techniques used in
296 this study could provide a framework to address this critical bottleneck.

297 One of the grand challenges in the life sciences is to link genes to ecosystems⁴⁵, based on the
298 posit that genes can have predictable ecological footprints at community and ecosystem
299 levels⁴⁶⁻⁴⁸. The extensive data sets from *Tara* Oceans have allowed us to predict as much as
300 89% of the variability in carbon export from the oligotrophic surface ocean with just a small
301 number of genes, largely with unknown functions, encoded by prokaryotes and viruses.
302 These findings can be used as a basis to include biological complexity and guide
303 experimental work designed to inform modeling of the global carbon cycle and to understand
304 how it influences and is influenced by changes in climate. Such statistical analyses scaling
305 from gene-to-ecosystems may open the way to the development of a new conceptual and
306 methodological framework to better understand the mechanisms underpinning key ecological
307 processes.

308 **References and Notes**

- 309 1 Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere:
310 Integrating terrestrial and oceanic components. *Science* **281**, 237-240,
311 doi:10.1126/Science.281.5374.237 (1998).
- 312 2 Boyd, P. W. & Newton, P. Evidence of the potential Influence of planktonic community structure on
313 the interannual variability of particulate organic-carbon flux. *Deep-Sea Res. I.* **42**, 619-639 (1995).
- 314 3 Guidi, L. *et al.* Effects of phytoplankton community on production, size, and export of large
315 aggregates: A world-ocean analysis. *Limnol. Oceanogr.* **54**, 1951-1963 (2009).
- 316 4 Kwon, E. Y., Primeau, F. & Sarmiento, J. L. The impact of remineralization depth on the air-sea
317 carbon balance. *Nat Geosci* **2**, 630-635 (2009).
- 318 5 IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the*
319 *Fifth Assessment Report of the Intergovernmental Panel on Climate Change.* (Cambridge University
320 Press, 2013).
- 321 6 Kitano, H. Biological robustness. *Nat Rev Genet* **5**, 826-837, doi:10.1038/Nrg1471 (2004).
- 322 7 Suweis, S., Simini, F., Banavar, J. R. & Maritan, A. Emergence of structural and dynamical properties
323 of ecological mutualistic networks. *Nature* **500**, 449-452, doi:10.1038/Nature12438 (2013).
- 324 8 Chow, C. E. T., Kim, D. Y., Sachdeva, R., Caron, D. A. & Fuhrman, J. A. Top-down controls on
325 bacterial community structure: microbial network analysis of bacteria, T4-like viruses and protists.
326 *ISME J.* **8**, 816-829, doi:10.1038/Ismej.2013.199 (2014).
- 327 9 Fuhrman, J. A. Microbial community structure and its functional implications. *Nature* **459**, 193-199,
328 doi:10.1038/Nature08058 (2009).
- 329 10 Lima-Mendez, G. *et al.* Determinants of community structure in the global plankton interactome.
330 *Science* **348**, doi:10.1126/science.1262073 (2015).
- 331 11 Giering, S. L. C. *et al.* Reconciliation of the carbon budget in the ocean's twilight zone. *Nature* **507**,
332 480-483 (2014).
- 333 12 Azam, F. Microbial control of oceanic carbon flux: The plot thickens. *Science* **280**, 694-696 (1998).
- 334 13 Agusti, S. *et al.* Ubiquitous healthy diatoms in the deep sea confirm deep carbon injection by the
335 biological pump. *Nat Commun* **6**, doi:10.1038/Ncomms8608 (2015).
- 336 14 Sancetta, C., Villareal, T. & Falkowski, P. Massive Fluxes of Rhizosolenid Diatoms - a Common
337 Occurrence. *Limnol. Oceanogr.* **36**, 1452-1457 (1991).
- 338 15 Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic north
339 Pacific gyre at station ALOHA. *Mar. Ecol. Prog. Ser.* **182**, 55-67, doi:10.3354/meps182055 (1999).
- 340 16 Omand, M. M. *et al.* Eddy-driven subduction exports particulate organic carbon from the spring
341 bloom. *Science* **348**, 222-225, doi:10.1126/science.1260062 (2015).
- 342 17 Richardson, T. L. & Jackson, G. A. Small phytoplankton and carbon export from the surface ocean.
343 *Science* **315**, 838-840 (2007).
- 344 18 Steinberg, D. K. *et al.* Bacterial vs. zooplankton control of sinking particle flux in the ocean's twilight
345 zone. *Limnol. Oceanogr.* **53**, 1327-1338 (2008).
- 346 19 Turner, J. T. Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump.
347 *Prog. Oceanogr.* **130**, 205-248, doi:10.1016/j.pocean.2014.08.005 (2015).
- 348 20 Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *Plos Biol.* **9**,
349 doi:10.1371/journal.pbio.1001177 (2011).
- 350 21 Strom, S. L. Microbial ecology of ocean biogeochemistry: A community perspective. *Science* **320**,
351 1043-1045, doi:10.1126/Science.1153527 (2008).
- 352 22 Worden, A. Z. *et al.* Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of
353 microbes. *Science* **347**, 1257594, doi:10.1126/Science.1257594 (2015).
- 354 23 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**,
355 doi:10.1126/science.1261359 (2015).
- 356 24 de Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**,
357 doi:10.1126/science.1261605 (2015).
- 358 25 Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348**,
359 doi:10.1126/science.1261498 (2015).
- 360 26 Bork, P. *et al.* Tara Oceans studies plankton at PLANETARY SCALE. *Science* **348**, 873-873,
361 doi:10.1126/science.aac5605 (2015).
- 362 27 Honjo, S., Manganini, S. J., Krishfield, R. A. & Francois, R. Particulate organic carbon fluxes to the
363 ocean interior and factors controlling the biological pump: A synthesis of global sediment trap
364 programs since 1983. *Prog. Oceanogr.* **76**, 217-285, doi:10.1016/j.pocean.2007.11.003 (2008).

365 28 Henson, S. A., Sanders, R. & Madsen, E. Global patterns in efficiency of particulate organic carbon
366 export and transfer to the deep ocean. *Global. Biogeochem. Cy.* **26**, doi:10.1029/2011GB004099
367 (2012).

368 29 Lê Cao, K. A., Rossouw, D., Robert-Granié, C. & Besse, P. A Sparse PLS for Variable Selection when
369 Integrating Omics Data. *Stat Appl Genet Mol* **7**, doi:10.2202/1544-6115.1390 (2008).

370 30 Chaffron, S., Rehrauer, H., Pernthaler, J. & von Mering, C. A global network of coexisting microbes
371 from environmental and whole-genome sequence data. *Genome Res.* **20**, 947-959,
372 doi:10.1101/Gr.104521.109 (2010).

373 31 Faust, K. & Raes, J. Microbial interactions: from networks to models. *Nat. Rev. Microbiol.* **10**, 538-
374 550, doi:10.1038/Nrmicro2832 (2012).

375 32 Aylward, F. O. *et al.* Microbial community transcriptional networks are conserved in three domains at
376 ocean basin scales. *Proceedings of the National Academy of Sciences*, doi:10.1073/pnas.1502883112
377 (2015).

378 33 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *Bmc*
379 *Bioinformatics* **9** (2008).

380 34 Fontanez, K. M., Eppley, J. M., Samo, T. J., Karl, D. M. & DeLong, E. F. Microbial community
381 structure and function on sinking particles in the North Pacific Subtropical Gyre. *Front Microbiol* **6**,
382 Artn 469, doi:10.3389/Fmicb.2015.00/169 (2015).

383 35 Biard, T. *et al.* *In situ* imaging reveals the biomass of large protists in the global ocean. *Nature*
384 (submitted).

385 36 Thomas, T. *et al.* Analysis of the *Pseudoalteromonas tunicata* Genome Reveals Properties of a
386 Surface-Associated Life Style in the Marine Environment. *PLoS ONE* **3**,
387 doi:10.1371/journal.pone.0003252 (2008).

388 37 Azam, F. & Malfatti, F. Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* **5**, 782-791,
389 doi:10.1038/nrmicro1747 (2007).

390 38 Shi, Y. M., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs
391 in the ocean's water column. *Nature* **459**, 266-U154, doi:10.1038/nature08055 (2009).

392 39 Yooshep, S. *et al.* The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of
393 protein families. *Plos Biol.* **5**, 432-466, doi:10.1371/journal.pbio.0050016 (2007).

394 40 Suttle, C. A. Marine viruses - major players in the global ecosystem. *Nat. Rev. Microbiol.* **5**, 801-812,
395 doi:10.1038/Nrmicro1750 (2007).

396 41 Weinbauer, M. G. Ecology of prokaryotic viruses. *Fems Microbiol Rev* **28**, 127-181,
397 doi:10.1016/j.femsre.2003.08.001 (2004).

398 42 Finkel, Z. V. *et al.* Phytoplankton in a changing world: cell size and elemental stoichiometry. *J.*
399 *Plankton Res.* **32**, 119-137 (2010).

400 43 Sommer, U. & Lewandowska, A. Climate change and the phytoplankton spring bloom: warming and
401 overwintering zooplankton have similar effects on phytoplankton. *Glob. Change Biol.* **17**, 154-162,
402 doi:10.1111/J.1365-2486.2010.02182.X (2011).

403 44 Behrenfeld, M. J. *et al.* Climate-driven trends in contemporary ocean productivity. *Nature* **444**, 752-
404 755 (2006).

405 45 DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's
406 interior. *Science* **311**, 496-503, doi:10.1126/Science.1120250 (2006).

407 46 Gianoulis, T. A. *et al.* Quantifying environmental adaptation of metabolic pathways in metagenomics.
408 *P. Natl. Acad. Sci. USA* **106**, 1374-1379, doi:10.1073/Pnas.0808022106 (2009).

409 47 Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes.
410 *Science* **277**, 1300-1302, doi:10.1126/Science.277.5330.1300 (1997).

411 48 Wymore, A. S. *et al.* Genes to ecosystems: exploring the frontiers of ecology with one of the smallest
412 biological units. *New Phytol* **191**, 19-36, doi:10.1111/J.1469-8137.2011.03730.X (2011).

414 **Figure Legends:**

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

418
419 **Figure 2 | Eukaryotic community associated to carbon export seen using standard**
420 **methods for regression-based modeling of high dimensional data.** Eukaryotic lineages
421 associated to carbon export as revealed by sPLS analysis. Correlations between lineages and
422 environmental parameters are depicted as a clustered heatmap and lineages with a correlation
423 to carbon export higher than 0.2 are highlighted.

424
425 **Figure 3 | Ecological networks reveal key taxa lineages associated with carbon export at**
426 **global scale.** The relative abundances of taxa in selected subnetworks were used to estimate
427 carbon export and to identify key lineages associated with the process. **a**, The selected
428 eukaryotic subnetwork ($n=49$, see Supplementary Information SI2) can predict carbon export
429 with high accuracy (PLS regression, LOOCV, $R^2=0.69$, see Extended data Fig. 3a). Lineages
430 with the highest VIP score (dots size is proportional to the VIP score in the scatter plot) in
431 the PLS are depicted as red dots corresponding to three Rhizaria (Collodaria, *Collozoum*
432 *inerme* and *Sticholonche* sp.), one copepod (*Oithona* sp.), one siphonophore (*Lilyopsis*),
433 three Dinophyceae and one ciliate (*Spirotontonia turbinata*). **b**, The selected prokaryotic
434 subnetwork ($n=109$, see Supplementary Information SI3) can predict carbon export with
435 good accuracy (PLS regression, LOOCV, $R^2=0.60$, see Extended data Fig. 3b). **c**, The
436 selected viral population subnetwork ($n=277$, see Supplementary Information SI4) can
437 predict carbon export with high accuracy (PLS regression, LOOCV, $R^2=0.89$, see Extended
438 data Fig. 3c). Two viral populations with a high VIP score (red dots) are predicted as
439 *Synechococcus* phages (see Supplementary Information SI4).

440 **Figure 4 | Plankton community network built from eukaryotic, prokaryotic and viral**
441 **subnetworks related to carbon export.** Major lineages were selected within the three
442 subnetworks ($\text{VIP} > 1$). Co-occurrences between all lineages of interest were extracted from
443 a previously established global co-occurrence network (see methods). Only lineages
444 discussed within the study are pinpointed. The resulting graph is composed of 329 nodes,
445 467 edges, with a diameter of 7, and average weighted degree of 4.6.

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

456 **Methods**

457 **Environmental data collection**

458 From 2009-2013, environmental data (Supplementary Information SI9) were collected across all
459 major oceanic provinces in the context of the *Tara* Oceans expeditions²⁰. Sampling stations were
460 selected to represent distinct marine ecosystems at a global scale⁴⁹. Note that Southern Ocean stations
461 were not examined herein because they were ranked as outliers due to their exceptional
462 environmental characteristics and biota^{23,24}. Environmental data were obtained from vertical profiles
463 of a sampling package^{50,51}. It consisted of conductivity and temperature sensors, chlorophyll and
464 CDOM fluorometers, light transmissometer (Wetlabs C-star 25cm), a backscatter sensor (WetLabs
465 ECO BB), a nitrate sensor (SATLANTIC ISUS) and a Hydroptic Underwater Vision Profiler (UVP;
466 Hydroptics⁵². Nitrate and fluorescence to chlorophyll concentrations as well as salinity were
467 calibrated from water samples collected with Niskin bottle⁵⁰. Net Primary Production (NPP) data
468 were extracted from 8 day composites of the Vertically Generalized Production Model (VGPM⁵³) at
469 the week of sampling⁵⁴. Carbon fluxes and carbon export, corresponding to the carbon flux at 150 m,
470 were estimated based on particle concentration and size distributions obtained from the UVP⁵¹ and
471 details are presented below.

472 **From particle size distribution to carbon export estimation**

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

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

476 where d is the particle diameter, and exponent k is defined as the slope of the number spectrum when
477 equation (2) is log transformed. This slope is commonly used as a descriptor of the shape of the
478 aggregate size distribution.
479

480
481 The carbon-based particle size approach relies on the assumption that the total carbon flux of
482 particles (F) corresponds to the flux spectrum integrated over all particle sizes:

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

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

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

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

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

487 where $\beta = g(\rho - \rho_0)(18\nu\rho_0)^{-1}$, g is the gravitational acceleration, ρ_0 the fluid density, and ν the
488 kinematic viscosity.
489

490 In addition, mass and settling rates of particles, $m(d)$ and $w(d)$, respectively, are often described as
491 power law functions of their diameter obtained by fitting observed data, $m(d) \cdot w(d) = Ad^B$. The

492 particles carbon flux can then be estimated using an approximation of Eq. 2 over a finite number (x)
 493 of small logarithmic intervals for diameter d spanning from 250 μm to 1.5 mm (particles $<250 \mu\text{m}$
 494 and $>1.5 \text{ mm}$ are not considered, consistent with the method presented by *Guidi et al.*, [2008]⁵⁷) such
 495 as

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

496

497 where $A=12.5\pm3.40$ and $B=3.81 \pm 0.70$ have been estimated using a global dataset that compared
 498 particle fluxes in sediment traps and particle size distributions from the UVP images.

499 **Genomic data collection**

500 For the sake of consistency between all available datasets from the *Tara* Oceans expeditions, we
 501 considered subsets of the data recently published in Science²³⁻²⁵. In brief, one sample corresponds to
 502 data collected at one depth (surface (SRF) or Deep Chlorophyll Maximum (DCM) determined from
 503 the profile of chlorophyll fluorometer) and at one station. To study the eukaryotic community in our
 504 current manuscript, we selected stations at which we had environmental data and carbon export
 505 estimated at 150 m with the UVP and all size fractions. Consequently a subset of 33 stations
 506 (corresponding to 56 samples) has been created compared to the 47 stations analyzed in *de Vargas et*
 507 *al.* [2015]. A similar procedure has been applied to the prokaryotic and viral datasets, reducing the
 508 *Sunagawa et al.* [2015] prokaryotic dataset to a subset of 104 samples from 62 stations and the *Brum*
 509 *et al.* [2015] viral dataset into a subset of 37 samples from 22 stations (See Supplementary
 510 Information SI10). In addition a detailed table is provided summarizing which samples (depth and
 511 station) are available for each domain (Supplementary Information SI11).

512 **Eukaryotic taxa profiling**

513 Photic-zone eukaryotic plankton diversity has been investigated through millions of environmental
 514 Illumina reads. Sequences of the 18S ribosomal RNA gene V9 region were obtained by PCR
 515 amplification and a stringent quality-check pipeline has been applied to remove potential chimera or
 516 rare sequences (details on data cleaning in *de Vargas et al.* [2015]²⁴). For 47 stations, and if possible
 517 at two depths (SRF and DCM), eukaryotic communities were sampled in the *piconano*- (0.8-5 μm),
 518 *micro*- (20-180 μm) and *meso*-plankton (180-2000 μm) fractions (a detailed list of these samples is
 519 given in Supplementary Information SI12). In the framework of the carbon export study, sequences
 520 from all size fractions were pooled in order to get the most accurate and statistically reliable dataset
 521 of the eukaryotic community. The 2.3 million eukaryotic ribotypes were assigned to known
 522 eukaryotic taxonomic entities by global alignment to a curated database²⁴. To get the most accurate
 523 vision of the eukaryotic community, sequences showing less than 97% identity with reference
 524 sequences were excluded. The final eukaryotic relative abundance matrix used in our analyses
 525 included 1,750 lineages (taxonomic assignment has been performed using a last common ancestor
 526 methodology, and had thus been performed down to species level when possible) in 56 samples from
 527 33 stations. Pooled abundance (number of V9 sequences) of each lineage has been normalized by the
 528 total sum of sequences in each sample.

529 **Prokaryotic taxa profiling**

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

533 0.22-3 μm size fraction. Prokaryotic taxonomic profiling was performed using 16S rRNA gene tags
534 directly identified in Illumina-sequenced metagenomes (mi tags) as described in *Logares et al.*,
535 [2014]⁵⁸. 16S mi tags were mapped to cluster centroids of taxonomically annotated 16S reference
536 sequences from the SILVA database⁵⁹ (release 115: SSU Ref NR 99) that had been clustered at 97%
537 sequence identity using USEARCH v6.0.307⁶⁰. 16S mi tag counts were normalized by the total reads
538 count in each sample (further details in *Sunagawa et al.* [2015]²³). The photic-zone prokaryotic
539 relative abundance matrix used in our analyses included 3,253,962 mi tags corresponding to 1,328
540 genera in 104 samples from 62 stations.

541

542 **Prokaryotic functional profiling**

543 For each prokaryotic sample, gene relative abundance profiles were generated by mapping reads to
544 the OM-RGC using the MOCAT pipeline⁶¹. The relative abundance of each reference gene was
545 calculated as gene length-normalized base counts. And functional abundances were calculated as the
546 sum of the relative abundances of these reference genes, annotated to OG functional groups. In our
547 analyses, we used the subset of the OM-RGC that was annotated to Bacteria or Archaea (24.4 M
548 genes). Using a rarefied (to 33 M inserts) gene count table, an OG was considered to be part of the
549 ocean microbial core if at least one insert from each sample was mapped to a gene annotated to that
550 OG. For further details on the prokaryotic profiling please refer to *Sunagawa et al.* [2015]²³. The final
551 prokaryotic functional relative abundance matrix used in our analyses included 37,832 OGs or
552 functions in 104 samples from 62 stations. Genes from functions of FNET1 and FNET2 subnetworks
553 were taxonomically annotated using a modified dual BLAST-based last common ancestor (2bLCA)
554 approach⁶². We used RAPsearch2⁶³ rather than BLAST to efficiently process the large data volume
555 and a database of non-redundant protein sequences from UniProt (version: UniRef_2013_07) and
556 eukaryotic transcriptome data not represented in UniRef (see Supplementary Information SI5, SI6,
557 for full annotations).

558 **Enumeration of prokaryotes by flow cytometry**

559 For prokaryote enumeration by flow cytometry, three aliquots of 1 ml of seawater (pre-filtered by
560 200- μm mesh) were collected from both SRF and DCM. The samples were fixed immediately using
561 cold 25% glutaraldehyde (final concentration 0.125%), left in the dark for 10 min at room
562 temperature, flash-frozen and kept in liquid nitrogen on board and then stored at -80°C on land. Two
563 subsamples were taken to separate counts of heterotrophic prokaryotes (not shown herein) and
564 phototrophic picoplankton. For heterotrophic prokaryote determination, 400 μl of sample was added
565 to a diluted SYTO-13 (Molecular Probes Inc., Eugene, OR, USA) stock (10:1) at 2.5 $\mu\text{mol l}^{-1}$ final
566 concentration, left for about 10 min in the dark to complete the staining and run in the flow
567 cytometer. We used a FACS Calibur (Becton & Dickinson) flow cytometer equipped with a 15 mW
568 Argon-ion laser (488 nm emission). At least 30,000 events were acquired for each subsample (usually
569 100,000 events). Fluorescent beads (1 μm , Fluoresbrite carboxylate microspheres, Polysciences Inc.,
570 Warrington, PA) were added at a known density as internal standards. The bead standard
571 concentration was determined by epifluorescence microscopy. For phototrophic picoplankton, we
572 used the same procedure as for heterotrophic prokaryote, but without addition of SYTO-13. Data
573 analysis was performed with FlowJo software (Tree Star, Inc.).

574 **Profiling of viral populations**

575 In order to associate viruses to carbon export we used viral populations as defined in *Brum et al.*
576 [2015]²⁵ using a set of 43 *Tara* Oceans viromes. Briefly, viral populations were defined as large
577 contigs (>10 predicted genes and >10 kb) identified as most likely originating from bacterial or
578 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 Bowtie2⁶⁴ 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 (sPLS⁶⁵ 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 pls⁶⁹. 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. & Sutcliffe, 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: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'Ajuts 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

769 genomic and statistical analyses. A.L., Y.D., L.G., S.C., Lu.B. and D.E. produced and analysed the networks.
770 E.K., C.B. and G.G. supervised the study. M.S., J.R., E.K., C.B. and G.G. provided constructive comments,
771 revised and edited the manuscript. *Tara* Oceans coordinators provided a creative environment and constructive
772 criticism throughout the study. All authors discussed the results and commented on the manuscript.

773 **Author Information**

774 Data described herein is available at EBI under the project identifiers PRJEB402, PRJEB6610 and PRJEB7988,
775 PANGAEA^{50,51,54}, and a companion website (<http://www.raeslab.org/companion/ocean-carbon-export.html>).
776 The data release policy regarding future public release of *Tara* Oceans data is described in *Pesant et al.*,
777 [2015]⁴⁹. All authors approved the final manuscript. Reprints and permissions information is available at
778 www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests
779 for materials should be addressed to lguidi@obs-vlfr.fr, samuel.chaffron@vib-kuleuven.be,
780 lucie.bittner@upmc.fr, damien.eveillard@univ-nantes.fr, Jeroen.Raes@vib-kuleuven.be, karsenti@embl.de,
781 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 ($s \times s$ OTUs (Operational Taxonomic Units)) and its corresponding environmental matrix ($s \times p$ (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 ($s \times s$ 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, $p = 9e^{-03}$). **c**, A single viral subnetwork ($n=277$, $N=5'476$) is strongly associated to carbon export (Pearson cor. 0.93, $p = 2e^{-15}$). **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, $p = 5e^{-16}$). **e**, The prokaryotic subnetwork structure correlates to carbon export (Pearson cor. = 0.47, $p = 5e^{-06}$). **f**, The viral population subnetwork structure correlates to carbon export (Pearson cor. = 0.88, $p = 6e^{-93}$).

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

836 (Pearson cor. = 0.64, $p = 4.0 \times 10^{-10}$). **c**, *Synechococcus* / *Prochlorococcus* cell counts ratio correlates
837 significantly to carbon export (Pearson cor. = 0.54, $p = 4.0 \times 10^{-07}$).
838

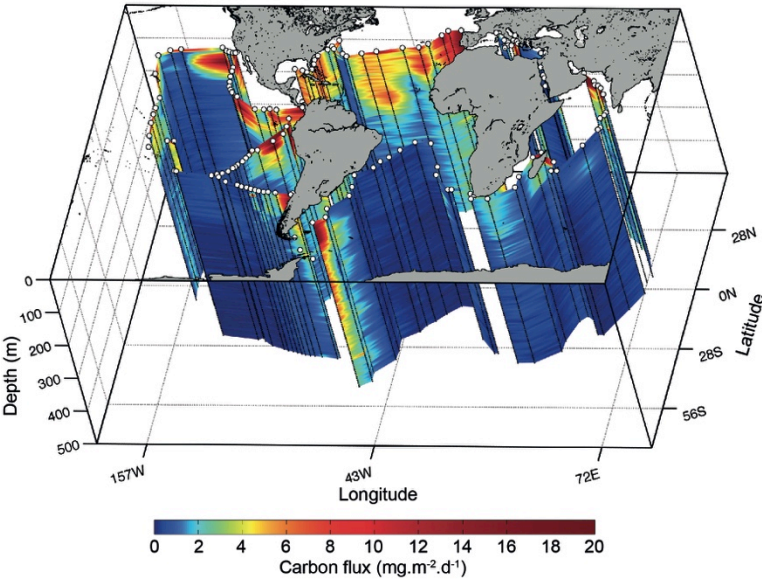
839 **Extended Data Figure 5:** Function and gene subnetworks associated to environmental parameters
840 and their structure correlate to carbon export. **a,b**, Global ecological networks were built for the
841 prokaryotic functions and viral PCs using the WGCNA methodology (see methods) and correlated to
842 classical oceanographic parameters as well as carbon export. Each global network is decomposed into
843 smaller coherent subnetworks (depicted by distinct colours on the y-axis) and their eigen vector is
844 correlated to all environmental parameters. Similar to a correlation at the network scale, this approach
845 directly links subnetworks to environmental parameters (*i.e.* the more the taxa contribute to the
846 subnetwork structure, the more their abundance are correlated to the parameter). The measure allows
847 to identify subnetworks for which the overall structure is related to the carbon export. **a**, Two
848 bacterial functional subnetworks ($n=441$ and $n=220$, $N=37'832$) are associated to carbon export
849 (Pearson cor. 0.54, $p = 1 \times 10^{-07}$ and 0.42, $p = 1 \times 10^{-04}$). **b**, Two viral PCs subnetworks ($n=1'879$ and
850 $n=2'147$, $N=4'678$) are strongly associated to carbon export (Pearson cor. 0.75, $p = 3 \times 10^{-07}$ and 0.91, $p =$
851 3×10^{-14}). **c,d** The WGCNA approach directly links subnetworks to environmental parameters, *i.e.* the
852 more the features contribute to the subnetwork structure (topology), the more their abundance are
853 correlated to the parameter. This measure allows to identify subnetworks for which the overall
854 structure, summarized as the eigen vector of the subnetwork, is related to the carbon export. **c**, The
855 bacterial function subnetwork structures correlates to carbon export (FNET1 Pearson cor. = 0.68, $p =$
856 3×10^{-61} , and FNET2 Pearson cor. = 0.47, $p = 6 \times 10^{-13}$). **d**, The viral PC subnetwork structures correlates to
857 carbon export (VNET1 Pearson cor. = 0.91, $p < 1 \times 10^{-200}$, and VNET2 Pearson cor. = 0.96, $p < 1 \times 10^{-200}$).
858

859 **Extended Data Figure 6:** Cumulative abundance of genus-level taxonomic annotations of genes
860 encoding functions from FNET1 and FNET2 subnetworks and Bacterial function subnetworks
861 predict carbon export. **a**, Genes contributing to the relative abundance of FNET1 and FNET2
862 subnetwork functions were taxonomically annotated by homology searches against a non-redundant
863 gene reference database using a last common ancestor (LCA) approach (see methods). **b,c**, PLS
864 regression was used to predict carbon export using abundances of functions (OGs) in selected
865 subnetworks. LOOCV was performed and VIP scores computed for each function. **b**, Light green
866 subnetwork (FNET1) functions predict carbon export with a R^2 of 0.41. **c**, Dark green subnetwork
867 (FNET2) functions predict carbon export with a R^2 of 0.48.
868

869 **Extended Data Figure 7:** Viral protein cluster networks reveal potential marker genes for carbon
870 export prediction at global scale. **a**, A viral protein cluster (PC) network was built using abundances
871 of PCs predicted from viral population contigs associated to carbon export (Fig. 3b) using the
872 WGCNA methodology (see methods) and correlated to classical oceanographic parameters. Two
873 viral PC subnetworks (light and dark orange, VNET1 and VNET2, left and right panel respectively)
874 are strongly associated to carbon export (VNET1: Pearson cor. 0.75, $p = 3 \times 10^{-07}$ and VNET2: 0.91, $p =$
875 3×10^{-14} , Extended data figure 5b). Size of dots is proportional to the VIP score computed for the PLS
876 regression. **b**, Viral PC subnetworks predict carbon export. PLS regression was used to predict
877 carbon export using abundances of viral protein clusters (PCs) in selected subnetworks. LOOCV was
878 performed and VIP scores computed for each PC. Light orange subnetwork (VNET1, left panel) PCs
879 predict carbon export with a R^2 of 0.55. Dark orange subnetwork (VNET2, right panel) PCs predict
880 carbon export with a R^2 of 0.89.
881

882 **Extended Data Figure 8:** WGCNA and PLS regression analyses for the full Eukaryotic dataset. **a**, A
883 single eukaryotic subnetwork ($n=58$, is strongly associated to carbon export (Pearson cor. 0.79, $p =$
884 3×10^{-14}). **b**, The eukaryotic subnetwork structure correlates to carbon export (Pearson cor. = 0.94, $p = 4 \times$
885 10^{-27}). **c**, The eukaryotic subnetwork predicts carbon export with a R^2 of 0.76. **d**, Lineages with the
886 highest VIP score (dots size is proportional to the VIP score in the scatter plot) in the PLS are
887 depicted as red dots corresponding to two rhizarian (Collodaria), one copepod (*Euchaeta*), and three
888 dinophyceae (*Noctiluca scintillans*, *Gonyaulax polygramma* and *Gonyaulax sp.* (clade 4)).

889



890

891 Figure 1

892

893

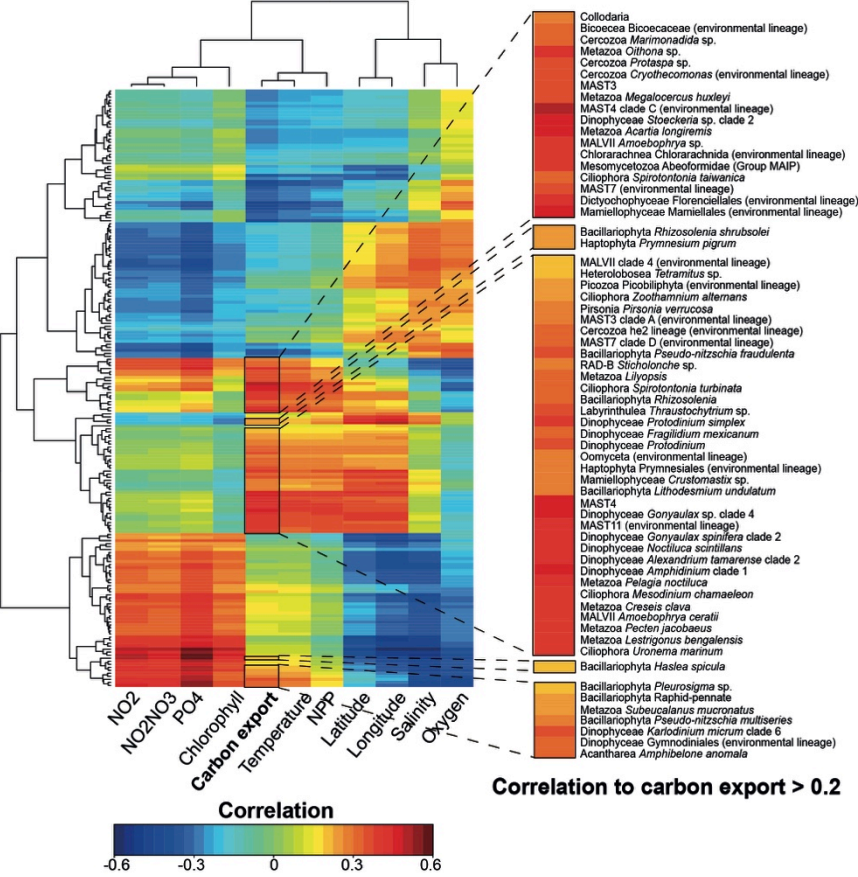
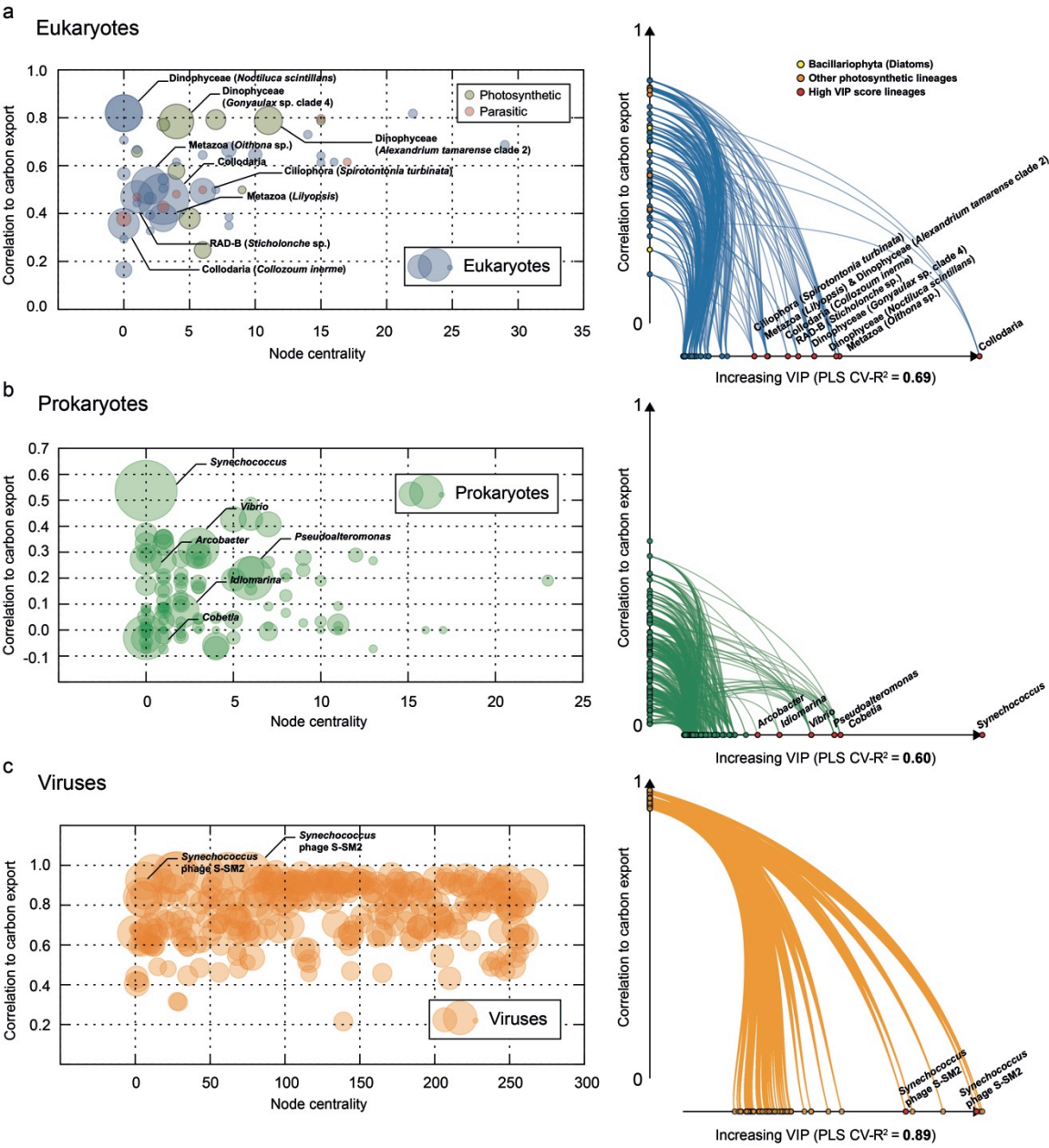


Figure 2

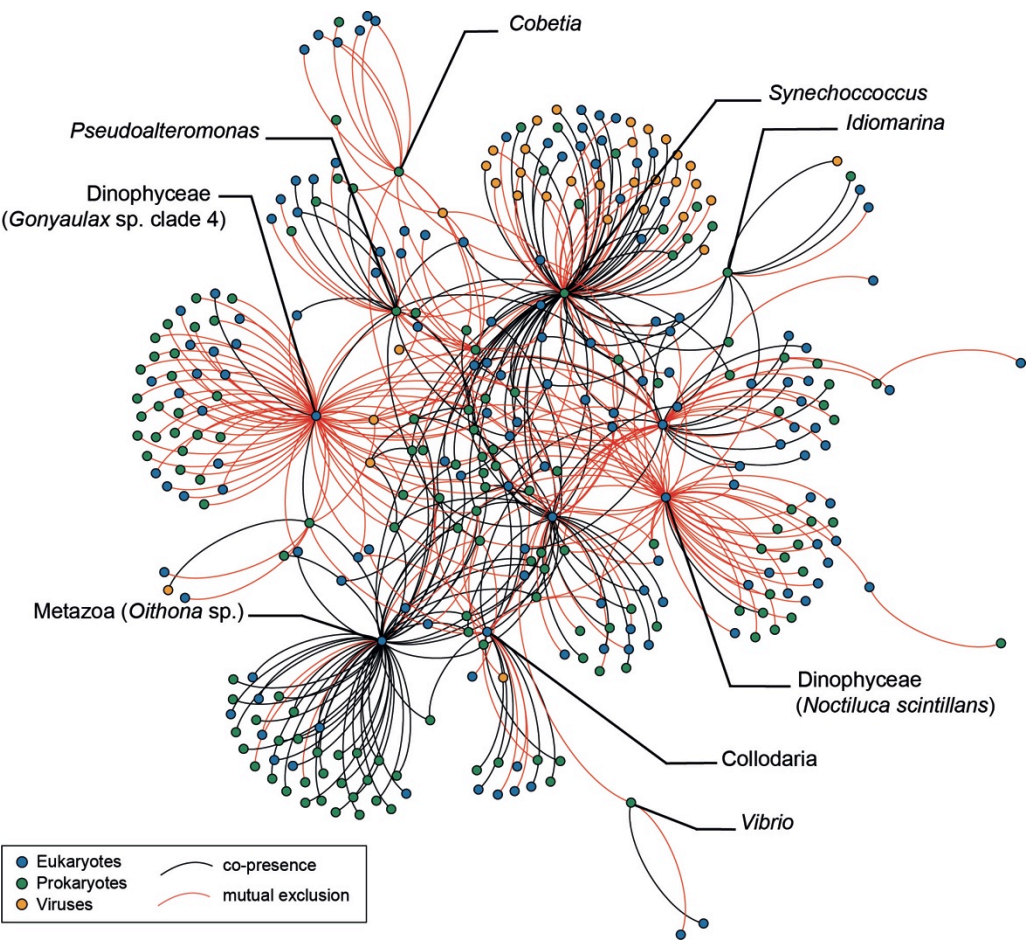
898
899



900
901
902

Figure 3

903
904

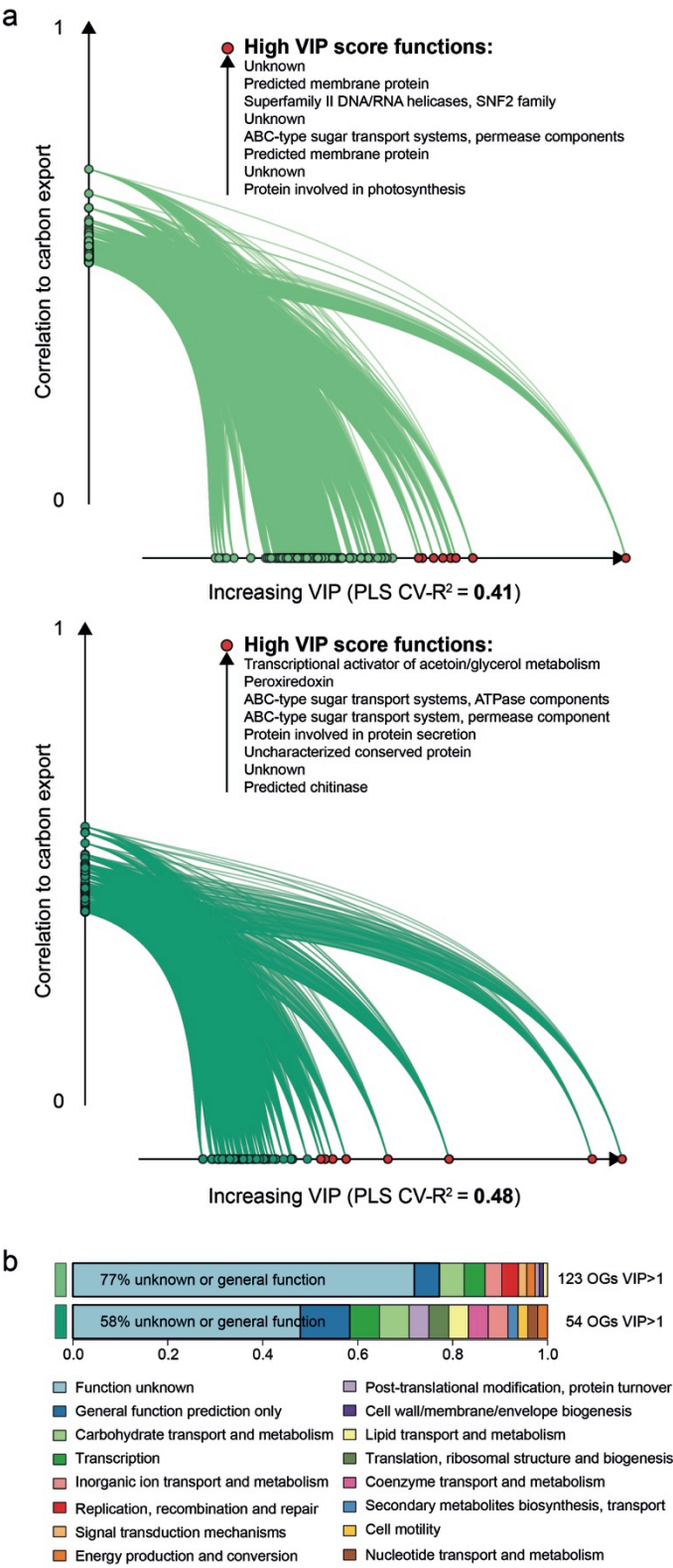


905

906 Figure 4

907

908
909

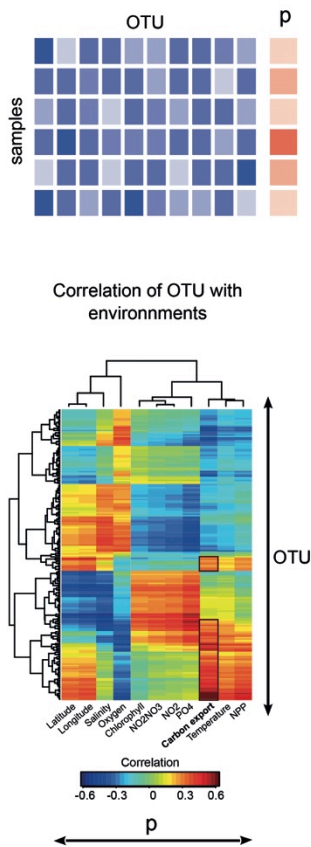


910

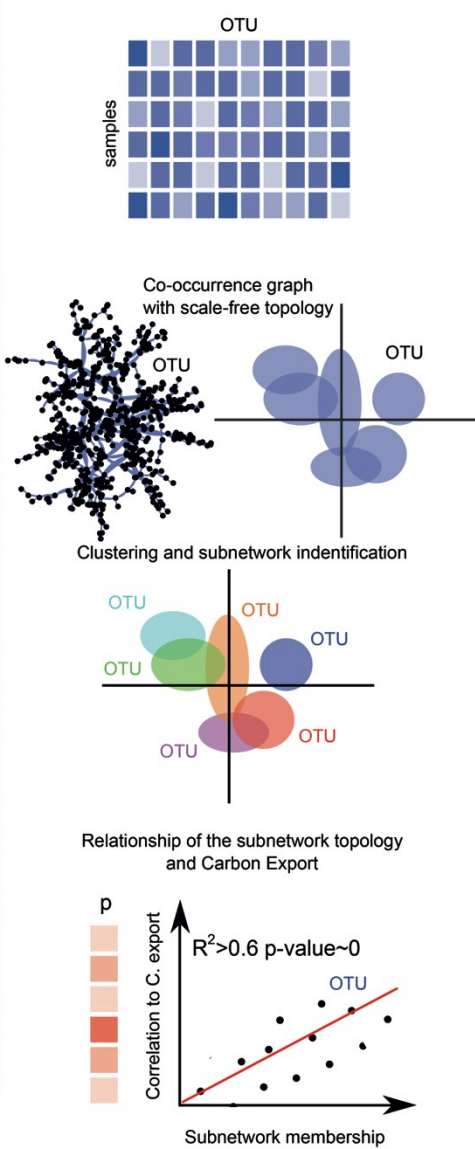
911 Figure 5

912

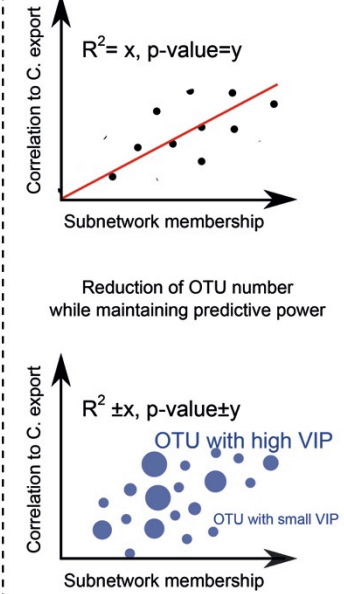
a) Pairwise approach



b) Graph-based approach (WGCNA)



c) Machine learning technique (PLS)



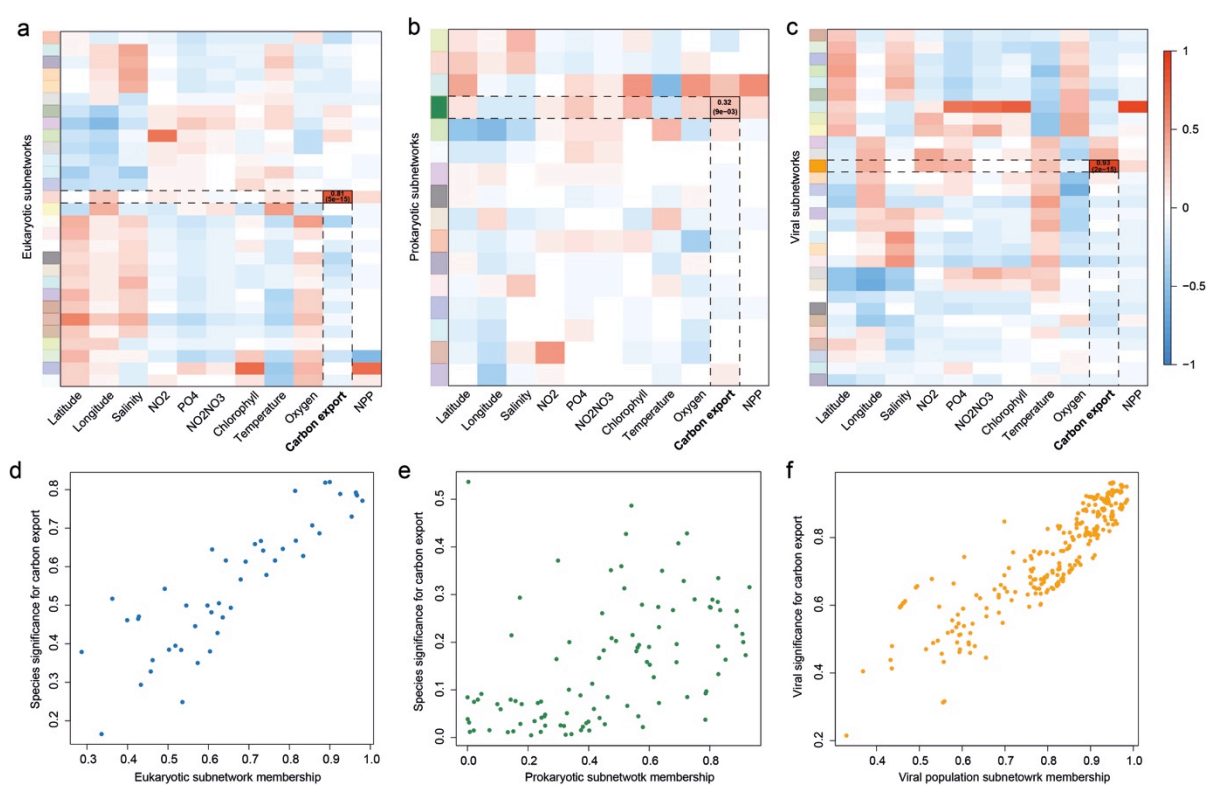
913

914 Extended Data Figure 1

915

916

917

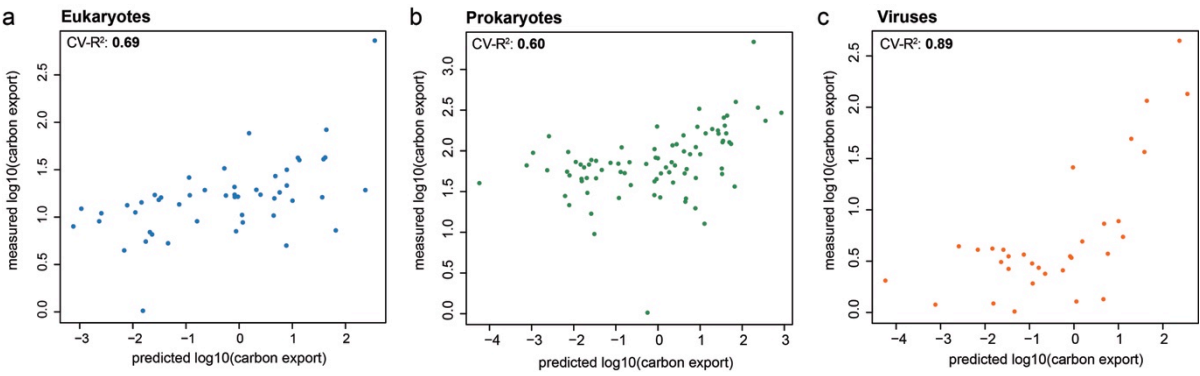


918

919 Extended Data Figure 2

920

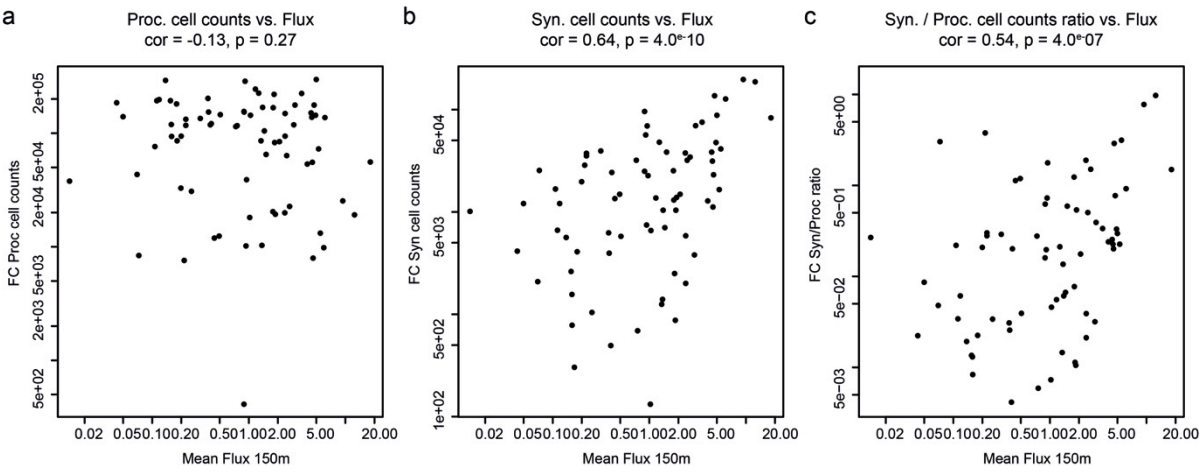
921
922



923
924
925

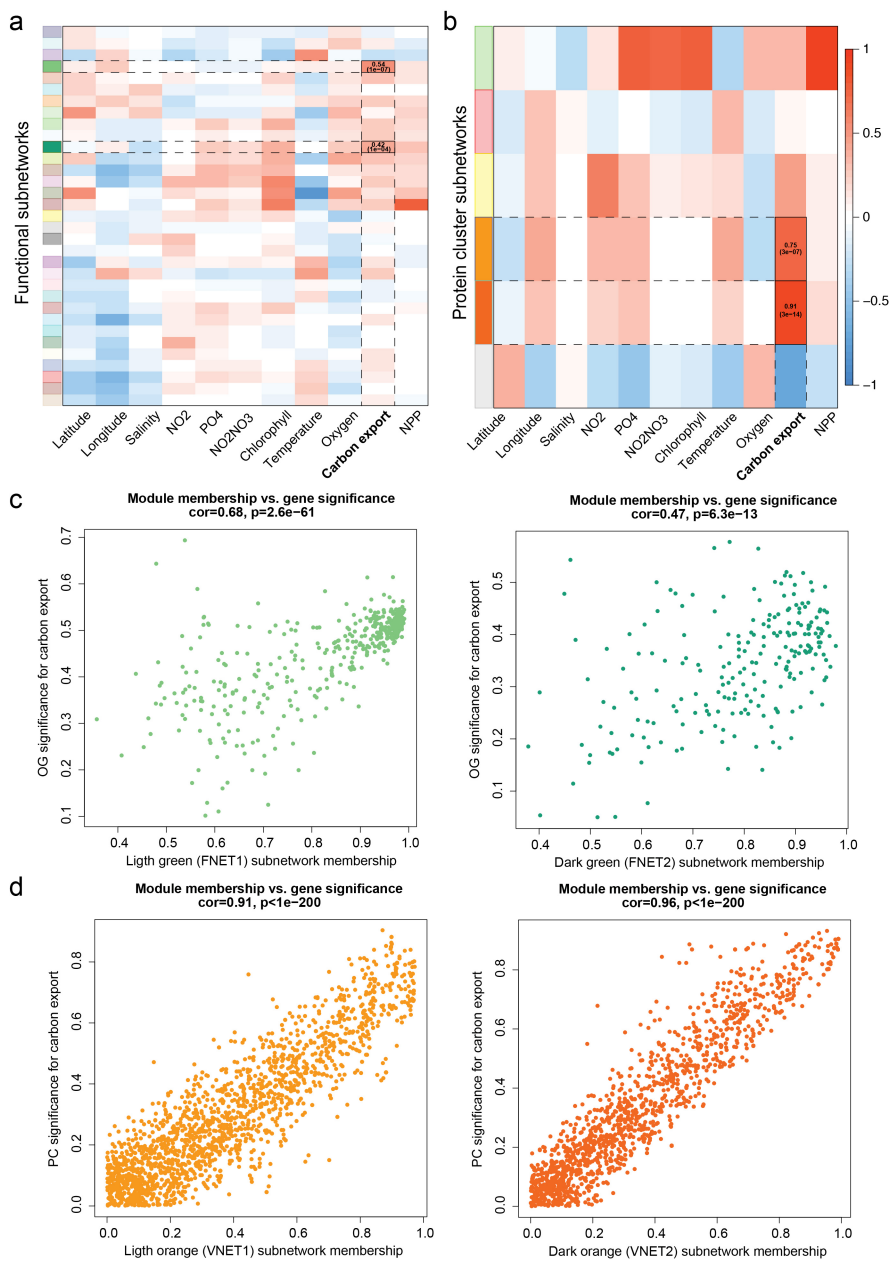
Extended Data Figure 3

926
927
928



929
930
931

Extended Data Figure 4



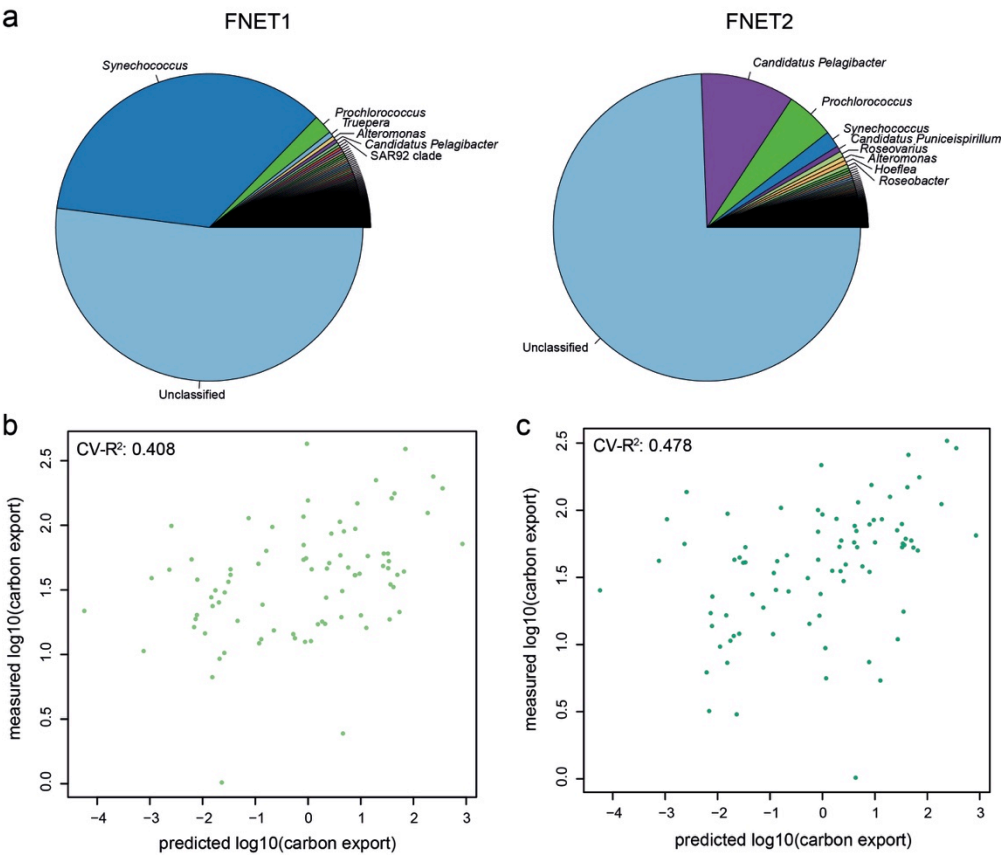
932

933

Extended Data Figure 5

934

935
936

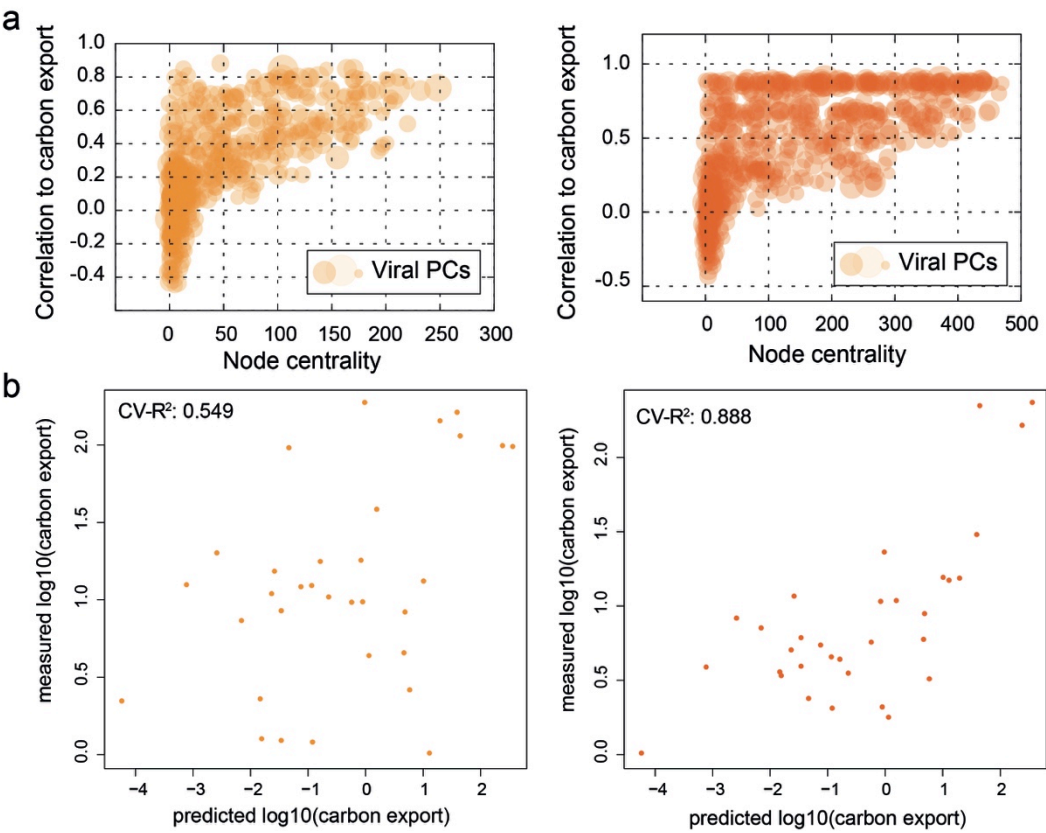


937

938 Extended Data Figure 6

939

940
941

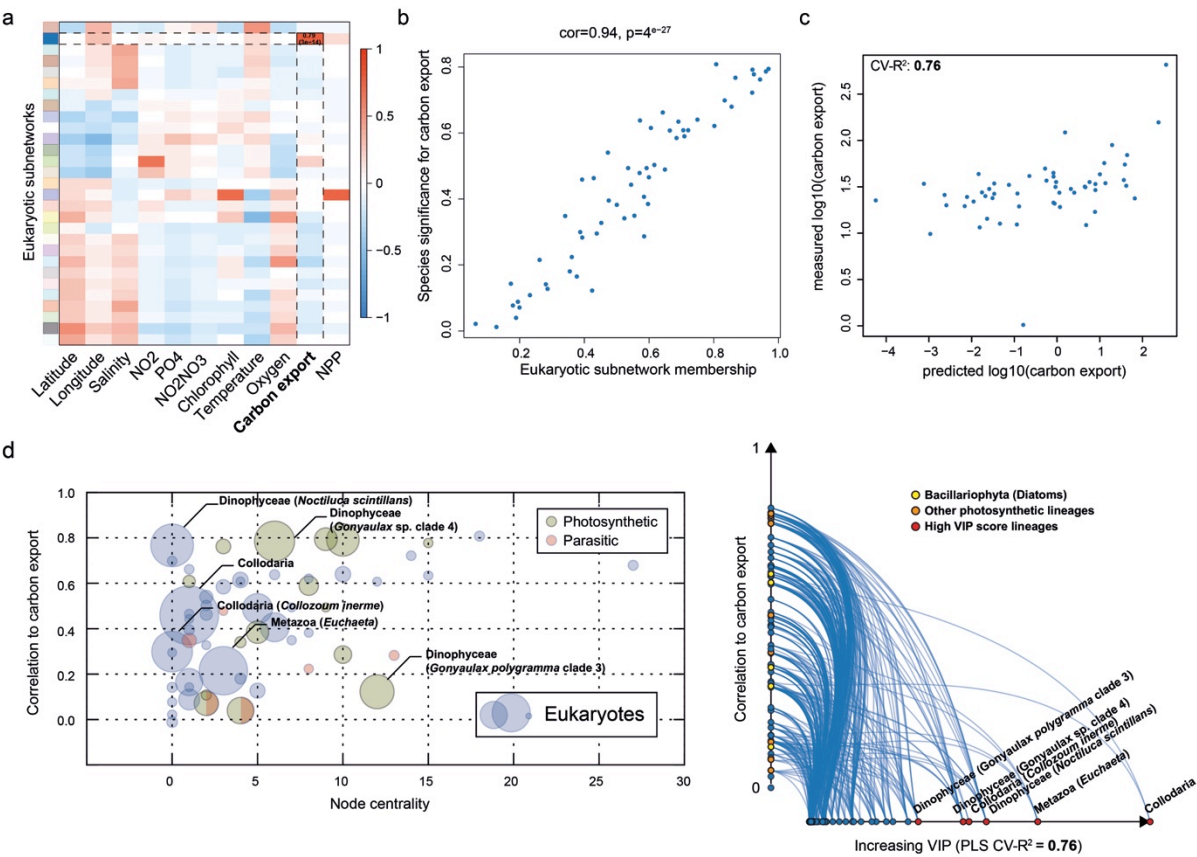


942

943 Extended Data Figure 7

944

945
946



947

948 Extended Data Figure 8