A global ocean atlas of eukaryotic genes

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th>Citation</th>
<th>Carradec, Quentin et al. “A Global Ocean Atlas of Eukaryotic Genes.” Nature Communications 9, 1 (January 2018): 373 © 2018 The Author(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1038/s41467-017-02342-1">http://dx.doi.org/10.1038/s41467-017-02342-1</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Nature Publishing Group</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Thu Apr 11 21:26:52 EDT 2019</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/118335">http://hdl.handle.net/1721.1/118335</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution 4.0 International License</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by/4.0/">http://creativecommons.org/licenses/by/4.0/</a></td>
</tr>
</tbody>
</table>
While our knowledge about the roles of microbes and viruses in the ocean has increased tremendously due to recent advances in genomics and metagenomics, research on marine microbial eukaryotes and zooplankton has benefited much less from these new technologies because of their larger genomes, their enormous diversity, and largely unexplored physiologies. Here, we use a metatranscriptomics approach to capture expressed genes in open ocean Tara Oceans stations across four organismal size fractions. The individual sequence reads cluster into 116 million unigenes representing the largest reference collection of eukaryotic transcripts from any single biome. The catalog is used to unveil functions expressed by eukaryotic marine plankton, and to assess their functional biogeography. Almost half of the sequences have no similarity with known proteins, and a great number belong to new gene families with a restricted distribution in the ocean. Overall, the resource provides the foundations for exploring the roles of marine eukaryotes in ocean ecology and biogeochemistry.
S

ingle-celled microeukaryotes and small multicellular zoo-

plankton account for most of the planktonic biomass in the

world’s ocean\(^1,2\). They are involved in various processes that
shape the biogeochemical cycles of the planet, from primary
production, recycling of organic matter by predation and para-
sitism, sequestration of carbon to a depth, and the transfer of
organic material to higher trophic levels in the food webs\(^3\). Yet,
their analysis is confounded because they are represented by
hundreds of thousands of different taxa belonging to almost all
phylogenetic groups of eukaryotes\(^4\), and the vast majority of them
cannot be cultured. Their highly variable genome sizes, spanning
at least four orders of magnitude\(^5\), and the predominance of
noncoding sequences are additional challenges that have impeded
their genomic exploration. Consequently, their study has been
limited principally to morphological description of diversity, as
well as taxonomic and biogeographic characterizations using
individual barcode genes\(^6,7\). By contrast, global surveys of the
functional potential of marine microbiota (≤5 μm) and double-
stranded DNA viruses are advancing rapidly because of the
availability of comprehensive gene catalogs\(^8\)–12, as has been
performed for the human gut\(^13\) To help assess gene function in
marine eukaryotes, transcriptome data sets from hundreds of
cultured marine eukaryotes\(^14\) have been generated, as well as
from some species of zooplankton\(^15\), which is helping to analyze
features of the global eukaryotic proteome and to interpret the
transcriptional responses of some components of eukaryotic
communities to localized stimuli\(^16,17\).

Herein, we use a metatranscriptomics approach using samples
collected from the global ocean during the Tara Oceans expedi-
tion\(^18\) to generate a global ocean reference catalog of genes from
planktonic microeukaryotes and to explore their expression patterns
with respect to biogeography and environmental conditions.

Results

The Tara Oceans catalog of expressed eukaryotic genes. To
identify and characterize the transcriptionally active genes from
the most abundant eukaryotic plankton in the global ocean, we
selected samples collected during the Tara Oceans expedition at
two main depths in the euphotic zone (subsurface (SRF) and deep
chlorophyll maximum (DCM)), at 68 different geographic loca-
tions across all the major oceanic provinces except the Arctic\(^19\)
(Fig. 1a). Four main organismal size fractions were sampled
independently\(^20\) to optimize the recovery of comprehensive
metatranscriptomes from piconanoplanktonic, nanoplancktonic,
microplanktonic, and mesoplanktonic communities, covering
protists to zooplankton and fish larvae. High-coverage polyA-
based (to avoid ribosomal, organellar, and bacterial RNA) RNA-
Seq was performed on a total of 441 size-fractioned plankton
communities (Fig. 1a), resulting in 16.5 terabases of raw data
from which residual ribosomal RNA sequences were removed.
The cDNA reads were individually assembled for each sample
and then clustered together at 95% sequence identity to create a
single, largely nonredundant resource of 116.8 million transcribed
sequences of at least 150 bases in length, hereafter termed unigenes,
with a N50 length of 635 bases. Rarefaction analysis revealed
that, despite its magnitude, the sampling effort did not
result in near saturation of the eukaryotic gene space, contrasting
with the results obtained from the smallest prokaryote-enriched
size fractions, analyzed by metagenomics from 243 Tara Oceans
samples\(^9\) (Fig. 1b). We estimate that the unigene curve would
reach saturation at 166–190 million sequences, if all ocean regions
would be taxonomically homogeneous (Supplementary Data 1).

Annotation of the >116 million unigenes (Methods and
Supplementary Fig. 1a) revealed that we could assign a taxonomy
level (from “cellular organism” to species name) to only 48.3% of
the unigenes (Fig. 2a and Supplementary Fig. 1b). By mapping
the unigenes onto known gene annotations from marine genomes, we
found a mean value of 2.20 (s.d. = 0.47) unigenes per gene
(Methods and Supplementary Data 2). We then estimated the
number of distinct transcriptomes (originating from different
species) that were present in the catalog by counting the mean
number of copies of conserved ribosomal protein genes, which
indicated that the catalog contains genes from 8823 (s.d. = 1532)
different organisms (Supplementary Data 3). These values
indicate that the unigenes are derived from around 53 (44–68)
million genes, with a mean of 6014 (4226–9223) genes per
sampled organism (Supplementary Data 4). All sequencing reads
from the 441 samples, as well as the reads from a parallel
metagenomics sequencing program, were mapped onto the
unigenes to provide relative expression and abundance for each
gene in every sample (Methods and Supplementary Fig. 1a).

With an equivalent sequencing effort, the complexity of the
metatranscriptomes decreased from the smallest piconano-
planktonic communities to the largest, mesoplanktonic assem-
bles (Fig. 1c), matching the pattern observed in extensive rDNA
metabarcoding data sets\(^9\). Rarefaction curves calculated individ-
ually per size fraction revealed the higher complexity of the
piconano and nanoplanktonic communities (Fig. 1b), and we found
that the 5–20 μm size fraction was the most gene rich, due to
intersample dissimilarity and the presence of more gene-rich
transcriptomes (Fig. 1b, c). All size fractions contained a
significant number of genes not found in the others (8.7–29%);
Supplementary Fig. 1c), indicating the importance of size
fractionation to describe the global eukaryote gene content of
the ocean. With the limitation that we are considering the most
expressed genes in our samples rather than the total gene content,
we observed that a breakdown of the rarefaction curve by oceanic
provinces shows consistent richness and undersaturation of the
gene space, with the notable exception of the Southern Ocean,
and to a lesser extent of the Mediterranean Sea (Fig. 1b). A high-
taxonomic level breakdown of the assignable unigenes across
Tara Oceans stations and organismal size fractions shows a
higher relative abundance of genes from photosynthetic protists
in the piconano plankton, and their progressive replacement by
metazoan transcripts in larger size fractions (Fig. 2b), confirming
the efficiency of the fractionation-based approach. We observed
1.13% of unigenes that are affiliated to prokaryotes. These were
not removed from the catalog, as they can be true nonpolyade-
ylated transcripts from this group, or alternatively to the low
level of eukaryotic annotations with respect to prokaryotes in
reference databases, or to horizontal gene transfers.

Our metatranscriptomic data also captured transcripts (or
RNA genomes) of viruses actively infecting their eukaryotic hosts.
Their activities were found to be pervasive across the geographic
and organismal size ranges examined in this study. Of the
taxonomically assignable unigenes, 33,870 (0.06%) were predicted
to be of eukaryotic virus origin, the vast majority of which (86%)
originated from nucleocytoplasmic large dsDNA viruses
(NCLDV\(^5\)) (Fig. 2c) likely due to the large number of genes
encoded in these viruses. Eukaryotic viral unigenes were
expressed (or present in the case of RNA viruses) in all
441 samples at a relative abundance ranging from 0.0006 to
0.4% (0.02% on average). NCLDV transcripts dominated the
piconano-planktonic communities, while RNA virus sequences
became dominant with increasing organism size (Supplementary
Fig. 2).

Factors discriminating the most expressed functional classes.
To investigate the functional structuring within eukaryotic
plankton communities, we defined the main parameters
discriminating the Pfam domain profiles using principal component analysis (PCA). The first two axes of the PCA are shown in Supplementary Fig. 3a. The main parameter explaining variance corresponded to differentiation between small- and large-size fractions (horizontal axis), and the second major component of variance (vertical axis) separated the Southern Ocean (SO) samples from all the others. A few Gene Ontology (GO) terms show consistent patterns across all size fractions, highlighting major functional and taxonomical differences between SO regions and temperate or tropical oceans (Supplementary Fig. 3b), that can be either due to geographic segregation or to specific parameters of SO, e.g., low iron bioavailability. Samples from this region also tend to be more enriched in diatoms than at the other stations (mean 13%, s.d. = 3.8 in austral stations vs. 3%, s.d. = 2.2, in other samples) (Fig. 2b).

When looking at the most enriched gene categories between size classes, we observed small fractions being enriched in light-based energetic processes (photosynthesis and proteorhodopsins), transport of nutrients, carbohydrate metabolism, and flagellar movement, whereas large size fractions were associated with functions related to multicellularity, cell–cell contact, chitin metabolism, and muscular movement (Fig. 3a and Supplementary Fig. 4). This result demonstrates that the metatranscriptomics data capture not only the taxonomic differences observed previously8 but also the functional repertoires in each size fraction. We also observed that the relative expression of photosynthesis genes (seen through chlorophyll-binding proteins) vs. proteorhodopsins (Bac_rhodopsin Pfam domain corresponding to type-I rhodopsins22,23) showed a strong preference for photosynthesis in groups dominated by autotrophs, supporting that rhodopsin is not a major way of using light energy in these groups in natural conditions (Supplementary Fig. 5a). To further investigate the distribution of the expression of the rhodopsins present in the catalog, we isolated all the unigenes with the Bac_rhodopsin motif) were assigned to that a large majority of annotated eukaryotic unigenes (82% of sequences to study their diversity (Methods section). We found that a large majority of annotated eukaryotic unigenes (82% of unigenes with the Bac_rhodopsin motif) were assigned to alveolates (73%), and contain conserved residues for proton-pumping activity, indicating that this group is the main contributor to proteorhodopsin-based light transduction in the open ocean. The three main clusters contain 55,325 unigenes (77%), and correspond to the three main groups observed based on references only24 (Fig. 3b). Cluster 1 contains xanthorhodopsin-like proteins with conserved residues implicated in proton pumping (Fig. 3b, c and Supplementary Fig. 5b). The 26,733 unigenes of this cluster are almost exclusively derived
from stramenopiles, alveolates, and haptophytes. This taxonomic distribution is consistent with the proposed single horizontal transfer from a bacterium to the common ancestor of the SAR group (Stramenopiles, Alveolates, and Rhizaria) and Haptista. The third cluster contains a large number of eukaryote references and most known sensory rhodopsins, but only 5641 unigenes with diverse taxonomies. Moreover, the proton acceptor residue E76, involved in the proton-pumping function, is not conserved, indicating that Cluster 3 proteins are likely to represent principally sensory rhodopsins (Fig. 3b, c and Supplementary Fig. 5c). Surprisingly, Cluster 2 contains only a few eukaryotic references but is the second largest with 22,951 sequences, and displays the consensus sequence consistent with a proton-pumping function (Fig. 3b, c and Supplementary Fig. 5d). Most of these appear to be derived from alveolates, including the syndiniales parasites. This indicates that one of the most important categories of proteorhodopsins in the ocean is currently underestimated, possibly because of the lack of cultivated organisms bearing it, and that it may link phototrophy with parasitism, a currently unexplored topic. Based on the hypothesis of a single lateral gene transfer event, the restricted taxonomic distribution of unigenes in Cluster 2 suggests
a more recent acquisition, which probably occurred before or during the radiation of the alveolate lineage. Interestingly, the consensus spectral tuning residue is different between Cluster 1 and Cluster 2: Cluster 1 protein sequences exhibit a leucine at position 10525, indicating a maximal absorption of green light, whereas Cluster 2 sequences bear a glutamic acid at this position, indicating a peak absorption of blue wavelengths (Fig. 3c).

Gene novelty. The majority (51.2%) of unigenes currently have no matches in public sequence databases, which limits the insights that can be derived from the gene catalog. Some sequences may be derived from non-coding genes or non-coding portions of coding genes, very short open reading frames, parts of genes where only another region is functionally known, or completely new open reading frames. To distinguish between these possibilities and better classify the catalog, we clustered all the unigenes according to a nucleic acid similarity threshold of >70% (Methods; Supplementary Fig. 6a). Despite its size, the gene catalog is not saturated, and accordingly we observed that 59.6% of unknown unigenes (UU) and 39.8% of known unigenes are represented by singletons (Fig. 4a, b). The clusters may thus be considered as being representative of gene family (GF) content of the catalog, with most singletons likely being derived from
**Fig. 4** Eukaryote gene catalog clustering and characterization of novel genes. **a** Global repartition of unigenes based on the gene catalog clustering. Unigenes were considered as singletons if they are in clusters of less than three units. Gene families are novel (nGF), taxonomically assigned (tGF), functionally assigned (fGF), or both (ftGF) (Methods). Numbers above each bar indicate the numbers of unigenes per cluster. **b** Distribution of unknown unigenes in the different categories described in **a**. **c** Ratio of tGFs vs. ftGFs in the main taxonomic groups. The total number of GFs assigned to each taxonomic group is indicated on the right. **d** Distribution of GF occupancy for the three main GF categories. GFs are classified according to their size (x-axis) and the y-axis indicates the number of stations where the GF family is expressed (at least one unigene detected with a coverage of more than 80% of the unigene length). Kolmogorov-Smirnov tests with p < 10^{-5} between occupancy distributions are indicated with red stars. **e** Distribution of mean expression levels of the three different categories of GFs among all samples. GFs are classified according to their size (x-axis). The expression of a GF in a sample was determined by the sum of the expression of its unigenes in RPKM.
smaller GFs that will grow with more sequencing effort. The 6.2 million GFs, encompassing 58.4 million unigenes, were subsequently subdivided into four classes based on taxonomic affiliation and functional annotation (see Methods; Fig. 4a-c): those with both functional and taxonomic assignments (ftGF), those with taxonomy-only assignments (tGF), those with function-only assignments (fGF), and those representing new GF (nGF). The ftGF category was not considered further because it contains too few clusters (1.43%).

We searched for fundamental differences between these three types of GFs by observing in how many stations they were detected (Methods section). Regardless of GF size, nGFs were present in less stations than fGFs, whereas tGFs showed intermediate occupancies (Fig. 4d and Supplementary Fig. 7a). This pattern was not due to higher mean expression levels of ftGFs or tGFs that would render them more detectable than nGFs (Fig. 4e). We conclude that the gene novelty detected corresponds to families that are present in fewer environments, yet are not less expressed than known families. Moreover, nGFs generally represent smaller GFs (6.3 unigenes per cluster) than fGFs (8.9) and ftGFs (11.4), suggesting that nGFs are conserved in a smaller range of species than characterized GFs (Fig. 4a and Supplementary Fig. 6b), or that they are present in less abundant taxonomic groups. It has been previously suggested that newly discovered genes are either biased taxonomically (which restrains their presence in databases), or that they correspond to genes that are necessary only in some conditions, potentially related to the adaptation of organisms to specific environments. We found evidence for both cases, as nGFs are more restricted in occupancy, whereas tGFs are more abundant in less-characterized phyla (Fig. 4c-e).

We further questioned whether the intermediate occupancies observed with tGFs can be due to an intrinsic property or to them being distributed between two types of families, looking either more like fGFs or more like nGFs. The distribution of occupancies in tGFs indeed appears to be bimodal, with a group containing fewer UUs resembling the ftGF distribution, and another group containing a high proportion of UUs resembling the nGF distribution (Supplementary Fig. 7b,c). We conclude that some of the tGFs likely represent widely occurring genes that have no predicted functions, most likely because of their limited taxonomic distribution in the global tree of eukaryotes. The others may represent GFs with characteristics of nGFs that have few members matching with references, generally reflecting efforts to gain information on environmentally-important organisms such as the MMETSP effort.

Although our metatranscriptomics sequencing effort is based on polyadenylated RNA and relatively shallow coverage per individual organism, and thus may not be able to capture non-coding RNAs significantly, we then consider the nGF category, asking if these new families can be coding. For this, we selected the central unigene of each cluster of more than 10 unigenes as a reference of the GF, then we looked for protein homologies with taxonomy-only assignments (tGF), those with function-only assignments (fGF), and those representing new GF (nGF). The fGF category was not considered further because it contains too few clusters (1.43%).

We searched for fundamental differences between these three types of GFs by observing in how many stations they were detected (Methods section). Regardless of GF size, nGFs were present in less stations than fGFs, whereas tGFs showed intermediate occupancies (Fig. 4d and Supplementary Fig. 7a). This pattern was not due to higher mean expression levels of ftGFs or tGFs that would render them more detectable than nGFs (Fig. 4e). We conclude that the gene novelty detected corresponds to families that are present in fewer environments, yet are not less expressed than known families. Moreover, nGFs generally represent smaller GFs (6.3 unigenes per cluster) than fGFs (8.9) and ftGFs (11.4), suggesting that nGFs are conserved in a smaller range of species than characterized GFs (Fig. 4a and Supplementary Fig. 6b), or that they are present in less abundant taxonomic groups. It has been previously suggested that newly discovered genes are either biased taxonomically (which restrains their presence in databases), or that they correspond to genes that are necessary only in some conditions, potentially related to the adaptation of organisms to specific environments. We found evidence for both cases, as nGFs are more restricted in occupancy, whereas tGFs are more abundant in less-characterized phyla (Fig. 4c-e).

The environmental footprint of gene expression in phytoplankton. To highlight how the annotated gene catalog can be useful for studying environmental gene expression, we examined the five principal photosynthetic groups (Fig. 2c), namely diatoms (Bacillariophyta), chlorophytes, dinoflagellates (Dinophyceae), haptophytes, and pelagophytes, for some of their most highly expressed functions and their variations according to two environmental parameters, specifically iron and net primary production (NPP). Obligate autotrophs, such as diatoms and chlorophytes, showed a higher correlation to NPP for genes involved in photosynthesis and carbon fixation than the other groups that also contain mixotrophic representatives. Additionally, we observed an apparent lack of correlation between expression of genes important for photosynthesis and carbon fixation in dinoflagellates in conditions of high NPP (Supplementary Fig. 9). Although this could be explained by low reliance on transcriptional regulation in this group, we observed an increased correlation of expression of genes encoding cell lytic components, such as proteases and lipases. Such changes in ecosystem function may be a consequence of alterations in the dominant dinoflagellates in the community or to switches in trophic strategy in mixotrophic species, and have significant implications for the functioning of marine food chains in different environmental conditions.

Differences in expression patterns of unigenes between two sampling stations can be linked to either (or both) changes in population composition and changes in expressed functions related to the environment. Comparison of metagenomes and metatranscriptomes allows assessment of the expression of genes from the catalog normalized to underlying gene abundances. To highlight this, we examined genes whose expression and/or copy number have been shown to be responsive to nutrient availability, specifically iron, an important yet often limiting nutrient in the ocean.

Phytoplankton are good models to study iron homeostasis as they have significant high demands of this metal due to its requirement for photosynthesis. One low iron response that occurs in the photosynthetic electron transport chain involves the replacement of the iron-sulfur containing electron carrier ferredoxin with flavodoxin, a less efficient protein that does not require iron. In addition to the canonical photosynthetic versions, there are a number of flavodoxins and ferredoxins involved in different metabolisms, or constituting functional domains of complex multidomain redox proteins. To study whether the flavodoxin/ferredoxin switch can be detected using our dataset, we carried out an analysis of the ferredoxin and flavodoxin families using the Pfam domains PF00111 and...
PF00258. These families not only include the photosynthetic versions but also other isoforms and domains, and there is an overlap of redox properties between different members of these two families, being potential isofunctional proteins in many reactions\textsuperscript{29}. Thus, we studied the relative levels of the two families of genes in the five major phytoplankton groups by calculating the ratio of their relative abundances and expression (Fig. 6).

With the exception of diatoms, gene abundances show little variations and only weak correlations with iron concentrations (Fig. 6a; “Metagenome” column and Supplementary Data\textsuperscript{5}). On the other hand, the ratios of relative expression show strong variations, particularly for chlorophytes, haptophytes and pelagophytes (Fig. 6; “Metatranscriptome” column), indicating that these three groups modulate the relative levels of ferredoxin and flavodoxin principally by regulation of mRNA levels. By contrast, diatoms tend to express flavodoxin genes more than ferredoxin genes, although a few mainly coastal stations showed a strong up-regulation of the latter. In this group, the metagenomics data indicate that diatom genomes display far more heterogeneity in ferredoxin/flavodoxin content than the other groups studied, suggesting that individual diatom species may be permanently adapted to specific iron regimes in the ocean rather than maintaining transcriptional flexibility, as observed in haptophytes, chlorophytes and pelagophytes. Unlike any other groups, dinoflagellates appear to rely only weakly on gene abundance or expression variations (Fig. 6), which may again be related to their low transcription flexibility. These results suggest that nutrient limitations are dealt with in different ways among these main photosynthetic taxa, either by a genotypic commitment to a specific regime, or by the maintenance of transcriptional flexibility, and that the Tara Oceans eukaryote gene catalog may be a useful resource to distinguish the strategies of any plankton group to adapt to these limitations when transcript regulation or gene copy number is implicated.

**Discussion**

The global ocean transcript catalog reported here represents a first resource to study extensively and uniformly the gene content of eukaryotes and the dynamics of their expression in the environment, and notably adds to previous DNA-based resources that describe the viral and prokaryotic components of the
The gene repertoire of planktonic eukaryotes is massive and diverse, much more so than the prokaryotic gene space. The impressive number of genes without functionally-characterized homologs in databases points to the large numbers of understudied yet widely distributed genera inhabiting marine ecosystems, for which even widely conserved GFs have yet to be investigated. The restricted distribution of totally new GFs highlights the need to develop methods for revealing their roles without the support of homology-based hypotheses. Because representatives of almost all of the eukaryote groups are abundant in oceanic plankton, they can likely inform us in new ways about the evolutionary trajectories of different eukaryotes, in particular those with parasitic and symbiotic lifestyles that have remained largely recalcitrant to study until now, although being a large part of the interacting species network within plankton ecosystems. The resource is also likely to be of great utility for exploring organisms within the zooplankton, including metazoans, that have to date been largely unexplored by genomics. As
we have shown for the principal groups of phytoplankton, it is well possible to obtain insights between adaptive and acclimatory processes underlying organismal responses to their environment using as proxies the contrasts between metagenomics and metatranscriptomics, paving the way for similar studies in other organisms.

Methods

Sampling of eukaryotic plankton communities. The biological samples were collected during the Tara Oceans expedition from 68 sampling sites. Typically, two depths were sampled in the photic zone: surface (SFE) and sub-surface (SRF) and a depth of chlorophyll maximum (DCM). A detailed description of all Tara Oceans field sampling strategy and protocols is available in Pesant at al.20. In short, planktonic eukaryote communities were collected in the 0.8–200 μm range and size-fractionated in four fractions (0.8–5 μm, 5–20 μm, 20–180 μm, and 180–2000 μm). A low-shear and non-intrusive industrial peristaltic pump was used for the 0.8–5 μm fraction and plankton nets for the others. The volumes of filtered seawater were scaled according to known organismal concentrations within each size fraction, from 0.1 m−2 for the most concentrated pico-plankton to 148 ± 136 m−2 for the most-dilute meso-plankton, in order to get near-exhaustive recovery of total eukaryotic bio-diversity in each sample. Water was filtered immediately after sampling. Whole-plankton communities were subsequently filtered on polycarbonate membranes, rapidly flash-frozen and preserved in liquid nitrogen on board Tara.

Physicochemical parameters measured during the expedition are available in the Pangaea database (https://www.pangaea.de/ and Supplementary Data 5) and described in Pesant at al. Due to the spatial remoteness of iron in the surface ocean, concentrations were derived from a global ocean simulation using the MITgcm ocean model configured with 18 km horizontal resolution and a biogeochemical simulation which resolves the cycles of nitrogen, phosphorus, iron and silicon14. The biogeochemical parameterizations, including iron, are detailed in Follows at al.22. Atmospheric deposition of iron was imposed using monthly simulations using the MITgcm ocean model configuration. The iron amount for a grid cell was calculated as the product of the above simulations using the MITgcm ocean model configuration and the atmospheric deposition (including river, sea and air-sea exchange) and the oceanic iron chemical cycle. The total iron available for the oceanic iron cycle was calculated as the sum of the above simulations, the atmospheric deposition and the oceanic iron chemical cycle.

Nucleic acid extraction, library construction and sequencing. DNA and RNA were extracted simultaneously by cryogenic grinding of cryopreserved membrane filters using a 6770 Freezer/Mill or 6870 Freezer/Mill instrument (SPEX Sample Prep, Metuchen, NJ) followed by nucleic acid extraction with NucleoSpin RNA Midi kits (Macherey-Nagel, Düren, Germany) combined with DNA Elution buffer kit (Macherey-Nagel). DNA and RNA were quantified by a fluorometric method using Qubit 2.0 Fluorometer (ThermoFisher Scientific, Waltham, MA). DNase treatments were applied to all RNA extractions. Metagenomic libraries were prepared manually or in a semi-automatic manner according to available DNA extraction kits. DNA inputs in fragmentation were 30 μg. For DNA samples, a Covaris E210 instrument (Covaris, Woburn, MA) was used for cDNA synthesis, followed by 12 cycles of PCR preamplification. The Illumina library preparation followed the manual protocol described for meta-transcriptomic reads of at least one Tara Oceans station. The gene catalog is accessible at http://www.genoscope.cns.fr/tara/.

Taxonomic assignment. To assign a taxonomic group to each unique gene, a reference database was built from UniRef90 (release of 2014–09–04)42, from the MMETSP project (release of 2014–07–30)34 manually curated to remove sequence redundancy, from Tara Oceans Single-cell Amplified Genomes (PRJEB6603). The database was supplemented with three Rhizaria transcriptomes (Phaeodae and Eucarytid, available through the European Nucleotide Archive under the reference PRJEB21821 (https://www.ebi.ac.uk/ena/data/view/PRJEB21821) and transcriptomes of Oithona nana)43. Sequence similarities between the gene catalog and the reference database were computed in protein space using Diamond (version 0.7.9)43 with the following parameters: -e 1e-5 -k 500 -a 8. Taxonomic affiliation was performed using a weighted Lowest Common Ancestor approach. For each unigene, all protein matches with a bitscore value ≥90% of the best match bitscore were kept. For each taxon, only matches with the highest bitscores were retained, and total LCA and weighted LCA (covering at least 67% of all bitscores), were further computed. In order to limit the number of false taxonomic assignments explained by the lack of reference genomes, the LCA result was corrected according to the percentage of identity of selected matches. The maximal taxonomic precision allowed was corrected as follows: >95% of identity = species, >95% of identity = genus, >80% of identity = family, >65% of identity = order, >50% of identity = class. The taxonomic assignment is accessible at http://www.genoscope.cns.fr/tara/.

The taxonomic assignment of eukaryotic viruses was performed as explained above but with the following modifications. First, subject sequences with viral taxonomic identifiers were removed and replaced by viral sequences of Virus-Host DB34 (as of 23 February 2017) to allow access to host type (species, genus and family) assignments. Viral families were also removed based on the basis of a positive BLAST hit against dedicated reference databases manually curated, and having matches with at least 70% identity over at least 80% of the unigene length or at least 300 bp long, or based on the presence of specific protein domains identified by CDD search. Domains COXI, COX2, COX3, COX2, TM, Cytochrome B_N_2, Cytochrome B_C, Cytochrome B_N, Oxidored_q1, Oxidored_q2, Oxidored_q3, Oxidored_q4, Oxidored_q5, Oxidored_q6, NADHd, NDH1_L_M, NDH1_L_1, and ATP_synth_6_0_A were used as signature for mitochondrial genes, domains and Photo_RC, PsaA_PsaB, PsII, Rubisco_large, and Rubisco_large_N for the chloroplastic ones, while unigenes also bearing domains Peptidase_M41, Gp_dh_N, or Gp_dh_C, GAPDH-1 were kept in the resource, being considered as nuclear genes. In summary a unigene as defined here is a complete or partial transcript assemble from metatranscriptomic reads of at least one Tara Oceans station.

Functional characterization of unigenes. Protein domain prediction was performed using the hmmsearch tool of the the HMMer package (version 3.1b2)52 against the Pfam-A database (release 28). Only matches exceeding the internal gathering threshold (cut off = 0.7) were retained. The list of associated Pfams is given in Supplementary Data 7. Functional characterization of unigenes is accessible at http://www.genoscope.cns.fr/tara/.

Jumping threshold (cut_off = 0.7) were retained. The list of associated Pfams is given in Supplementary Data 7. Functional characterization of unigenes is accessible at http://www.genoscope.cns.fr/tara/.
Expression and abundance of unigenes. In order to estimate the abundance and expression of each unigene in each sample, cleaned reads (from metagenomes and metatranscriptomes) were mapped against the reference catalog using the bwa tool (version 0.7.4)\(^{15}\). The following parameters were used: bwa aln -t 30 -O 11 -R 1; bwa.sampe -a 20000 -n 1; samtools. rmdup. Low complexity reads were removed. Reads covering at least 80% of read length with at least 95% of identity were retained for further analysis. In the case of several possible best matches, a random one was picked. Mapping results are summarized in Supplementary Data 8. Unigene expression values and genomic occurrences were computed in RPMK (kilobase cover per million of the calculated reads). RPMK values for each Unigenes in each sample are accessible at http://www.genoscope.cns.fr/tara/.

The abundance or expression of each unigene was normalized and formulated in two different ways. (i) The gene expression/abundance relative to the expression/abundance of all genes from the same taxonomic group in percentage, e.g., the expression of Pela- gophyceae genes (Pfl) was compared with the expression of all Pelagophyceae transcrip tions. (ii) The fraction of the gene expression/abundance attributed to a particular taxonomic group, e.g. 24.3% of ferredoxin genes are expressed/present in Pelagophyceae transcriptions. These normalized values of expression and abundance are calculated for all unigenes grouped by Pfams or GO term (Biological Processes) and a list of taxonomic groups: Haptophytes, Pelagophyceae, Pela- gophyceae, Bacillariophyta, Dictyochophyceae, O/U Stramenopiles, Chlorophyta, Dinophyceae, Ciliophora, O/ U Alveolata, Rhizaria, Copepoda, O/ U Protostomia, Tunicata, O/ U Deuterostomia, O/ U Metazoa, O/ U Eukarya, Bacteria, root (unigenes with matches in at least two of the Eukarya, Archaea, Bacteria, and Virus superkingdoms), unknown (unigenes that have no similarities in amino acid databases), O/ U = unigenes for which taxonomic affiliation ended at the indicated level or belonged to minor classes of the affiliation.

Estimation of transcriptome diversity. A total of 24 ribosomal genes, single copy, highly expressed and universally distributed\(^{29}\), were selected to estimate the number of different transcriptions in each sample: COG0049, COG0052, COG0057, COG0081, COG0087, COG0088, COG0091-COG0094, COG0096-COG0100, COG0102, COG0103, COG0184-COG0197, COG0200, COG0256, COG0522. The average number of unigenes covering each of these COG domains was used to estimate the number of different transcriptions. A unigene was considered to be present in a sample if at least 80% of its length was covered by sample reads with at least 95% identity. Reference genomes and their annotation used to calculate the diversity of the gene catalog and refine transcriptome diversity estimations were downloaded from Ensembl Protists (http://protists.ensembl.org/index.html) for Emiliana huxleyi, Thalassiosira oceanica, Aureococcus anophagefferens, Acanthamoeba castellanii strain N, Neif and Monosiga brevicollis, from Orcae (http://bioinformatics.psb.ugent.be/orcae/) for Batchyscythius prasinus and Microcystis pusilla and from Genoscope (http://www.genoscope.cns.fr/externe/GenomeBrowser/) for Oikopleura dioica and Oithona nana. The gene catalog was aligned (BLAST v3.2.11) against predicted genes from reference genomes with a minimum of 70% of identity over at least 80% of the length of the smallest sequence of the pair (Supplementary Data 2), then fully overlapping unigenes have been removed and reference genome scores. Orcae database was also used to map each gene and ribosomal proteins listed above were calculated. The mean of the result for each genome was used as an estimation of the catalog redundancy.

Construction of eukaryote gene families. Nucleic acid homologies between all unigenes of the eukaryotic gene catalog were calculated with BLAST (v. 36) (min 70% of identity and 100 bp). The 1609 million matches obtained were clustered with MCL (v. 14–137) into 6,225,695 clusters of 3 unigenes or more, named GFs (Supplementary Fig. 6a, steps 1–2). Clusters were classified into four categories according to their percentage of unigenes with a taxonomic affiliation and/or a functional characterization. Functionally and taxonomically assigned GFs (fGFs) comprise >5% of unigenes with matches and taxonomies; taxonomically assigned GFs (tGFs) comprise >5% of unigenes with matches but no predicted matches; new GFs (nGFs) have <5% of unigenes with matches or domains; and functionally assigned GFs (fGFs) have >5% of unigenes with domains and <5% with matches (Supplementary Fig. 6a, step 3). The most precise taxonomic affiliation carried by more than 50% of known unigenes of a given tGF or fGF was chosen to determine its taxonomic affiliation. A representative unigene for each GF with a minimum of 10 unigenes and assignment of homologous unigenes (Supplementary 3.1b2)52. NCBI sequences carrying the Pfam motif were retrieved through the NCBI sequences were clustered with the MCL algorithm\(^{31}\) using the -log(value) as edge weights and an inflation parameter of 1.4. For each of the three largest clusters, protein sequences were aligned using MAFFT 7.310\(^{14}\) and positions with more than 50% of gaps were discarded. Logo consensus sequences were created using weblogo 3 program\(^{55}\). Global phylogenetic tree was constructed from a global consensus sequences\(^{56}\). Global phylogenetic inference was made using the dominant zooplankter of the North Atlantic Ocean.

References


Acknowledgements

We thank the commitment of the following people and sponsors who made this singular expedition possible: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genocea/CEA, the French Government ‘Investissement d’Avenir’ programs Oceanomics (ANR-11-BTBR-0008), MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-11-IDEX-0001–02), and FRANCE GÉNOMIQUE (ANR-10-INSBS-09), Fund for Scientific Research—Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects PHTYBACT/ANR-2010–1709–01, POSEIDON/ANR-09-BLAN-0348, PROMETHES/ANR-09-PCS-GENM-217, TARA-GERUS/ANR-09-PCS-GENM-218, EU FP7 (MicroBio3/No.287589), ERC Advanced Grant Award (Diatomic 294823), the LouisD institute of the Foundation de France, a Radcliffe Institute Fellowship from Harvard University to CB, JSPS/MEXT KAKENHI (Nos. 26430184, 15H06437, 16H06429, 16K21723, 16K0020), The Canon Foundation (No. 20131400025), agnès b, the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L’Orient, the EDF Foundation EDF Diversiterre, Pirestrome, the Prince Albert II de Monaco Foundation, Etienne Bourges, the Tara schooner, and its captain and crew. Tara Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge C. Scarpelli for support in high-performance computing. Computations were performed using the plaqe, titane and curc HPC machine provided through GENCI grants (20120176389, 20120176389, 2012014036389, 2012015036389, and 2012016636389). This article is contribution number 62 of Tara Oceans.

Author contributions


Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-00742-1. See details in the Competing interests section.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018

Quentin Carradec1,2,3, Eric Pelletier1,2,3, Corinne Da Silva1, Adriana Alberti1, Yoann Seeleuthner1,2,3, Romain Blanc-Mathieu4, Gipsi Lima-Mendez5,6,21,22, Fabio Rocha7, Leila Tirichine7, Karine Labadie1, Amos Kirilovsky1,2,3,7, Alexis Bertrand1, Stefan Engelken1, Mohammed-Amin Madoui1,2,3, Raphaël Méheust7, Julie Poulin1, Sarah Romac8,9, Daniel J. Richter1,8,9, Genki Yoshikawa4, Céline Dimier7,8,9, Stefanie Kandels-Lewis10,11, Marc Picheral12, Sarah Searson13, Tara Oceans Coordinators, Olivier Jaillon4, Jean-Marc Auryn1, Eric Karsenti7,11,12, Matthew B. Sullivan14, Shinichi Sunagawa1,2,3, Corinne Da Silva1, Adriana Alberti1, Yoann Seeleuthner1,2,3, Peer Bork10,16,17,18, Silence G. Acinas23, Emmanuel Boss24, Michael Follows25, Gabriel Gorsky12, Nigel Grimsley26,27, Lee Karp-Boss24, Uros Krzic28, Stephane Pesant29,30, Emmanuel G. Reynaud31, Christian Sardet12,32, Mike Sieracki33, Sabrina Speich34,35, Lars Stemmann12, Didier Velayoudon36 & Jean Weissenbach1,2,3

1CEA - Institut de Biologie François Jacob, Genoscope, Evry, 91057, France. 2CNRS UMR Metabolic Genomics, Evry, 91057, France. 3Univ Evry, Evry, 91057, France. 4Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. 5Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, Leuven, 3000, Belgium. 6VIB Center for Microbiology, Herestraat 49, Leuven, 3000, Belgium. 7Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, Paris, F-75005, France. 8CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, 29680, France. 9Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, Roscoff, 29680, France. 10Structural and Computational Biology, European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg, 69117, Germany. 11Directors' Research European Molecular Biology Laboratory, Meyerhofstr. 1, Heidelberg, 69117, Germany. 12Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire d’Océanographie de Villefranche (LOV) Observatoire Océanologique, Villefranche-sur-Mer, 06230, France. 13Department of Oceanography, University of Hawaii, Honolulu, 96844 Hawaii, USA. 14Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA. 15Department of Biology, Institute of Microbiology, Vladimir-Prelog-Weg 4, Zürich, 8093, Switzerland. 16Molecular Medicine Partnership Unit, University of Heidelberg and European Molecular Biology Laboratory, Heidelberg, 69120, Germany. 17Max Delbrück Centre for Molecular Medicine, Berlin, 13125, Germany. 18Department of Bioinformatics, University of Wuerzburg, Würzburg, 97074, Germany. 19Aix Marseille Univ, Université de Toulon, CNRS, IRD, MIO, Marseille, 13284, France. 20Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, 80121, Italy. 21Present address: Cellular and Molecular Microbiology, Faculté des Sciences, Université Libre, de Bruxelles (ULB), Belgium. 22Present address: Interuniversity Institute for Bioinformatics in Brussels, ULB-VUB, Boulevard du Triomphe CP 263, 1050 Brussels, Belgium. Quentin Carradec and Eric Pelletier contributed equally to this work.

Tara Oceans Coordinators

Silvia G. Acinas23, Emmanuel Boss24, Michael Follows25, Gabriel Gorsky12, Nigel Grimsley26,27, Lee Karp-Boss24, Uros Krzic28, Stephane Pesant29,30, Emmanuel G. Reynaud31, Christian Sardet12,32, Mike Sieracki33, Sabrina Speich34,35, Lars Stemmann12, Didier Velayoudon36 & Jean Weissenbach1,2,3

23Department of Marine Biology and Oceanography, Institut de Ciències del Mar (CSIC), E-08003 Barcelona, Catalonia, Spain. 24School of Marine Sciences, University of Maine, Orono, ME 04469, USA. 25Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, 02138 MA, USA. 26CNRS UMR 7232, BIOM, Avenue du Fontaulé, Banyuls-sur-Mer, 66650, France. 27Sorbonne Universités Paris 06, O08 UPMC, Avenue du Fontaulé, Banyuls-sur-Mer, 66650, France. 28Cell Biology and Biophysics, European Molecular Biology Laboratory, Meyerhofstrasse 1, Heidelberg, 69117, Germany. 29MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, 28359, Germany. 30PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, 28359, Germany. 31Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland. 32CNRS, UMR 7009 Biodiv, Observatoire Océanologique, Villefranche-sur-Mer, F-06230, France. 33National Science Foundation, Arlington, VA 22230, USA. 34Laboratoire de Physique des Océans, UBO-IUEM, Place Copernic, Plouzané, 29920, France. 35Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, Paris Cedex 05, 75231, France. 36DVIP Consulting, Sèvres, 92310, France.