Molecular Ecology
May 2018, Volume 27 Issue 10 Pages 2365-2380
http://dx.doi.org/10.1111/mec.14579
http://archimer.ifremer.fr/doc/00444/55550/

Analysis of the genomic basis of functional diversity in dinoflagellates using a transcriptome-based sequence similarity network

Meng Arnaud [1,*], Corre Erwan [2], Probert Ian [3], Gutierrez-Rodriguez Andres [4], Siano Raffaele [5], Annamale Anita [6,7,8], Alberti Adriana [6,7,8], Da Silva Corinne [6,7,8], Wincker Patrick [6,7,8], Le Crom Stephane [1], Not Fabrice [9,*], Bittner Lucie [1,*]

[1] Sorbonne Univ, UPMC Univ Paris 06, Univ Antilles Guyane, Univ Nice Sophia Antipolis,CNRS,EPS IBPS, Paris, France.
[2] UPMC, CNRS, FR2424, ABiMS,Stn Biol, Roscoff, France.
[3] UPMC, CNRS, FR2424, Roscoff Culture Collect,Stn Biol Roscoff, Pl Georges Teissier, Roscoff, France.
[4] Natl Inst Water & Atmospher Res NIWA Ltd, Wellington, New Zealand.
[5] IFREMER, Ctr Brest, DYNECO PELAGOS, Plouzane, France.
[6] CEA, Inst Genom, GENOSCOPE, Evry, France.
[7] CNRS, UMR8030, Evry, France.
[8] Univ Evry Val dEssonne, Evry, France.
[9] CNRS, UMR 7144, Stn Biol Roscoff, Pl Georges Teissier, Roscoff, France.

* Corresponding authors : email addresses : arnaud.meng@etu.upmc.fr ; not@sb-roscoff.fr ; lucie.bittner@upmc.fr

Abstract :

Dinoflagellates are one of the most abundant and functionally diverse groups of eukaryotes. Despite an overall scarcity of genomic information for dinoflagellates, constantly emerging high-throughput sequencing resources can be used to characterize and compare these organisms. We assembled de novo and processed 46 dinoflagellate transcriptomes and used a sequence similarity network (SSN) to compare the underlying genomic basis of functional features within the group. This approach constitutes the most comprehensive picture to date of the genomic potential of dinoflagellates. A core-predicted proteome composed of 252 connected components (CCs) of putative conserved protein domains (pCDs) was identified. Of these, 206 were novel and 16 lacked any functional annotation in public databases. Integration of functional information in our network analyses allowed investigation of pCDs specifically associated with functional traits. With respect to toxicity, sequences homologous to those of proteins found in species with toxicity potential (e.g., sxtA4 and sxtG) were not specific to known toxin-producing species. Although not fully specific to symbiosis, the most represented functions associated with proteins involved in the symbiotic trait were related to membrane processes and ion transport. Overall, our SSN approach led to identification of 45,207 and 90,794 specific and constitutive pCDs of, respectively, the toxic and symbiotic species represented in our analyses. Of these, 56% and 57%, respectively (i.e., 25,393 and 52,193 pCDs), completely lacked annotation in public databases. This stresses the extent of our lack of knowledge, while emphasizing the potential of SSNs to identify candidate pCDs for further functional genomic characterization.

Keywords : genomics, proteomics, microbial biology, molecular evolution, protists, transcriptomics

INTRODUCTION

Dinoflagellates are unicellular eukaryotes belonging to the Alveolata lineage (Bachvaroff et al., 2014). This group encompasses a broad diversity of taxa that have a long and complex evolutionary history, play key ecological roles in aquatic ecosystems, and have significant economic impacts (reviewed in Murray et al. 2016; Janouškovec et al. 2016). The ecological success of dinoflagellates in the marine planktonic environment is assumed to be due to their ability to exhibit various survival strategies associated with an extraordinary physiological diversity (Murray et al., 2016). Nearly half of dinoflagellates have chloroplasts, but most of these are likely mixotrophic, combining photosynthetic and heterotrophic modes of nutrition (reviewed in Jeong et al. 2010; Stoecker et al. 2017). Many dinoflagellates produce toxins and form long-lasting harmful algal blooms with deleterious effects on fisheries or aquaculture (reviewed in Flewelling et al. 2005). Some species of the genus *Alexandrium* can produce toxins that effect higher trophic levels in marine ecosystems (*i.e.* copepods, fish) and are harmful to humans (Kohli et al., 2016; Murray et al., 2016; Orr et al., 2013). Members of the genus *Symbiodinium* are known to establish mutualistic symbioses with a wide diversity of benthic hosts, sustaining reef ecosystems worldwide (Goodson et al., 2001; Lin et al., 2015). Interactions between dinoflagellates and other marine organisms are extremely diverse, including (photo)symbioses (Decelle et al., 2015), predation (Jeong et al., 2010), kleptoplasty (Gast et al., 2007), and parasitism (Siano et al., 2011). Dinoflagellates have been highlighted as important members of coastal and open-ocean protistan communities based on environmental molecular barcoding surveys (Le Bescot et al., 2016; Massana et al., 2015) and the parasitic syndiniales in particular have been identified as key players that drive in situ planktonic interactions in the ocean (Lima-Mendez et al., 2015).

Along with metabarcoding surveys based on taxonomic marker genes, environmental investigations of protistan ecology and evolution involve genomic and transcriptomic data. Interpretation of such large datasets is limited by the current lack of reference data from unicellular eukaryotic planktonic organisms, resulting in a high proportion of unknown sequences (Caron et al., 2016; Sibbald and Archibald, 2017). This is particularly significant for dinoflagellates as this taxon remains poorly explored at the genome level, with only three full genome sequences published so far (Aranda et al., 2016; Lin et al., 2015; Shoguchi et al., 2013). Their genomes are notoriously big (0.5 to 40x larger than the human haploid genome) and have a complex organization (Jaeckisch et al., 2011; Murray et al., 2016; Shoguchi et al., 2013). Consequently, most recent studies investigating functional diversity of dinoflagellates rely on transcriptomic data to probe these non-model organisms.

The Moore Foundation Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP, https://www.ncbi.nlm.nih.gov/bioproject/248394, (Keeling et al., 2014)) provided the opportunity to produce a large quantity of reference transcriptomic data (Sibbald and Archibald, 2017). Among the 650 transcriptomes released, 56 were from 24 dinoflagellate genera encompassing 46 distinct strains (Keeling et al., 2014). This dataset constitutes a unique opportunity to investigate the genomic basis of the major evolutionary and ecological traits of dinoflagellates (Janouškovec et al., 2016). Performing a global analysis of such a large dataset (~3 million sequences) is challenging and requires innovative approaches. Most studies published so far have targeted specific biological processes and pathways, focusing on a small subset of the available data (Dupont et al., 2015; Kohli et al., 2016; Meyer et al., 2015). In one recent study a 101-protein dataset was used to produce a multiprotein phylogeny of dinoflagellates (Janouškovec et al., 2016). As a large fraction of the sequences produced in the MMETSP project do not have any distant homologues in current reference databases, almost half (46%) of the data remains unannotated.

With the advent of high-throughput sequencing technologies and its inherent massive production of data, sequence similarity network (SSN) approaches (Atkinson et al., 2009; Cheng et al., 2014; Méheust et al., 2016) offer an alternative to classical methods, enabling inclusion of unknown sequences in the global analysis (Forster et al., 2015; Lopez et al., 2015). In a functional genomic context, SSNs facilitate large-scale comparison of sequences, including functionally unannotated sequences, and hypothesis design based on both model and non-model organisms. For instance, SSN has been used to define enolase protein superfamilies and assign function to nearly 50% of sequences composing the superfamilies that had unknown functions (Gerlt et al., 2012). Here we used a SSN approach involving 42 *de novo* assembled transcriptomes from the MMETSP project as well as new transcriptomes of four recently described dinoflagellates to unveil the core-, accessory-, and pan-proteome of dinoflagellates and to define gene sets characteristic of selected functional traits.

RESULTS

Dataset metrics overview

A total of 46 transcriptomes were assembled and retained for further analyses using our protocol and a proteome was predicted for each transcriptome (Tab. 1). Globally, more than half of the protein-coding domains matched with functional annotations in InterPro (58%: 746,074 of 1,275,911) of which 549,459 had an identified Gene Ontology (GO) annotation. All individually assembled transcriptomes, derived proteomes and their corresponding functional annotations are available at https://figshare.com/projects/Dinoflagellate_SSN/28410.

Our SSN involves 1,275,911 vertices (protein-coding domains or, for short thereafter, domains) linked by 6,142,013 edges (pairwise sequence identity value ≥ 60%). The network consisted of 350,267 connected components (CCs) with 11,568 of these having a size from 10 to 100 vertices. It encompassed 46 proteomes having a mean of 60,661 domains with an average length of 307 bp. According to InterPro functional annotations, 50.5% of the CCs were composed of unannotated sequences only.

Identification of core / accessory / pan connected components

Global comparison analysis has been processed on 43 of the 46 proteomes that have a comparable number of domains. The analysis revealed 252 core CCs, 160,431 accessory CCs, and 347,551 pan CCs (Fig. 1A). The trend of the core proteome CC number was extrapolated using a non-linear regression model. The best-fit function was $y = a / x$, with y the predicted number of core CCs, x the number of proteomes and a an estimated parameter. For 2 to 43 proteomes, this model had a Pearson correlation coefficient of 0.97 (p-value of estimated parameter a < 2e-16). The number of core CCs for 50, 60 and 70 proteomes was extrapolated to 170, 144 and 123 CCs respectively, without displaying a saturation to a fixed number of core CCs. The Pielou diversity indices shew a mean value of 0.96, indicating the core CCs were evenly structured, i.e. rarely being dominated by a single proteome.

Functional annotation revealed that 91,4% of core domains matched to the InterPro database. According to GOslim functional categories, the most abundant annotations correspond to “ribosomal proteins” having a role in RNA translation (i.e. 7,968 of 37,842 core domains) followed by protein involved in phosphorylation, in signal transduction and in cell redox homeostasis (Fig. S2). The 37,842 core domains were further analyzed by comparison to other reference databases: the proportion of matches reached 12.5% (involved in 51 CCs) against BUSCO (Simão et al., 2015), 79.6% (involved in 190 CCs) against UniProtKB/Swiss-Prot and 93.7% (involved in 236 CCs) (Simão et al., 2015) against nr (Fig. 1B). 16 CCs (*i.e.* 946 domains) did not have any match (Fig. 1B). 101 orthologous alignments used for a recent phylogeny (Janouškovec et al., 2016) were compared to the core domains : 1606 domains from 46 CCs matched with at least one of the 101 alignments (Fig. S3, Tab. S15), but no homology was found with the domains from our 16 unknown core CCs.

Dinoflagellate functional traits investigations

In the SSN based on the 46 proteomes, the number of CCs exclusively composed of domains from species tagged with a single functional trait (trait-CCs) has been reported for each trait (Tab. S1-S9), as well as the percentage of trait-CCs (*e.g.* trait-CC including at least one InterPro functional annotation). As expected considering the taxonomic coverage of our dataset, a maximum number of trait-CCs were found for the “chloroplast” trait (336,099 CCs) whereas a minimum number was found for the “parasitism” trait (826 CCs). The “chloroplast” trait had the highest percentage of annotated trait-CCs (93%) while the “parasitism” trait had the lowest (23%) (Fig. S4). Among the trait-CCs, a total of 5 “toxicity potential” trait-CCs involving 7 of 14 possible proteomes were detected. Likewise, 2 “symbiosis” trait-CCs including 8 of 12 possible proteomes were identified (Tab. S4 & S6).

Focus on the “toxicity potential” functional trait

Well-described proteins involved in dinoflagellate toxicity, the polyketide synthases (PKS) and saxitoxins (STX) were sought within our dataset. 36 PKS homologs were identified in 17 “toxicity potential” trait-CCs (composed of 45 domains) (Tab. S10) whereas 646 PKS homologs were found in 165 non-“toxicity potential” CCs (composed of 1,144 domains). The 1,189 corresponding domains (i.e. 45 + 1,144) had either a Thiolase-like functional annotation (1,159 domains), which corresponds to the superfamily of KS enzyme domains of PKS, or lacked annotation (30 domains) according to the InterPro database. The *sxtA4* and *sxtG* genes have been reported to be found in potentially toxic species. No *stxA4* or *stxG* (i.e. genes associated with saxitoxin producing species (Stüken et al., 2011)) homolog was found in "toxicity potential" trait-CCs (Tab. S11). In contrast, 4 homologs of *stxA4* and 3 homologs of *stxG* were identified in non-"toxicity potential" trait-CCs. *sxtA4* hits correspond to 1 CCs (composed of 6 domains), and *sxtG* hits belonged to 1 CC composed of 3 domains. The 4 *sxtA4* homologs matched the InterPro annotation "pyridoxal phosphate-dependent transferase" and the 2 remaining domains of the CC were unannotated. A single InterPro annotation was found for the CC composed by *sxtG* homologs and corresponded to an amidinotransferase known as a *sxtG* protein domain (Tab. S11).

GO functional annotations of all "toxicity potential" trait-CCs revealed that the cellular component functional level, "membrane" and "integral component of membrane", annotations represented 51% and 27% of the domains respectively (Fig. 2A). At the biological process annotation level, 14% of the domains were linked to "ion transport" (Fig. 2A). At the molecular function annotation level, 24% corresponded to "protein binding" (Fig. 2A). Differential composition of functional annotations between proteomes revealed that "ion transport" protein domains occurred 7 times more often in "toxicity potential" trait-CCs whereas pentatricopeptide repeat, C2 domain, P-loop containing nucleoside triphosphate hydrolase, Pyrrolo-quinoline quinone beta-propeller repeat, Quinonprotein alcohol dehydrogenase-like and Thrombospondin type 1 repeat domains occurred 1 to 2 times more often in "toxicity potential" trait-CCs (Fig. 2B).

CCs involving most toxic representatives were investigated to reveal functions shared among toxic species only (Fig. 2C). Five core "toxicity potential" trait-CCs (corresponding to a total of 49 domains) encompassed 7 of the 14 toxic dinoflagellate proteomes considered in our analysis. Not a single of these 49 domains had a GO annotation. Based on InterPro annotations, 3 of the 5 CCs are respectively composed of 14 “nucleotide-binding alpha-beta plait” domains, 7 “P-loop containing nucleoside triphosphate hydrolase” domains and 8 “nucleotide-diphospho-sugar transferase” domains. The remaining two of these 5 CCs were entirely composed of 7 and 15 unannotated domains. Supplementary results about the taxonomic and functional composition of the core “toxicity potential” trait-CCs can be found on https://figshare.com/projects/Dinoflagellate_SSN/28410.

Among the 45,207 “toxicity potential” trait-CCs, 69% of them (*i.e.* 31,496 CCs corresponding to 70,359 domains) completely lacked InterPro functional annotations. Additional alignments to the nr database (using an e-value of 1e-3 and a sequence identity higher than 80%) revealed 6,103 hits including 283 domains, which finally lowered the number of “toxicity potential” trait-CCs without functional annotation to 25,393.

Focus on the “symbiosis” functional trait

A large range of dinoflagellates, expresses genes identified in the literature as potentially involved in symbiotic processes (Tab. S12). 150 of these gene sequences were sought in our datasets. 8 domains from 5 “symbiosis” trait-CCs were identified as proteins involved in symbiosis establishment (nodulation protein noIO and phosphoadenosine phosphosulfate reductase), cell recognition processes (merozoite surface protein), and highlighted in cnidarian-algal symbiosis (peroxiredoxin, ferritin) (Tab. S12). Similarly, 71 domains (spread across 21 CCs) were found in non-“symbiosis” trait-CCs. Functions of these 71 domains are involved in symbiosis establishment (P-type H+-ATPase, phosphoadenosine phosphosulfate reductase), cell recognition processes (merozoite surface protein 1) and exposed in cnidarian-algal symbiosis (superoxide dismutase, catalase, peroxiredoxin, glutathione peroxidase, g-glutamylcysteine synthetase).

GO functional annotations from all “symbiosis” trait-CCs (Fig. 2D) revealed that at the cellular component level, 83% of the annotations were “membrane proteins”. At the biological process level, 21% of the annotations were “ion transport” domains and 18% were involved in “protein phosphorylation”. At the molecular function level, 39% of the annotations were “protein-binding” domains, 10% were involved in “ion channel activity” and 9.9% in “calcium ion binding”. Differential composition of functional annotations between proteomes revealed 4 annotations occurring 2 to 10 times more in symbiotic lineages: ion transport, ankyrin repeat, EF-hand and zinc finger, and CCCH-type (Fig. 2E).

Two core CCs of the “symbiosis” trait involving a maximum of 8 distinct proteomes and 187 core “symbiosis” trait-CCs involving 7 proteomes (of the 12 proteomes symbiotic species available) were identified (Fig. 2F). GO annotations of these 189 core “symbiosis” trait-CCs revealed that the majority of the domains (*i.e.* 1400 out of 1896) could not be functionally annotated. Among those that could be annotated, 73.8% of the domains corresponded to “membrane proteins” (cellular component), and the remainder corresponded to “proteins of photosystem I”, “extracellular region” and “spliceosomal complex”. With respect to biological process, 31.9% of the domains were involved in ion transport while 23.8% were involved in proteolytic processes (Tab. S13). Supplementary results about the taxonomic and functional composition of the core “symbiosis” trait-CCs can be found on https://figshare.com/projects/Dinoflagellate_SSN/28410.

Among the 90,794 “symbiosis” trait-CCs, 57% of them (*i.e.* 52,491 CCs corresponding to 130,673 domains) completely lacked InterPro functional annotations. Additional alignments to the nr database (using an e-value of 1e-3 and a sequence identity higher than 80%) revealed matches for 495 domains, which finally lowered the number of “toxicity potential” trait-CCs without functional annotation to 52,193.

DISCUSSION

An efficient analysis pipeline to study non-model organisms and their dark matter

Our *de novo* assembly and downstream pipeline analysis of multiple dinoflagellate transcriptomes overcame several biases inherent to *de novo* assembly processes (Fig. S5). For instance, the domain prediction step selected transcripts involving ORFs and protein domains and allowed removal of truncated or chimeric transcripts (Yang and Smith, 2013). Data derived from high quality transcriptomes (cf. definition in the Material and Methods section) enabled construction of sequence similarity networks to focus on shared domains among multiple proteomes. Considering our 46 proteomes, a mean value of 60,661 domains was found, which is consistent with the previously estimated range of 34,156 to 75,461 genes in dinoflagellates (Murray et al., 2016). The median length of the domains was 307 bp, also consistent with the median protein length of 361 bp from genomes of 5 model species (*Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*) (Brocchieri and Karlin, 2005).

Sequence similarity networks represent an informative and pragmatic way to study massive datasets (Alvarez-Ponce et al., 2013; Atkinson et al., 2009; Cheng et al., 2014; Forster et al., 2015; Méheust et al., 2016). In (Cheng et al., 2014), 84 genome-derived proteomes of prokaryotes (i.e. 128,628 sequences) were used to study the impact of redox state changes on their gene content and evolution. The authors found that the core CCs revealed a correlation between their network structure and differences in respiratory phenotypes. Our SSN has allowed simultaneous exploration of 46 transcriptome-derived proteomes (1,275,911 sequences), including their overwhelming "dark matter" (*i.e.* here domains totally lacking functional annotation). High identity and coverage threshold values used to filter alignments ensured that only high quality alignments were included in the network (Bittner et al., 2010). The integration of 4 new dinoflagellate proteomes represented an increase of 14% of domains in the SSN and overall the dataset represents the most comprehensive picture to date of the genomic potential of dinoflagellates. This new resource and comparative genomic approach allow generation and testing of original hypotheses about the genomic basis for evolutionary history and life style, functional traits, and specificities of dinoflagellates.

Large-scale comparison of dinoflagellate proteomes confirms the extent of our lack of knowledge

The SSN analyses allowed characterization of the core and accessory proteomes for this large dataset of non-model organisms. Because our analysis relied on a *de novo* assembled, transcriptome-derived, proteome SSN rather than classical knowledge-based genomics, it also promoted discovery of new CCs, each of which can be functionally assimilated to a single putative conserved protein-domain (pCD) in such non-model organisms (Lopez et al., 2015) (Fig. S6).

The core dinoflagellate proteome identified in our analysis was composed of 252 pCDs (Fig. 1A), a size that falls in the range of the latest estimates for bacteria (352 core genes) (Yang et al., 2015) and eukaryotes (258 core genes in CEGMA, and more recently 429 single-copy orthologs in BUSCO) (Parra et al., 2007; Simão et al., 2015). The extrapolation of the number of core CCs does not saturate, suggesting that the number of core CCs for dinoflagellates could be less than 256. Our comparative analysis with the most up-to-date eukaryotic orthologous gene database BUSCO strongly stresses the need to generate more gene and protein data for non-model marine organisms in order to populate reference databases (Armengaud et al., 2014). The small overlap between core dinoflagellate pCDs and the BUSCO database suggests that essential functions expressed by dinoflagellates are distantly related to those of current model eukaryotes.

Our SSN constitutes a strong basis for exploration and refinement of functional annotations as our dataset encompassed a broad range of dinoflagellate taxa according to recent phylogenetic analyses (Bachvaroff et al., 2014; Janouškovec et al., 2016). However, the identified core proteome can only be considered partial as our dataset i- did not include representatives of all described dinoflagellate lineages, and ii- relied on transcriptomic (i.e. gene expression) data that can vary according to eco-physiological conditions and/or life-cycle stage. The content of our SSN can be however updated permanently to refine these estimates as new dinoflagellate genomic data are accumulated (Aranda et al., 2016; Lin et al., 2015; Shoguchi et al., 2013). 236 (93%) core CCs involving one or more functionally annotated domains (Fig. 2B) can be exploited to extend annotation to other aligned domains within each CC. For instance, looking for the HSP70 conserved protein domain, which is ubiquitous in all eukaryotic organisms (Germot and Philippe, 1999), 320 domain annotated as HSP70, all belonging to a single CC composed of 328 domain. The 8 remaining domain sequences were either imprecisely annotated as chaperone DnaK (1 sequence), cyclic nucleotide-binding domain (2 sequences), heat shock protein 70 family (3 sequences) or annotation was simply missing (2 sequences) (Tab. S14). As HSP70 represented 97% of the annotations, it is reasonable to extend it to all sequences forming the CC. Considering only CCs that were at least half composed of annotated domain sequences, this approach could be applied to complement the functional characterization of 49 CCs (583 unannotated domains).

Janouškovec et al., 2016 used for the first time a multi-protein dataset providing a robust phylogeny for dinoflagellates. The comparison of the 101 orthologous alignments (Janouškovec et al., 2016) with our 252 pCDs revealed that 206 of them could constitute good new candidates for refining dinoflagellate phylogeny, increasing by nearly 200% the quantity of information available for such studies.

Among the 176,958 distinct CCs entirely composed of unannotated domains, 16 CCs or pCDs (composed of 946 domain) belonged to our core dinoflagellate proteome (Fig. 1B). This highlights that many fundamental genomic features remain to be characterized in this lineage. These unknown groups of homologous domains are excellent potential candidate markers to further investigate dinoflagellate genomics at a broad scale and might also be useful for identification of dinoflagellates within complex environmental genomic datasets.

Confirmation and the new insights about the genomic bases of the toxicity

Toxic dinoflagellates represent about 80% of toxic eukaryotic phytoplankton species (Janouškovec et al., 2016). Production of toxins by dinoflagellates is well known and can cause major health and economic problems. *Karenia brevis*, for example, is known to produce brevetoxins which cause fish mortality and can affect human health through the consumption of contaminated seafood or direct exposure to harmful algal blooms (HABs) (Flewelling et al., 2005). To date, several dinoflagellate toxins have been chemically and genetically characterized (Cusick and Sayler, 2013; Kellmann et al., 2010; Stüken et al., 2011; Wang, 2008). In our SSN analyses, PKS homologs were identified in CCs composed of domains from both “toxicity potential” and non-“toxicity potential” species. This result validates a previous report that PKS proteins are not exclusive to toxic species), but are in fact involved in the production of a variety of natural products such as small acids, acetyl-CoA or propionyl-Co (Khosla et al., 2014). Spreading information among unannotated domains in both “toxicity potential” and non-“toxicity potential” trait-CCs in which PKS were identified allowed extension of the potential PKS-like annotation to 9 and 498 domains respectively (Tab. S15). PKS domains for 4 extra species (*Alexandrium catenella*, *Kryptoperidinium foliaceum*, *Protoceratium reticulatum* and

Crypthecodinium cohnii) were also detected compared to the database from (Kohli et al., 2016) (Tab. S13) (Kohli et al., 2016).

With respect to saxitoxin production, as no *sxtA4* and *sxtG* (i.e. the combination of genes associated with saxitoxin producing species (Stüken et al., 2011)) homologs were found in "toxicity potential" trait-CCs, it suggests that such proteins are also not exclusively expressed by toxic species and/or are not constitutively expressed. As Murray et al., 2015, we detected *sxtA4* and *sxtG* proteins in the transcriptomes of the toxic species *Pyrodinium bahamense* and *Gymnodinium catenatum* (Tab. S11). However, our results also differed somewhat from this previous study even if it is based on the same initial MMETSP dataset. Specifically, we were not able to detect *sxtA4* in *Alexandrium fudyense* (Murray et al., 2015) whereas *sxtA4* domains were detected *Pelagodinium beii* (Murray et al., 2015) (Tab. S11). These differences may be due to the use of distinct *de novo* assembly tools and pCD prediction processes, illustrating the requirement to ultimately combine *in vitro* and *in silico* methods in order to unambiguously characterize toxic species. We also confidently detected 1 *sxtA4* homolog and 1 *sxtG* homolog in *P. beii*, an a priori non-toxic species that has never been reported as a STX-producer. *sxtG* has previously been identified in non-toxic species (Orr et al., 2013), but the presence of both domains (*sxtA4* and *sxtG*) in a non-toxic species would be a first recorded discovery. If this pattern would not be confirmed in the future by *in silico* and *in vitro* analyses, such result might be a consequence from a contamination. MMETSP transcriptomes contaminations is furthermore a recurrent debate in the protistology community (e.g. Dorrell et al., 2017), however as our SSN vertices are labelled with the taxonomy and the strain names, it is possible and easy, whenever one decides that a strain is doubtful, to remove its corresponding vertices and edges. From an evolutionary point of view, as PKS and STX genes are also found in species currently described as non-toxic, it seems that like for snake venoms, dinoflagellate toxins might have evolved by recruitment of genes encoding regular proteins followed by gene duplication and neo-functionalization of the domains (Vonk et al., 2013).

Composition of “toxicity potential” trait-CCs showed that membrane protein and more specifically ion transport proteins are important components of toxic species. This is in agreement with that ion channel proteins and proteins involved in neurotransmission are mediators of dinoflagellate toxicity (Cusick and Sayler, 2013; Wang, 2008). Finally, 2 of the 5 CCs with the most toxic representatives (i.e. 7 species) were exclusively composed of unannotated domains, representing essential functions constitutively expressed by toxic species only and for which further investigations are required to better characterize toxic dinoflagellates.

From the study of symbiosis to the detection of genomic markers

The “symbiotic” gene set compiled from the literature based on their involvement in the establishment and maintenance of symbiosis (Lehnert et al., 2014; Lin et al., 2015) was found here in both “symbiosis” trait-CCs and in non-“symbiosis” trait-CCs (Tab. S12), suggesting that these proteins are constitutively expressed by all dinoflagellate species. This result may reflect the fact that the transcriptomes of dinoflagellate strains were not directly isolated from symbiotic conditions, but rather from their free-living stages maintained in culture. Symbiotic genes identified from the literature were originally inferred from studies on holobionts (*i.e.* host and symbionts) but proved here not to be exclusive to symbiotic dinoflagellates when performing global comparison of multiple datasets.

Functional annotations of “symbiosis” trait-CCs revealed an overall clear domination of proteins involved in phosphorylation and ion transport domains (*e.g.* sodium, potassium and calcium ion channel proteins) located within membrane compartments (Fig. 2D). The 4 most prominent functions that occurred 2 to 10 times more often in “symbiosis” trait-CCs (Fig. 2E) were related to ion transport domains and regulation processes. Protein phosphorylation is known to take part in cellular mechanisms in response to the environment (Day et al., 2016) and play a key role in signal transduction to other cells in plant parasitism and symbiosis models (Lionetti and Metraux, 2015). The specific dominant presence of ion transport domains (also involved in cell signalling and cell adaptation to the environment) in symbiotic dinoflagellates could represent a constitutive characteristic of symbiotic species facilitating establishment and maintenance of the symbiosis. Notably, the role of ion channel proteins has been highlighted as essential in plant root endosymbiosis (Charpentier et al., 2008; Matzke et al., 2009). This suggests that symbiotic species are likely to be constitutively better adapted for environmental adaptations.

45% of the domains associated to symbiotic species were unknown (Tab. S16) and 129,754 domains from 52,193 “symbiosis” trait-CCs remained unannotated according to the InterPro and nr databases. The 2 “symbiosis” trait-CCs encompassing 8 distinct species were exclusively composed of unannotated domains, suggesting that they represent pCDs with fundamental, yet unknown, functions constitutively expressed by symbiotic species. Overall, our analyses demonstrate that SSN has significant potential to reveal the variety of annotated and unknown pCDs that constitute good candidates for further study to characterize and understand the genomic basis of symbioses involving dinoflagellates.

CONCLUSION

Our efficient analysis pipeline and our innovative analysis strategy allowed us to study the genomic of non-model organisms, here dinoflagellates, and their dark matter on a massive scale. We confirmed that genes currently listed as implied in the “toxicity potential” or “symbiosis” functional traits, were not specific from toxic or symbiotic lineages, thus implying that these sequences have evolved by recruitment of genes encoding regular proteins followed by gene duplication and neo-functionalization of these domains.

By contrast, our approach, also identified candidate putative conserved protein domains for further genomic characterization of these functional traits. These markers are to date working hypotheses which will have to be further confirmed by future molecular studies (at the bench using more samples and differential expression analyses, PCR and qPCR), and also by mining directly environmental meta-omics datasets.

M&M

Dataset building

The dataset used in our study included all dinoflagellate transcriptomes available in the MMETSP project repository (https://www.ncbi.nlm.nih.gov/bioproject/248394) as well as 4 transcriptomes generated for this study (more details in the following section) (Fig. S7). This represented a total of 60 datasets (Fig. S7). Two *Pelagodinium beii* RCC1491 datasets appeared (one produced by the MMETSP, and one produced in the framework of our analysis), we nevertheless analysed them separately as sequencing experiments were performed in distinct institutes (cf. recommendations in Keeling et al 2014). Furthermore, transcriptomes from the same species but produced from different strains were pooled when the number of reads were insufficient to create two transcriptomes of “high quality” (*n.b.* a definition of the “high quality” transcriptome is given a few lines below) if the sequencing experiments were performed in the same institute. Consequently, the two *Oxyrrhis marina* strains (NA and LB1974), the two *Prorocentrum minimum* strains (CCMP1329 and CCMP2233) and the two *Polarella glacialis* strains (CCMP1383 and CCMP2088) were pooled; whereas we did not pool the *Brandtodinium* *nutricula* (RCC3387 and RCC3468) which were involving both enough reads to perform reliable assemblies.

These 60 datasets (Fig. S7) correspond to 48 distinct species from 34 genera, 18 families, and 11 of the 21 current dinoflagellate taxonomic orders according to the taxonomic framework of the WoRMS database (http://www.marinespecies.org/index.php) and Algaebase database (Guiry and Guiry, 2018). Taxonomy and functional traits information (*i.e.* chloroplast occurrence and origin, trophic mode, toxicity potential, ability to live in symbiosis, to perform kleptoplasty, to be a parasite or to be toxic for fauna) were indicated for each organism (Fig S7).

Cultivation and RNA sequencing for four dinoflagellate strains

Free-living clonal strains of the dinoflagellate species *Brandtodinium nutricula* (RCC3468) (Probert et al., 2014) and *Gymnoxanthella radiolariae* (RCC3507) (Yuasa et al., 2016) isolated from symbiotic Radiolaria, *Pelagodinium beii* (RCC1491) (Siano et al., 2010) isolated from a foraminiferan host, and the non-symbiotic *Heterocapsa* sp. (RCC1516) were obtained from the Roscoff Culture Collection (www.roscoff-culture-collection.org). Triplicate 2^{-L} acid-washed, autoclaved polycarbonate Nalgene bottles were filled with 0.2 micron filter-sterilized (Stericup-GP, Millipore) seawater with K/2 (-Tris,-Si) medium supplements (Keller et al., 1987) and inoculated with an exponentially growing culture of each strain. All cultures were maintained at 18°C, ~80 µmol photon m^{-2} s^{-1} light intensity and 14:10 light:dark cycle. Cell abundance was monitored daily by flow cytometry with a FACSAria flow cytometer (Becton Dickinson, San José, CA, USA) and derived cell division rates were used to monitor the growth phase of the culture. Light and dark phase samples for transcriptome analyses were taken from exponential and stationary phase cultures. 100 mL aliquots from each culture were filtered onto 3 micron pore-size polycarbonate filters with an autoclaved 47 mm glass vacuum filter system (Millipore) and a hand-operated PVC vacuum pump with gauge to maintain the vacuum pressure below

5 mm Hg during filtration. The filter was then placed in a sterile 15 mL falcon tube filled with ca. 5 ml TriZol and stored at -80°C.

Total RNA was purified directly from the filters stored in TriZol using the Direct-zol RNA Miniprep kit (ZymoResearch, Irvine, CA). First, the tube containing the filter immersed in TriZol was incubated for 10 min at 65°C. Then, after addition of an equal volume of 100% EtOH and vortexing, the mixture was loaded into a Zymo-SpinIIC column and centrifuged for 1 min at 12,000 g. The loading and centrifugation steps were repeated until exhaustion of the mixture. RNA purification was completed by prewash and wash steps following the manufacturer's instructions and RNA was directly eluted in 45 µL nuclease-free water. The in-column DNAse step was replaced by a more efficient post-extraction DNAse treatment using the Turbo DNA-free kit (Thermo Fisher Scientific, Waltham, MA) according to the manufacturer's rigorous DNase treatment procedure. After two rounds of 30 minutes incubation at 37°C, the reaction mixture was purified with the RNA Clean and Concentrator-5 kit (ZymoResearch) following the procedure described for retention of >17nt RNA fragments. Total RNA, eluted in 20 µL nuclease-free water, was quantified with RNA-specific fluorimetric quantification on a Qubit 2.0 Fluorometer using Qubit RNA HS Assay (ThermoFisher Scientific). RNA quality was assessed by capillary electrophoresis on an Agilent Bioanalyzer using the RNA 6000 Pico LabChip kit (Agilent Technologies, Santa Clara, CA).

RNA-Seq library preparations were carried out from 1 µg total RNA using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA), which allows mRNA strand orientation. Briefly, poly(A)+ RNA was selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. Then, the second strand was generated to create double-stranded cDNA. Strand specificity was achieved by quenching the second strand during final amplification thanks to incorporation of dUTP instead of dTTP during second strand synthesis. Then, ready-to-sequence Illumina libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina libraries (KapaBiosystems, Wilmington, MA), and library profiles evaluated with an Agilent 2100 Bioanalyzer (Agilent Technologies). Each library was sequenced using 101 bp paired-end read chemistry on a HiSeq2000 Illumina sequencer.

Data filtering and de novo assembly

Using Trimmomatic (Bolger et al., 2014), reads with quality below 30 Q on a sliding window size of 10 were excluded. Remaining reads were assembled with the *de novo* assembler Trinity version 2.1.1 (Grabherr et al., 2011) using default parameters for the paired reads method (strand-specific read orientation RF). Of the initial 60 transcriptome datasets (56 from the MMETSP repository and 4 produced in this study), 57 were successfully assembled. The assembly process could not be completed properly for 3 datasets due to a computation error from the assembly software (*Karenia brevis* strain CCMP 2229, Wilson SP1 and SP3 as a combined assembly, *Oxyrrhis marina* strain CCMP1795 and *Symbiodinium kawagutii* strain CCMP2468). Assembled transcripts were then evaluated based on: (i) sequence metrics, and (ii) read remapping rates calculated respectively with homemade scripts and Bowtie 2 in local mode (Langmead et al., 2009) (Tab. 1). Two classes of assembly quality were defined: those with >30,000 transcripts with a N50 > 400 bp and read remapping rate >50% were tagged as “high quality” transcriptomes whereas the remainders were tagged as “low quality” transcriptomes. An exception was made for one poor quality transcriptome corresponding to the species *Oxyrrhis marina* (LB1974 and NA strain) composed of 18,275 assembled transcripts that was intentionally tagged as a “high quality” transcriptome because this basal species holds a key evolutionary and ecological position among dinoflagellates (Bachvaroff et al., 2014; Lee et al., 2014; Montagnes et al., 2011).

Coding domain prediction and functional annotation

For each transcriptome, coding domain prediction of assembled transcripts was conducted with Transdecoder version 2.0.1 (Haas et al., 2013) to obtain peptide sequences of corresponding domains. We defined each set of predicted protein domains as a proteome. The optional step of Transdecoder consisting in the identification of ORFs in the protein domain database Pfam was not executed in order to avoid a comparative approach that would result in a limited discovery of new sequences. The predicted coding domains were then processed with the InterProScan 5 functional annotation program version 5.11-51.0 (Jones et al., 2014) to scan for protein signatures. Default parameters were used to obtain each proteome. Finally, to get a broad overview of the ontology content of our datasets, GO slims were retrieved from the Gene Ontology Consortium to build a summary of the GO annotations without the detail of the specific fine-grained terms (http://geneontology.org/page/go-slim-and-subset-guide).

Sequence similarity network building and exploration

A sequence similarity network (SSN) is a graph in which vertices are genomic sequences and the edges represent similarity between sequences. A SSN is composed of connected components (CC) (subgraphs or subnetworks, including at least two vertices disconnected from other subgraphs in the total network). As information can be linked to sequences (*e.g.* in our study: taxonomy, functional annotation, functional traits), the SSN and its structure can be explored accordingly. Using predicted protein domain sequences, a SSN was constructed with the BLASTp alignment method (Altschul et al., 1990) with an e-value of 1e-25 using the DIAMOND software (Buchfink et al., 2015). Similarities satisfying query and subject sequence coverages higher than 80% were kept.

Whenever domains aligned together forming a CC it can be assumed that they potentially share a similar molecular function (Marchler-Bauer et al., 2005) and form putative conserved domains (pCDs). SSN exploration and analyses were performed using R (version 3.2.3) personal scripts and functions implemented in the igraph R package (version 1.0.1) (Csárdi and Nepusz, 2006). Biological information related to the species considered were mapped on each vertex, and missing information were marked as <NA>. All scripts and the SSN (as well as the information linked to each vertices) can be found on https://figshare.com/projects/Dinoflagellate_SSN/28410.

In our approach, CC number, structure and composition were impacted when edge sequence identity cut off was shifted. We thus tested different similarity thresholds and chose an optimal threshold according to the two following criteria: maximizing the number of large CCs (*i.e.* minimum of 30 vertices) and the number of CCs involving a single homogeneous functional annotation (*i.e.* a unique GOslim term at the Biological Process level). An optimal sequence identity threshold at 60% similarity with our dataset was inferred (Fig. S1). As a last filtering step, we chose to consider only vertices and edges of proteomes that fitted the optimal threshold defined above (Fig. S7), which resulted in a dataset of 46 proteomes (Tab. 1).

43 proteomes composed of comparable numbers of protein domains (*i.e.* a minimum of 9,000 domains) (Fig. S8) were used to define the core-, accessory- and pan-proteomes. The core-proteome corresponds to the CCs composed of sequences from every single proteome considered, whereas the accessory-proteome corresponds to the CCs composed of sequences from a single proteome. The pan-proteome corresponds to the total number of CCs identified in the network. To build Fig. 1, proteomes were compared from the largest to the smallest: the two biggest datasets were first selected to

calculate the core/accessory/pan values; then the biggest remaining dataset was added to calculate the core/accessory/pan values for 3 proteomes. etc. until considering the comparison of all 43 proteomes. In addition to the InterProScan annotation process, sequences belonging to core CCs were compared to 3 databases: (i) the BUSCO core eukaryotic gene set (Simão et al., 2015), (ii) the UniProtKB/Swiss-Prot database, and (iii) the nr database, using BLASTp and an e-value of 1e-25.

To further explore the composition and structure of the CCs, we computed the Pielou equitability index (Mulder et al., 2004), classically used in ecology in order to estimate the richness and/or evenness of species in a sample. Here the Pielou index was used to estimate the contribution of each proteome in a CC, and more precisely for assessing whether a CC is mainly composed of domains from a limited number of proteomes. The index ranges from 0 to 1, and a high index corresponds to an homogeneous contribution of the proteomes.

Investigating functional traits for dinoflagellates

Analyses of functional traits were based on the SSN encompassing the 46 proteomes derived from “high quality” transcriptomes. The information about 10 selected functional traits was retrieved from the literature (Tab. 1). The details about plastid origin and presence were retrieved from (Caruana and Malin, 2014). Dinoflagellates that are capable of mixotrophy were listed in (Jeong et al., 2010). The information on species with a human (AZP, DSP, NSP, PSP, CFP syndromes) or to marine fauna (ichyotoxicity) toxicity potential was obtained from the Taxonomic Reference List of Harmful MicroAlgae of the IOC-UNESCO (http://www.marinespecies.org/hab/index.php). Dinoflagellate plastidy is reviewed in (Gagat et al., 2014). Dinoflagellates which have the capacity to produce DMSP in high cellular concentration were described in (Caruana et al., 2012). Presence of the theca, characteristic of thecate dinoflagellates, has been studied in (Lin, 2011; Orr et al., 2012). In (Rengefors et al., 1998) authors studied dinoflagellates species that go through a cyst stage during their life cycle. Symbiotic taxa are characterized in (Decelle et al., 2012; Probert et al., 2014; Siano et al., 2010; Trench and Blank, 1987; Yuasa et al., 2016). We later focused on CCs that are specific to a given trait, called "trait-CCs", defined by CCs exclusively composed of vertices tagged with this single trait (and excluding <NA> tags).

Following an exploratory approach, among trait-CCs, CCs including a maximum of distinct proteomes were sought (except for the "parasite" trait, as only one parasite proteomes is represented in the network). In this study, we examined more specifically the functional composition for the "toxicity potential" and "symbiosis" trait-CCs. To validate the SSN capacity to detect trait-CCs characteristic for a given function, we followed a knowledge-based approach searching for sequence similarities through BLASTp (e-value 1e-3) to well-known genes from the literature.

Focus on the "toxicity potential" functional trait

Specific studies on toxic dinoflagellate species have led to the establishment of defined gene sets likely related to toxin production (Snyder et al. 2003; Monroe & Van Dolah 2008; Wang 2008; Sheng et al. 2010; Kellmann et al. 2010; Stüken et al. 2011; Salcedo et al. 2012; Hackett et al. 2013; Cusick & Sayler 2013; Lehnert et al. 2014; Perini et al. 2014; Zhang et al. 2014; Kohli et al. 2015, 2016; Meyer et al. 2015; Murray et al. 2015; Beedessee et al. 2015). PKS genes are present in all dinoflagellates (Kohli et al., 2015) but many of the toxic metabolites produced by some dinoflagellate species are of polyketide origin (Kellmann et al., 2010). 2,632 polyketide synthase (PKS) peptide sequences from (Kohli et al., 2016) (supplementary data 3) were compared to sequences from "toxicity potential" trait-CCs as well as non-"toxicity potential" trait-CCs as a control (retained alignments show 80% sequence identity and 80% sequence coverage). Previous studies have also identified *sxtA4* and *sxtG* genes as related with the STX biosynthesis pathway (Orr et al., 2013). Our investigations in the "toxicity potential" and non-“toxicity potential” trait-CCs (retained alignments with 80% sequence identity and 90% sequence coverage) on were based on 26 *sxtA4* and 20 *sxtG* sequences from (Murray et al., 2015) (Tab. S17). The differential composition of functional annotations between “toxicity potential” and non-“toxicity potential” trait-CCs was investigated to detect functions that are likely more represented in toxic species. The counts of each annotation found in each functional category were respectively normalized by the total number of sequences that composed both trait-CCs. Finally, the difference of pair normalized counts for the same annotation in “toxicity potential” and non-“toxicity potential” trait-CCs was calculated (Fig. 2B).

Focus on “symbiosis” functional trait

In this study, three additional transcriptomes of symbiotic species were added to the MMETSP data to increase the number of transcriptomes of symbiotic species from 9 to 12. Following a similar strategy as for the “toxicity potential” functional trait, investigation of the “symbiosis” trait in our network was based on reported sets of genes potentially involved in the symbiotic lifestyle for *Symbiodinium kawagutii* (Lin et al., 2015) and coral symbiotic relationships (Tab. S18). We combined this set with other putative proteins highly up-regulated in anemone-dinoflagellate symbiosis (Lehnert et al., 2014). The distribution of 150 “symbiotic” marker sequences was studied across “symbiosis” trait-CCs (Tab. S12). The differential composition of functional annotations between “symbiosis” and non-“symbiosis” trait-CCs was investigated as previously described for “toxicity potential” trait-CCs.

DATA ACCESSIBILITY

SRA numbers for raw files of *Brandtodinium nutricula* (RCC3468): ERP106907, *Gymnoxanthella radiolariae* (RCC3507): available soon, *Pelagodinium beii* (RCC1491):

ERP106909, *Heterocapsa* sp. (RCC1516): ERP106906 are available on NCBI SRA database.

Personal R scripts, SSN file and attribute files for vertices and edges, Fasta files for each assembly of the MMETSP datasets and the corresponding functional annotations, Fasta files and CCs structure files corresponding to trait-CCs for each functional trait, as well as most advanced and extra analyses can be found on figshare: https://figshare.com/projects/Dinoflagellate_SSN/28410

Acknowledgments

The authors thank Gaëlle Lelandais, Laure Guillou and Éric Pelletier for their support and critical discussions. The authors We are also grateful to the RCC staff for providing dinoflagellate cultures as well as the Roscoff Bioinformatic platform ABiMS (http://abims.sb-roscoff.fr) for providing computational resources. The authors thank the three anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. This work was supported by a 3-year Ph.D. grant from “Interface Pour le Vivant” (IPV) program at the Université Pierre et Marie Curie (UPMC), Paris, and this project was supported by grants from Région Ile-de-France.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Alvarez-Ponce, D., Lopez, P., Bapteste, E., and McInerney, J.O. (2013). Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc. Natl. Acad. Sci. U. S. A. 110, E1594-1603.

Aranda, M., Li, Y., Liew, Y.J., Baumgarten, S., Simakov, O., Wilson, M.C., Piel, J., Ashoor, H., Bougouffa, S., Bajic, V.B., et al. (2016). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Sci. Rep. 6.

Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumot, A., and Hartmann, E.M. (2014). Non-model organisms, a species endangered by proteogenomics. J. Proteomics 105, 5–18.

Atkinson, H.J., Morris, J.H., Ferrin, T.E., and Babbitt, P.C. (2009). Using Sequence Similarity Networks for Visualization of Relationships Across Diverse Protein Superfamilies. PLoS ONE 4.

Bachvaroff, T.R., Gornik, S.G., Concepcion, G.T., Waller, R.F., Mendez, G.S., Lippmeier, J.C., and Delwiche, C.F. (2014). Dinoflagellate phylogeny revisited: Using ribosomal proteins to resolve deep branching dinoflagellate clades. Mol. Phylogenet. Evol. 70, 314–322.

Beedessee, G., Hisata, K., Roy, M.C., Satoh, N., and Shoguchi, E. (2015). Multifunctional polyketide synthase genes identified by genomic survey of the symbiotic dinoflagellate, Symbiodinium minutum. BMC Genomics 16.

Bittner, L., Halary, S., Payri, C., Cruaud, C., de Reviers, B., Lopez, P., and Bapteste, E. (2010). Some considerations for analyzing biodiversity using integrative metagenomics and gene networks. Biol. Direct 5, 47.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics btu170.

Brocchieri, L., and Karlin, S. (2005). Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. 33, 3390–3400.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12, 59–60.

Caron, D.A., Alexander, H., Allen, A.E., Archibald, J.M., Armbrust, E.V., Bachy, C., Bell, C.J., Bharti, A., Dyhrman, S.T., Guida, S.M., et al. (2016). Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat. Rev. Microbiol. advance online publication.

Caruana, A.M.N., and Malin, G. (2014). The variability in DMSP content and DMSP lyase activity in marine dinoflagellates. Prog. Oceanogr. 120, 410–424.

Caruana, A.M.N., Steinke, M., Turner, S.M., and Malin, G. (2012). Concentrations of dimethylsulphoniopropionate and activities of dimethylsulphide-producing enzymes in batch cultures of nine dinoflagellate species. Biogeochemistry 110, 87–107.

Charpentier, M., Bredemeier, R., Wanner, G., Takeda, N., Schleiff, E., and Parniske, M. (2008). Lotus japonicus CASTOR and POLLUX Are Ion Channels Essential for Perinuclear Calcium Spiking in Legume Root Endosymbiosis. Plant Cell 20, 3467–3479.

Cheng, S., Karkar, S., Bapteste, E., Yee, N., Falkowski, P., and Bhattacharya, D. (2014). Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front. Ecol. Evol. 2.

Csárdi, G., and Nepusz, T. (2006). The igraph software package for complex network research. InterJournal Complex Syst.

Cusick, K.D., and Sayler, G.S. (2013). An Overview on the Marine Neurotoxin, Saxitoxin: Genetics, Molecular Targets, Methods of Detection and Ecological Functions. Mar. Drugs 11, 991–1018.

Day, E.K., Sosale, N.G., and Lazzara, M.J. (2016). Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. Curr. Opin. Biotechnol. 40, 185–192.

Decelle, J., Probert, I., Bittner, L., Desdevises, Y., Colin, S., Vargas, C. de, Galí, M., Simó, R., and Not, F. (2012). An original mode of symbiosis in open ocean plankton. Proc. Natl. Acad. Sci. 109, 18000–18005.

Decelle, J., Colin, S., and Foster, R.A. (2015). Photosymbiosis in Marine Planktonic Protists. In Marine Protists, S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not, eds. (Springer Japan), pp. 465–500.

Dorrell, R.G., Gile, G., McCallum, G., Méheust, R., Bapteste, E.P., Klinger, C.M., Brillet-Guéguen, L., Freeman, K.D., Richter, D.J., and Bowler, C. (2017). Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. ELife 6, e23717.

Dupont, C.L., McCrow, J.P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., Roth, R., Hogle, S.L., Bai, J., Johnson, Z.I., et al. (2015). Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. ISME J. 9, 1076–1092.

Flewelling, L.J., Naar, J.P., Abbott, J.P., Baden, D.G., Barros, N.B., Bossart, G.D., Bottein, M.-Y.D., Hammond, D.G., Haubold, E.M., Heil, C.A., et al. (2005). Brevetoxicosis: Red tides and marine mammal mortalities. Nature 435, 755–756.

Forster, D., Bittner, L., Karkar, S., Dunthorn, M., Romac, S., Audic, S., Lopez, P., Stoeck, T., and Bapteste, E. (2015). Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. BMC Biol. 13, 16.

Gagat, P., Bodył, A., Mackiewicz, P., and Stiller, J.W. (2014). Tertiary Plastid Endosymbioses in Dinoflagellates. In Endosymbiosis, W. Löffelhardt, ed. (Springer Vienna), pp. 233–290.

Gast, R.J., Moran, D.M., Dennett, M.R., and Caron, D.A. (2007). Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition? Environ. Microbiol. 9, 39–45.

Gerlt, J.A., Babbitt, P.C., Jacobson, M.P., and Almo, S.C. (2012). Divergent Evolution in Enolase Superfamily: Strategies for Assigning Functions. J. Biol. Chem. 287, 29–34.

Germot, A., and Philippe, H. (1999). Critical Analysis of Eukaryotic Phylogeny: A Case Study Based on the HSP70 Family. J. Eukaryot. Microbiol. 46, 116–124.

Goodson, M.S., Whitehead, L.F., and Douglas, A.E. (2001). Symbiotic dinoflagellates in marine Cnidaria: diversity and function. Hydrobiologia 461, 79–82.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome (Trinity). Nat. Biotechnol. 29, 644–652.

Guiry, M.D., and Guiry, G.M. (2018). AlgaeBase.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8, 1494–1512.

Hackett, J.D., Wisecaver, J.H., Brosnahan, M.L., Kulis, D.M., Anderson, D.M., Bhattacharya, D., Plumley, F.G., and Erdner, D.L. (2013). Evolution of Saxitoxin Synthesis in Cyanobacteria and Dinoflagellates. Mol. Biol. Evol. 30, 70–78.

Jaeckisch, N., Yang, I., Wohlrab, S., Glöckner, G., Kroymann, J., Vogel, H., Cembella, A., and John, U. (2011). Comparative Genomic and Transcriptomic Characterization of the Toxigenic Marine Dinoflagellate Alexandrium ostenfeldii. PLOS ONE 6, e28012.

Janouškovec, J., Gavelis, G.S., Burki, F., Dinh, D., Bachvaroff, T.R., Gornik, S.G., Bright, K.J., Imanian, B., Strom, S.L., Delwiche, C.F., et al. (2016). Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proc. Natl. Acad. Sci. 201614842.

Jeong, H.J., Yoo, Y.D., Kim, J.S., Seong, K.A., Kang, N.S., and Kim, T.H. (2010). Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Sci. J. 45, 65–91.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinforma. Oxf. Engl. 30, 1236–1240.

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. PLOS Biol 12, e1001889.

Keller, M.B., Lavori, P.W., Friedman, B., Nielsen, E., Endicott, J., McDonald-Scott, P., and Andreasen, N.C. (1987). The Longitudinal Interval Follow-up Evaluation. A comprehensive method for assessing outcome in prospective longitudinal studies. Arch. Gen. Psychiatry 44, 540–548.

Kellmann, R., Stüken, A., Orr, R.J.S., Svendsen, H.M., and Jakobsen, K.S. (2010). Biosynthesis and Molecular Genetics of Polyketides in Marine Dinoflagellates. Mar. Drugs 8, 1011–1048.

Khosla, C., Herschlag, D., Cane, D.E., and Walsh, C.T. (2014). Assembly Line Polyketide Synthases: Mechanistic Insights and Unsolved Problems. Biochemistry (Mosc.) 53, 2875–2883.

Kohli, G.S., John, U., Figueroa, R.I., Rhodes, L.L., Harwood, D.T., Groth, M., Bolch, C.J.S., and Murray, S.A. (2015). Polyketide synthesis genes associated with toxin production in two species of Gambierdiscus (Dinophyceae). BMC Genomics 16, 410.

Kohli, G.S., John, U., Van Dolah, F.M., and Murray, S.A. (2016). Evolutionary distinctiveness of fatty acid and polyketide synthesis in eukaryotes. ISME J.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Le Bescot, N., Mahé, F., Audic, S., Dimier, C., Garet, M.-J., Poulain, J., Wincker, P., de Vargas, C., and Siano, R. (2016). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. Environ. Microbiol. 18, 609–626.

Lee, R., Lai, H., Malik, S.B., Saldarriaga, J.F., Keeling, P.J., and Slamovits, C.H. (2014). Analysis of EST data of the marine protist Oxyrrhis marina, an emerging model for alveolate biology and evolution. BMC Genomics 15, 122.

Lehnert, E.M., Mouchka, M.E., Burriesci, M.S., Gallo, N.D., Schwarz, J.A., and Pringle, J.R. (2014). Extensive Differences in Gene Expression Between Symbiotic and Aposymbiotic Cnidarians. G3 GenesGenomesGenetics 4, 277–295.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., et al. (2015). Determinants of community structure in the global plankton interactome. Science 348, 1262073.

Lin, S. (2011). Genomic understanding of dinoflagellates. Res. Microbiol. 162, 551–569.

Lin, S., Cheng, S., Song, B., Zhong, X., Lin, X., Li, W., Li, L., Zhang, Y., Zhang, H., Ji, Z., et al. (2015). The Symbiodinium kawagutii genome illuminates dinoflagellate gene expression and coral symbiosis. Science 350, 691–694.

Lionetti, V., and Metraux, J.-P. (2015). Plant cell wall in pathogenesis, parasitism and symbiosis (Frontiers Media SA).

Lopez, P., Halary, S., and Bapteste, E. (2015). Highly divergent ancient gene families in metagenomic samples are compatible with additional divisions of life. Biol. Direct 10, 64.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., et al. (2005). CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res. 33, D192–D196.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., Chambouvet, A., Christen, R., Claverie, J.-M., Decelle, J., et al. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. Environ. Microbiol. 17, 4035–4049.

Matzke, M., Weiger, T.M., Papp, I., and Matzke, A.J.M. (2009). Nuclear membrane ion channels mediate root nodule development. Trends Plant Sci. 14, 295–298.

Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P., and Bapteste, E. (2016). Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc. Natl. Acad. Sci. 113, 3579–3584.

Meyer, J.M., Rödelsperger, C., Eichholz, K., Tillmann, U., Cembella, A., McGaughran, A., and John, U. (2015). Transcriptomic characterisation and genomic glimps into the toxigenic dinoflagellate Azadinium spinosum, with emphasis on polykeitde synthase genes. BMC Genomics 16.

Monroe, E.A., and Van Dolah, F.M. (2008). The Toxic Dinoflagellate Karenia brevis Encodes Novel Type I-like Polyketide Synthases Containing Discrete Catalytic Domains. Protist 159, 471–482.

Montagnes, D.J.S., Lowe, C.D., Roberts, E.C., Breckels, M.N., Boakes, D.E., Davidson, K., Keeling, P.J., Slamovits, C.H., Steinke, M., Yang, Z., et al. (2011). An introduction to the special issue: Oxyrrhis marina, a model organism? J. Plankton Res. 33, 549–554.

Mulder, C.P.H., Bazeley-White, E., Dimitrakopoulos, P.G., Hector, A., Scherer-Lorenzen, M., and Schmid, B. (2004). Species evenness and productivity in experimental plant communities. Oikos 107, 50–63.

Murray, S.A., Diwan, R., Orr, R.J.S., Kohli, G.S., and John, U. (2015). Gene duplication, loss and selection in the evolution of saxitoxin biosynthesis in alveolates. Mol. Phylogenet. Evol. 92, 165–180.

Murray, S.A., Suggett, D.J., Doblin, M.A., Kohli, G.S., Seymour, J.R., Fabris, M., and Ralph, P.J. (2016). Unravelling the functional genetics of dinoflagellates: a review of approaches and opportunities. Perspect. Phycol. 37–52.

Orr, R.J.S., Murray, S.A., Stüken, A., Rhodes, L., and Jakobsen, K.S. (2012). When Naked Became Armored: An Eight-Gene Phylogeny Reveals Monophyletic Origin of Theca in Dinoflagellates. PLoS ONE 7, e50004.

Orr, R.J.S., Stüken, A., Murray, S.A., and Jakobsen, K.S. (2013). Evolutionary Acquisition and Loss of Saxitoxin Biosynthesis in Dinoflagellates: the Second “Core” Gene, sxtG. Appl. Environ. Microbiol. 79, 2128–2136.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinforma. Oxf. Engl. 23, 1061–1067.

Perini, F., Galluzzi, L., Dell’Aversano, C., Dello Iacovo, E., Tartaglione, L., Ricci, F., Forino, M., Ciminiello, P., and Penna, A. (2014). SxtA and sxtG Gene Expression and Toxin Production in the Mediterranean Alexandrium minutum (Dinophyceae). Mar. Drugs 12, 5258–5276.

Probert, I., Siano, R., Poirier, C., Decelle, J., Biard, T., Tuji, A., Suzuki, N., and Not, F. (2014). Brandtodinium gen. nov. and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. J. Phycol. 50, 388–399.

Rengefors, K., Karlsson, I., and Hansson, L.-A. (1998). Algal cyst dormancy: a temporal escape from herbivory. Proc. R. Soc. B Biol. Sci. 265, 1353–1358.

Salcedo, T., Upadhyay, R.J., Nagasaki, K., and Bhattacharya, D. (2012). Dozens of Toxin-Related Genes Are Expressed in a Nontoxic Strain of the Dinoflagellate Heterocapsa circularisquama. Mol. Biol. Evol. 29, 1503–1506.

Sheng, J., Malkiel, E., Katz, J., Adolf, J.E., and Place, A.R. (2010). A dinoflagellate exploits toxins to immobilize prey prior to ingestion. Proc. Natl. Acad. Sci. 107, 2082–2087.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., Takeuchi, T., Hisata, K., Tanaka, M., Fujiwara, M., et al. (2013). Draft Assembly of the Symbiodinium minutum Nuclear Genome Reveals Dinoflagellate Gene Structure. Curr. Biol. 23, 1399–1408.

Siano, R., Montresor, M., Probert, I., Not, F., and de Vargas, C. (2010). Pelagodinium gen. nov. and P. béii comb. nov., a dinoflagellate symbiont of planktonic foraminifera. Protist 161, 385–399.

Siano, R., Alves-de-Souza, C., Foulon, E., Bendif, E.M., Simon, N., Guillou, L., and Not, F. (2011). Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. Biogeosciences 8, 267–278.

Sibbald, S.J., and Archibald, J.M. (2017). More protist genomes needed. Nat. Ecol. Evol. 1, 0145.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics btv351.

Snyder, R.V., Gibbs, P.D.L., Palacios, A., Abiy, L., Dickey, R., Lopez, J.V., and Rein, K.S. Polyketide Synthase Genes from Marine Dinoflagellates. Mar. Biotechnol. 5, 1–12.

Stoecker, D.K., Hansen, P.J., Caron, D.A., and Mitra, A. (2017). Mixotrophy in the Marine Plankton. Annu. Rev. Mar. Sci. 9, 311–335.

Stüken, A., Orr, R.J.S., Kellmann, R., Murray, S.A., Neilan, B.A., and Jakobsen, K.S. (2011). Discovery of Nuclear-Encoded Genes for the Neurotoxin Saxitoxin in Dinoflagellates. PLOS ONE 6, e20096.

Trench, R.K., and Blank, R.J. (1987). Symbiodinium Microadriaticum Freudenthal, S. Goreauii Sp. Nov., S. Kawagutii Sp. Nov. and S. Pilosum Sp. Nov.: Gymnodinioid Dinoflagellate Symbionts of Marine Invertebrates 1. J. Phycol. 23, 469–481.

Vonk, F.J., Casewell, N.R., Henkel, C.V., Heimberg, A.M., Jansen, H.J., McCleary, R.J.R., Kerkkamp, H.M.E., Vos, R.A., Guerreiro, I., Calvete, J.J., et al.

(2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proc. Natl. Acad. Sci. 110, 20651–20656.

Wang, D.-Z. (2008). Neurotoxins from Marine Dinoflagellates: A Brief Review. Mar. Drugs 6, 349–371.

Yang, Y., and Smith, S.A. (2013). Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14, 328.

Yang, L., Tan, J., O’Brien, E.J., Monk, J.M., Kim, D., Li, H.J., Charusanti, P., Ebrahim, A., Lloyd, C.J., Yurkovich, J.T., et al. (2015). Systems biology definition of the core proteome of metabolism and expression is consistent with high-throughput data. Proc. Natl. Acad. Sci. 112, 10810–10815.

Yuasa, T., Horiguchi, T., Mayama, S., and Takahashi, O. (2016). Gymnoxanthella radiolariae gen. et sp. nov. (Dinophyceae), a dinoflagellate symbiont from solitary polycystine radiolarians. J. Phycol. 52, 89–104.

Zhang, Y., Zhang, S.-F., Lin, L., and Wang, D.-Z. (2014). Comparative Transcriptome Analysis of a Toxin-Producing Dinoflagellate Alexandrium catenella and Its Non-Toxic Mutant. Mar. Drugs 12, 5698–5718.

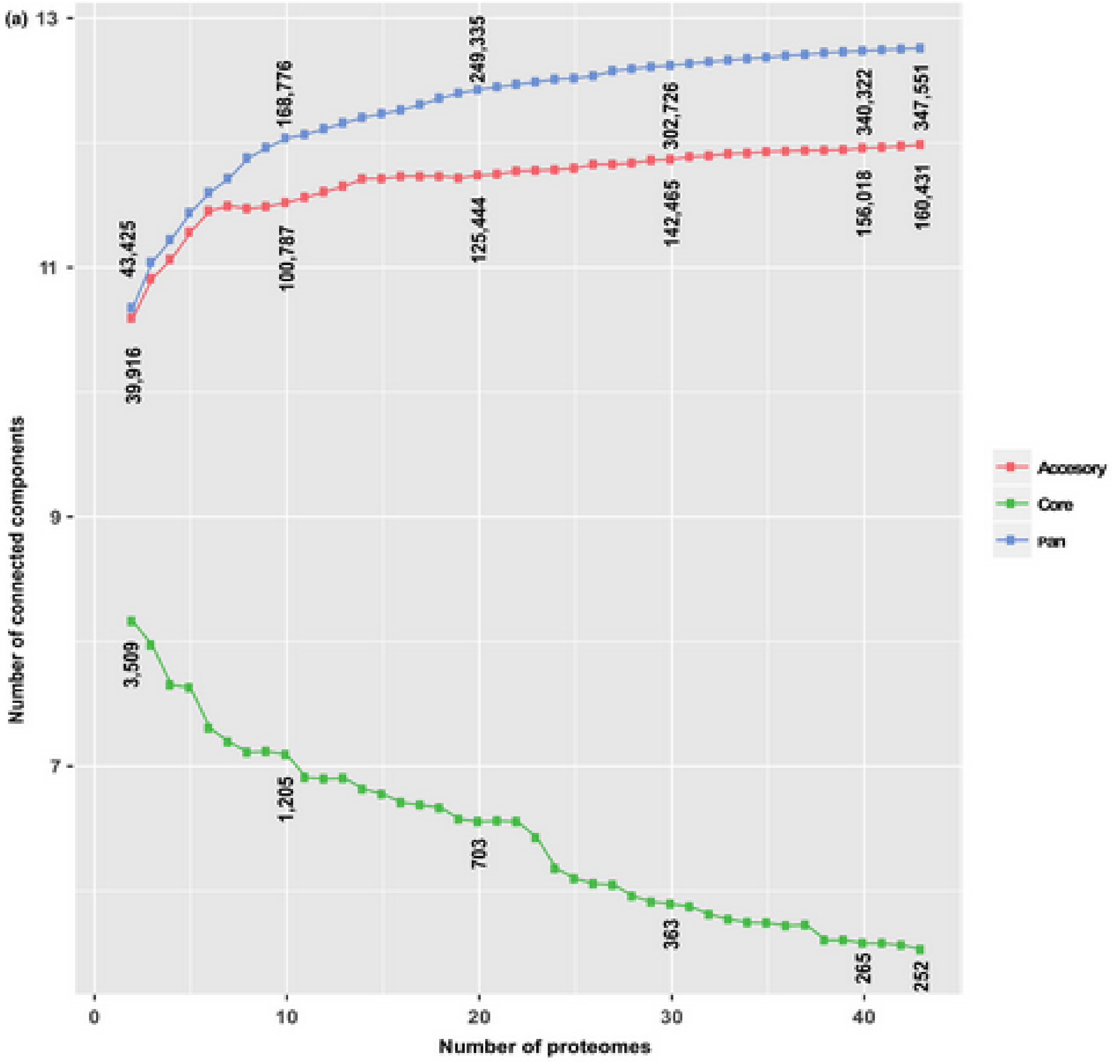
Tab 1. Transcriptomes taxonomy, assembly metrics and functional traits
Summary table of the 46 transcriptomes (and the corresponding strains) analyzed in this study, ranked based on their taxonomy. For 3 datasets, strain has been pooled: *Oxyrrhis marina* strains (NA and LB1974), the two *Prorocentrum minimum* strains (CCMP1329 and CCMP2233) and the two *Polarella glacialis* strains (CCMP1383 and CCMP2088). In contrast we kept separated strains from MMETSP and our datasets: the two *Pelagodinium beii* RCC1491 strains (see M&M section). Assembly metrics are reported for each transcriptome encompassing: the number of assembled contigs, N50, the remapping rate of initial reads, the number of predicted protein domains found in transcript sequences and the number of functional annotations identified through Interproscan 5. The proteomes derived from the 46 transcriptomes presented here are included in the sequence similarity network. Based on a literature survey, information about functional traits for each species included in the dataset is provided: chloroplast type (P: peridinin, H: haptophyte-like, C: cryptomonad-like, D: diatom-like, R: remnant or absent plastid, NC: non-consitutive chloroplast), mixotrophy, toxicity potential (DSP:Diarrhetic shellfish poisoning, CFP:Ciguatera Fish Poisoning, PSP:Paralytic shellfish poisoning, AZP:Azaspiracid Shellfish Poisoning, NSP:Neurologic Shellfish Poisoning), ability to be symbionts, kleptoplasty, ichyotoxicity, parasitism, ability to produce DSMP, presence of a theca, ability to form cysts during life-cycle. <NA> corresponds to a lack of information.

Guiry, M.D. & Guiry, G.M. 2018. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 08 February 2018.

Transcriptome taxonomy, metrics & function traits

ID	TAXONOMY					ASSEMBLY METRICS					FUNCTIONAL TRAITS									
	order	family	genus	specie	strain	# contigs	N50	remapping rates	# protein coding domains	# annotations	chloroplast type	kleptoplasty	mixotrophy	toxicity potential	symbiont	ichyotoxicity	parasitism	DMSP	thecate	cyst forming
1	Dinophysiales	Dinophysiaceae	*Dinophysis*	*acuminata*	DAEP01	123 473	747	73,07	57 612	21 401	NC	y	y	DSP	n	n	n	y	y	n
2	Gonyaulacales	Ceratiaceae	*Ceratium*	*fusus*	PA161109	147 425	1 234	81,94	77 757	28 582	P	n	y	n	n	n	n	n	n	n
3	Gonyaulacales	Crypthecodiniaceae	*Crypthecodinium*	*cohnii*	Seligo	102 139	1 396	84,64	37 992	15 703	R	n	n	n	n	n	n	y	y	n
4	Gonyualacales	Goniodomataceae	*Gambierdiscus*	*australes*	CAWD149	95 306	812	83,2	47 902	17 321	P	n	n	CFP	n	n	n	n	n	n
5	Gonyaulacales	Goniodomataceae	*Pyrodinium*	*bahamense*	pbaha01	142 061	772	75,18	73 648	26 001	P	n	n	PSP	n	n	n	n	n	y
6	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*andersoni*	CCMP2222	97 010	438	56,64	41 556	13 166	P	n	y	PSP	n	n	n	n	y	y
7	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*catenella*	OF101	95 316	570	69,07	51 078	17 364	P	n	y	PSP	n	n	n	n	y	y
8	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*margalefi*	AMGDE01CS-322	145 973	825	80,58	87 070	29 537	P	n	n	n	n	n	n	n	y	y
9	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*minutum*	CCMP113	21 364	550	41	10 572	3 817	P	n	y	PSP	n	n	n	y	y	y
10	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*monilatum*	CCMP3105	114 652	1 404	84,63	75 921	26 620	P	n	y	PSP	n	n	n	n	y	y
11	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*tamarense*	CCMP1771	176 197	1 065	79,45	91 414	36 592	P	n	y	PSP	n	n	n	y	y	y
12	Gonyaulacales	Gonyaulacaceae	*Gonyaulax*	*spinifera*	CCMP409	70 621	634	73,93	33 151	12 928	P	n	y	n	n	n	n	y	y	y
13	Gonyaulacales	Gonyaulacaceae	*Lingulodinium*	*polyedra*	CCMP1738	131 324	1 278	86,04	80 900	28 396	P	n	y	DSP	n	n	n	y	y	y
14	Gonyaulacales	Gonyaulacaceae	*Protoceratium*	*reticulatum*	CCCM535=CCMP1889	96 484	650	70,13	50 156	18 700	P	n	n	DSP	n	n	n	n	n	y
15	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*carterae*	CCMP1314	60 662	1 590	74	37 749	15 747	P	n	y	n	n	y	n	y	n	n
16	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*massartii*	CS-259	76 973	1 280	85,32	40 678	16 207	P	n	y	n	n	n	n	n	n	n
17	Gymnodiniales	Gymnodiniaceae	*Gymnodinium*	*catenatum*	GC744	124 421	836	74,72	54 459	22 417	P	n	y	PSP	n	n	n	n	n	y
18	Gymnodiniales	Gymnodiniaceae	*Gymnoxanthella*	*radiolariae*	RCC3507	160 971	1 683	93,11	102 709	39 515	P	n	n	n	y	n	n	y	n	n
19	Gymnodiniales	Gymnodiniaceae	*Togula*	*jolla*	CCCM725	73 075	1 054	81,34	35 840	15 309	P	n	NA	n	n	n	n	n	n	n
20	Gymnodiniales	Gymnodiniaceae	*Karlodinium*	*micrum*	CCMP2283	142 286	1 330	84,41	65 800	28 395	H	n	y	n	n	y	n	n	n	n
21	Dinophyceae incertae sedis	Noctilucaceae	*Noctiluca*	*scintillans*	SPMC136	66 050	1 230	84,07	33 017	14 223	R	n	n	n	n	n	n	n	n	n
22	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	NA LB1974	18 275	569	42,15	5 189	2 402	R	n	n	n	n	n	n	n	n	y
23	Peridiniales	Heterocapsaceae	*Heterocapsa*	*sp.*	RCC1516	225 203	1 289	88,62	107 673	36 966	P	n	y	n	n	n	n	y	y	n
24	Peridiniales	Heterocapsaceae	*Heterocapsa*	*arctica*	CCMP445	62 237	628	66,3	33 122	12 078	P	n	NA	n	n	n	n	n	y	n
25	Peridiniales	Heterocapsaceae	*Heterocapsa*	*rotundata*	SCCAPK-0483	69 955	774	72,65	39 543	14 077	P	n	y	n	n	n	n	y	y	n
26	Peridiniales	Heterocapsaceae	*Heterocapsa*	*triquetra*	CCMP448	89 751	698	68,45	44 370	16 265	P	n	y	n	n	n	n	y	y	n
27	Peridiniales	*Amphidomataceae*	*Azadinium*	*spinosum*	3D9	152 890	1 269	83,7	76 500	30 065	P	n	NA	AZP	n	n	n	n	y	n
28	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3387	92 032	672	66,47	59 250	19 378	P	n	n	n	y	n	n	y	y	n
29	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3468	187 598	1 199	89,84	115 229	36 197	P	n	n	n	y	n	n	y	y	n
30	Peridiniales	Peridiniaceae	*Durinskia*	*baltica*	CSIRO_CS-38	158 433	836	77,96	71 415	29 330	D	n	NA	n	n	n	n	n	y	NA
31	Peridiniales	Peridiniaceae	*Glenodinium*	*foliaceum*	CCAP1116/3	154 714	746	76,33	82 653	29 409	P	n	NA	n	n	n	n	n	n	y
32	Peridiniales	Peridiniaceae	*Kryptoperidinium*	*foliaceum*	CCMP1326	254 192	792	70,28	135 557	48 835	D	n	NA	n	n	n	n	y	y	n
33	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei*	SHTV-5	194 233	1 526	86,93	114 374	37 917	P	n	n	n	n	n	n	n	y	y
34	Peridiniales	Peridiniaceae	*Scrippsiella*	*trochoidea*	CCMP3099	160 890	1 386	82,78	90 198	34 206	P	n	y	n	n	n	n	y	y	y
35	Prorocentrales	Prorocentraceae	*Prorocentrum*	*minimum*	CCMP1329 CCMP2233	110 115	710	66,53	45 564	17 693	P	n	y	DSP	n	n	n	n	y	n
36	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	154 473	1 513	92,2	111 658	36 604	P	n	n	n	y	n	n	y	n	n
37	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	99 728	946	75,84	44 901	18 705	P	n	n	n	y	n	n	y	n	n
38	Suessiales	Suessiaceae	*Polarella*	*glacialis*	CCMP1383 CCMP2088	108 029	794	68,56	46 056	17 766	P	n	n	n	n	n	n	y	n	y
39	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	D1a	142 720	493	53,9	52 578	20 131	P	n	n	n	y	n	n	y	y	n
40	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP421	136 116	965	75,98	81 880	29 878	P	n	n	n	y	n	n	y	y	n
41	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C15	101 453	1 024	80,66	48 884	18 687	P	n	n	n	y	n	n	y	y	n
42	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C1	89 177	1 181	84,07	52 745	21 085	P	n	n	n	y	n	n	y	y	n
43	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP2430	79 016	1 082	87,26	50 160	19 992	P	n	n	n	y	n	n	y	y	n
44	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	Mp	74 565	1 416	89,35	46 513	18 550	P	n	n	n	y	n	n	y	y	n
45	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	cladeA	72 446	947	80,77	41 846	14 844	P	n	n	n	y	n	n	y	y	n
46	Syndiniales	Amoebophryaceae	*Amoebophrya*	*sp.*	Ameob2	26 721	1 856	87,46	5 548	2 075	R	n	n	n	n	n	y	n	n	n

Fig. 1. Unveiling the core, accessory and pan proteome of 43 dinoflagellates proteomes.
(A) Number of connected components (CCs) in the core (green), accessory (red) and pan (blue) dinoflagellate proteomes, considering 2 to 43 proteomes. (B) Comparison of the 37,842 protein domains included in the 252 core dinoflagellates CCs to BUSCO, UniProtKB/Swiss-Prot and nr databases. The number and percentage of core sequences with at least one match in each database, and the number and percentage of their corresponding CCs.



(b)

		Number of sequence	Percentage of sequences (*)	Number of involved components	Percentage involved components (**)
sequence homology	BUSCO version: 06/2016	4,737	12.5%	51	20.2%
	UniProtKB/SwissProt version: 06/2016	30,138	79.6%	190	75.4%
	nr version: 12/12/2015	35,455	93.7%	236	93.7%
protein domain	InterPro version: 06/2016	334,573	91.4%	226	89.7%
	Remaining unnanotated domains	946	2.5%	16	6.3%

* Total number of core sequences = 37,842
** Total number of core components = 252

Fig. 2. Exploring functions in toxicity potential trait-CCs and symbiosis trait-CCs.

(A and D) Top 10 functional annotations (GOslim levels) of sequences belonging to the 45,207 “toxicity potential” trait-CCs (A) and to the 90,794 “symbiosis” trait-CCs (D). (B and E) Differential composition of functional annotations between “toxicity potential” and non-“toxicity potential” trait-CCs (B) and “symbiosis” and non-“symbiosis” trait-CCs (E). (C and F) The circular barplot shows the number of connected components that include 1 to 14 proteome(s) of the transcriptomes assigned to toxic species (C) and the number of connected components that include 1 to 12 proteome(s) of the transcriptomes assigned to symbiotic species (F).

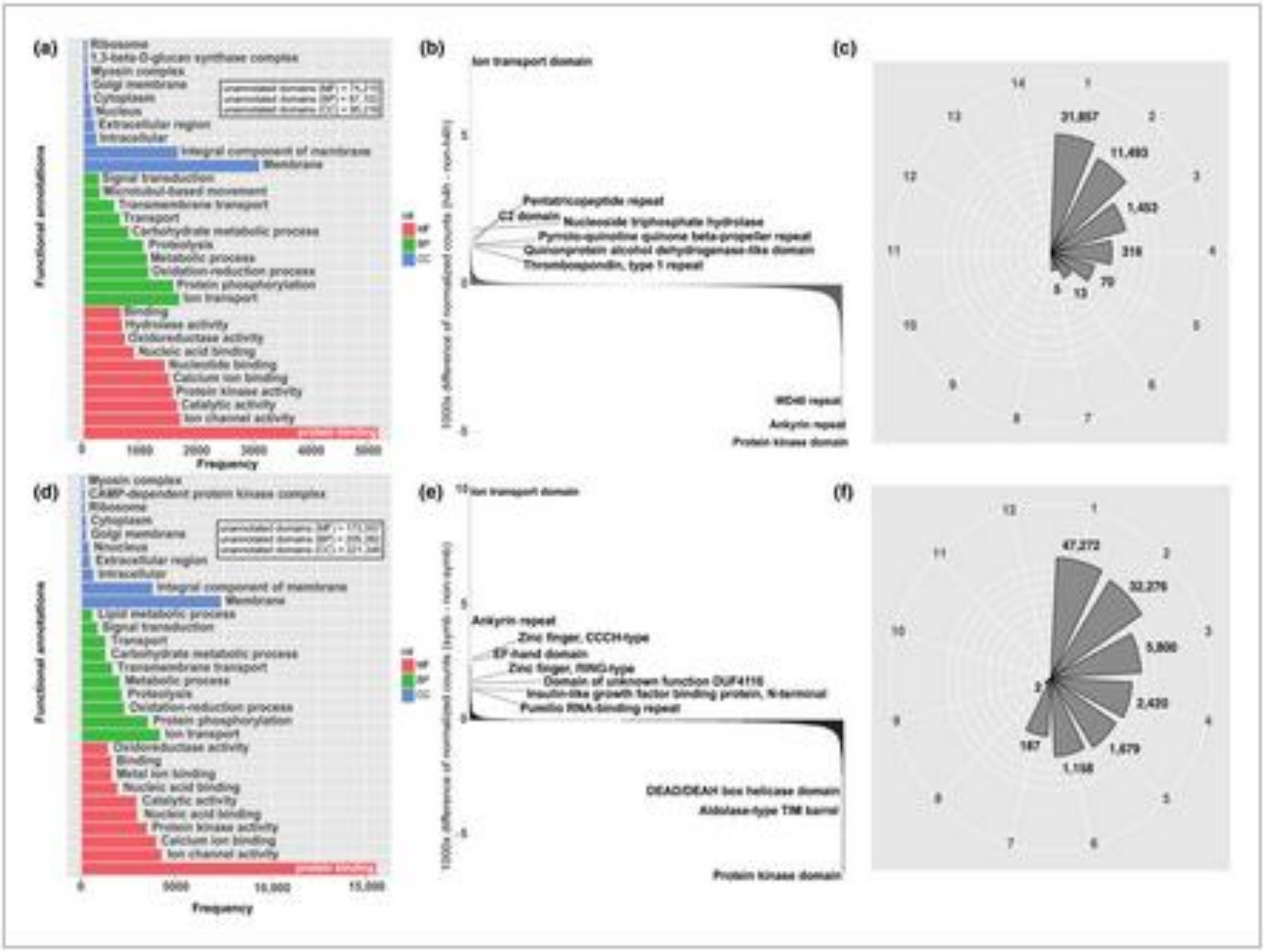


Fig. S1: Optimal sequence identity threshold selection. The cutoff was chosen such that: (A) the network contains a maximum of connected components with homogeneous functional annotation (i.e. a unique GO Slim term for all annotated protein coding domains in each CC) and (B) the network conserved a maximum of « large » connected components.

Fig. S2: Top 10 functional annotations (GOslim levels) in all core components. The three levels of annotation are represented: Molecular Function level (MF, red), Biological Process (BP, green) and Cellular Component (CC, blue).

Fig. S3: Overlap between the 101 multiple protein alignments from (2) (used for phylogenomics) and our core dinoflagellate proteome. The number of copies of the 101 proteins is represented here in a heat map: the blue color gradient represents the number of orthologous sequences available of a particular protein (y-axis) for each studied dinoflagellate species in (2) (x-axis). The overlap between the data set from (2) and our core proteome is represented by the red boxes: i.e. each protein sequence from (2) that aligned to at least a core protein coding domain is here delimited in red.

Fig. S4: (A) Number of connected components for each functional trait. (B) Proportion of annotated sequences of connected components for each functional trait.

Fig. S5: Pipeline diagram of our analysis composed of 5 distinct steps (for more details see Material & Methods): (1) Preprocessing step including read quality evaluation and filtering; (2) *De novo* assembly step in which assembled contigs were generated from cleaned reads with Trinity (ref. 57). (3) Quality evaluation of the previously assembled contigs. (4) Downstream analysis divided into two parts, with first detection of likely coding domains within contig sequences and then functional annotation of previously detected domains. (5) Construction of a sequence similarity network based on *de novo* assembly and downstream analysis results.

Fig. S6: A connected component outline. At the top, a multiple alignment of 3 protein coding domain sequences A, B and C. The 3 alignments respect sequence identity threshold (>60%) and sequence coverage threshold (>80%). At the bottom, a sketch of the corresponding connected component where protein coding domain sequences are represented by vertices and each alignment between two sequences is represented by and edge.

Table. S7: Table of all datasets used in this work with MMETSP IDs. Taxonomy, functional traits information and the presence in the SSN has been indicated in the table for each entry.

Fig. S8: Number of peptide sequences per proteome derived from “high quality” transcriptomes. Red line represents the minimum number of sequences threshold (9,000 peptide sequences) required to perform core/accessory/pan proteome analysis.

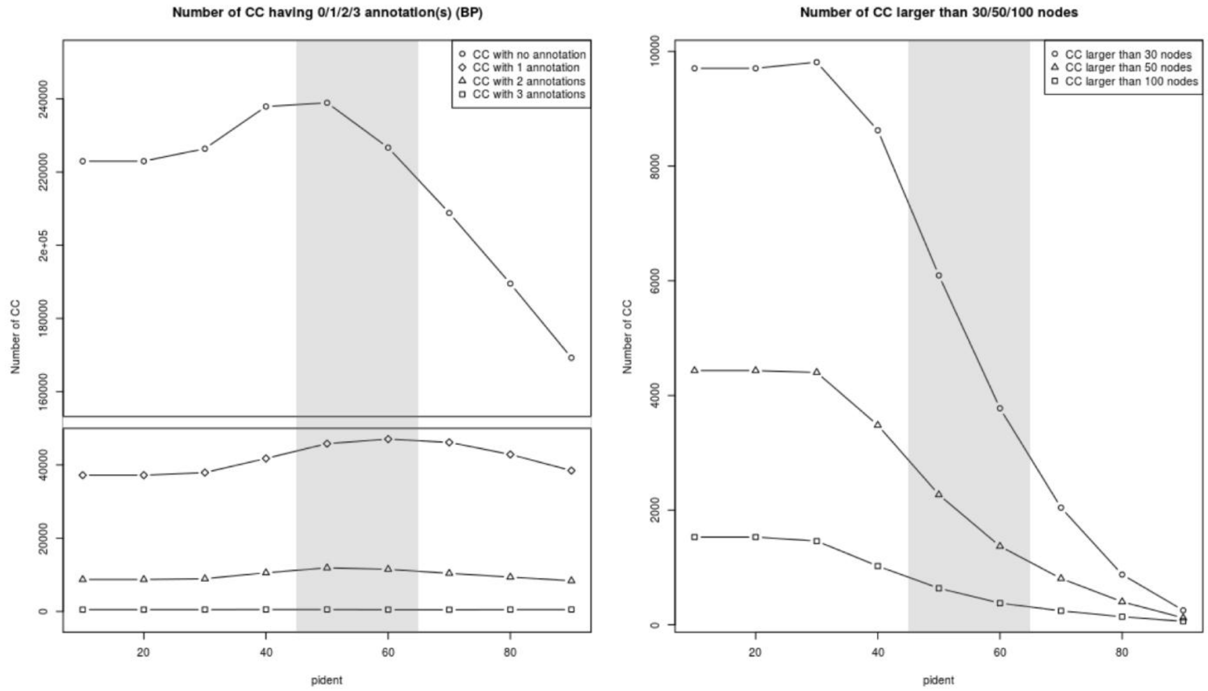


Fig. S 1: Optimal sequence identity threshold selection. The cutoff was chosen such that: (A) the network contains a maximum of connected components with homogeneous functional annotation (i.e. a unique GO Slim term for all annotated protein coding domains in each CC) and (B) the network conserved a maximum of « large » connected components.

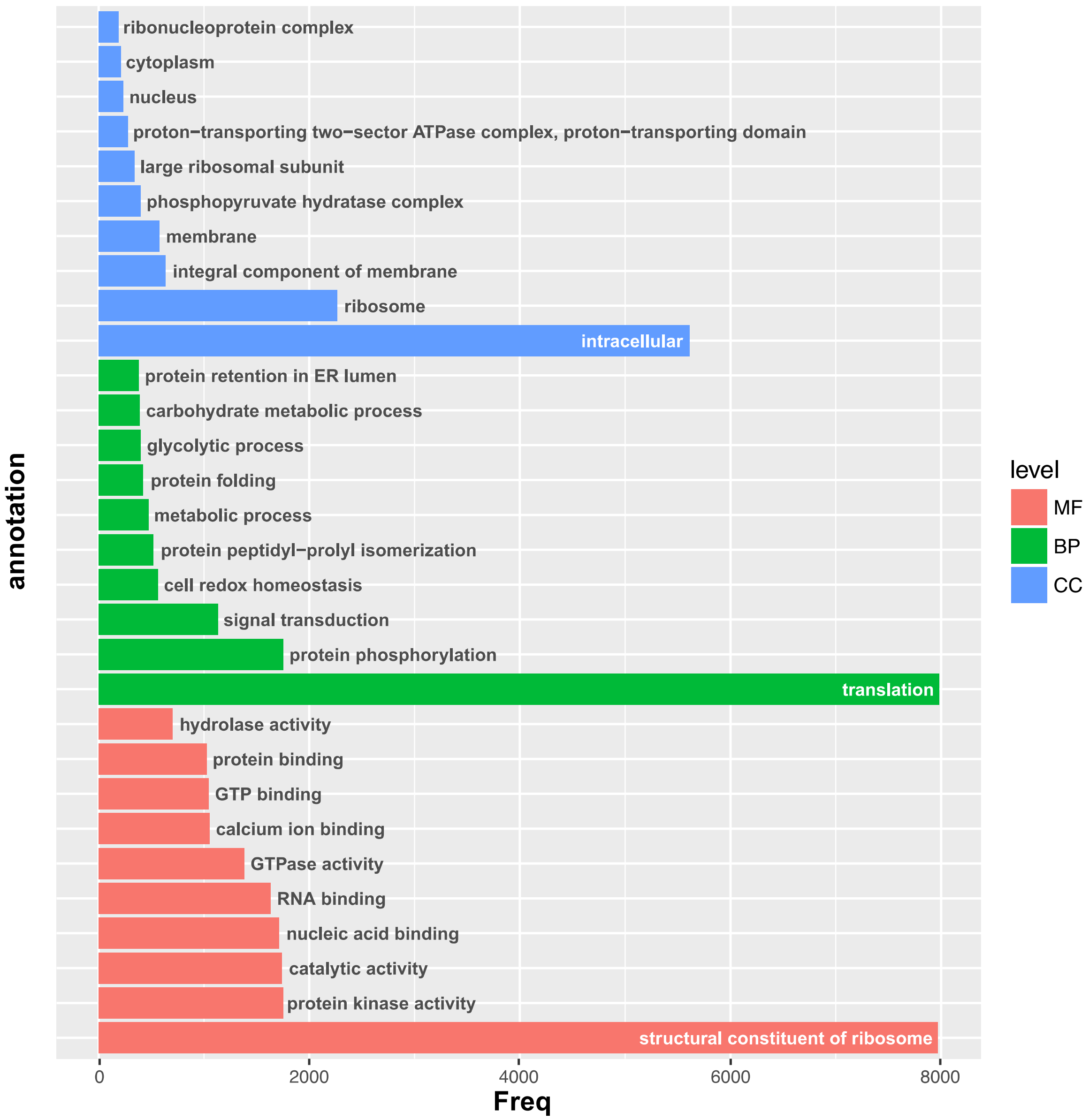


Fig. S 2 : Top 10 functional annotations (GOslim levels) in all core components. The three levels of annotation are represented: Molecular Function level (MF, red), Biological Process (BP, green) and Cellular Component (CC, blue).

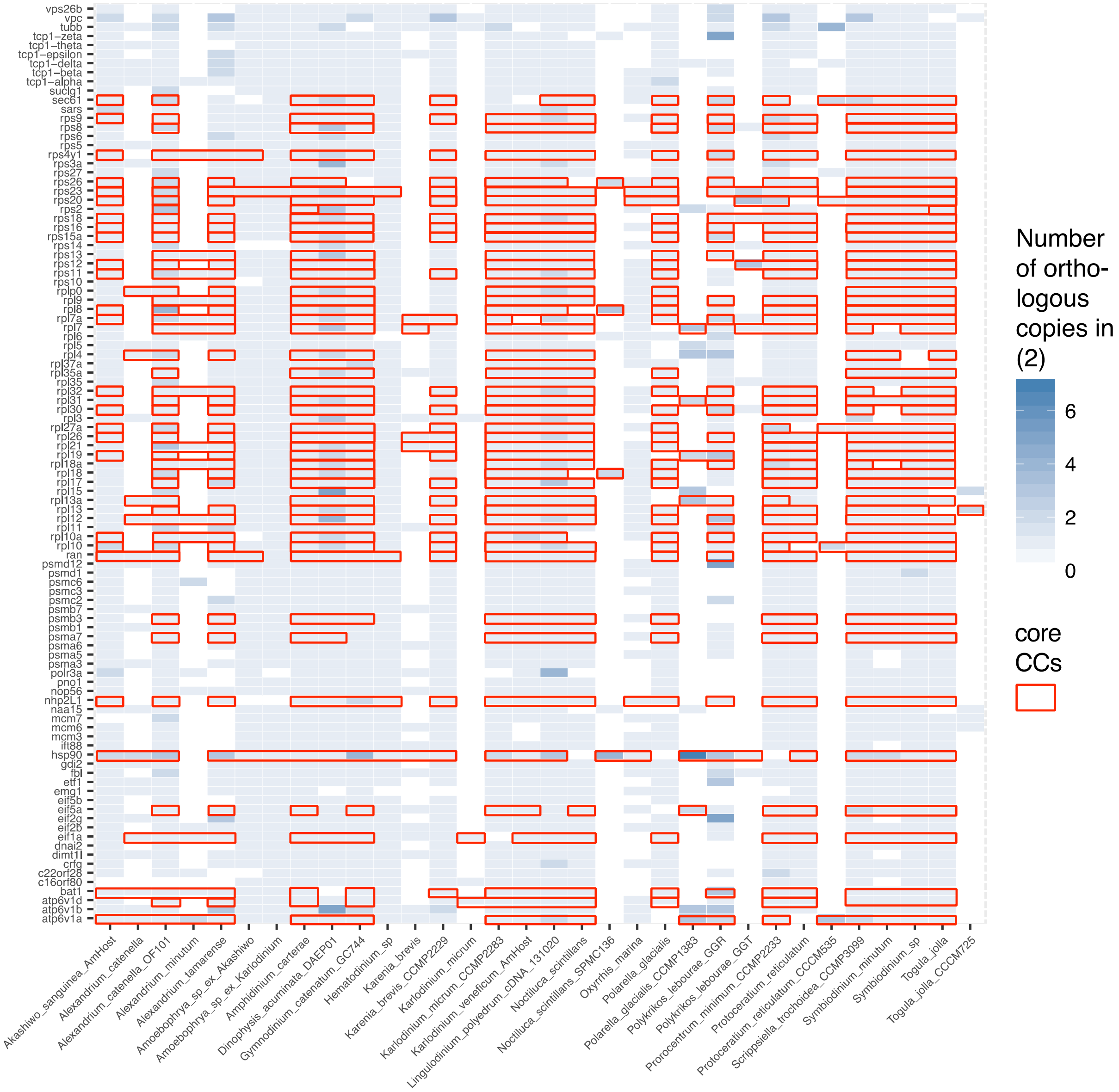


Fig. S 3 : Overlap between the 101 multiple protein alignments from (2) (used for phylogenomics) and our core dinoflagellate proteome. The number of copies of the 101 proteins is represented here in a heat map: the blue color gradient represents the number of orthologous sequences available of a particular protein (y-axis) for each studied dinoflagellate species in (2) (x-axis). The overlap between the data set from (2) and our core proteome is represented by the red boxes: i.e. each protein sequence from (2) that aligned to at least a core protein coding domain is here delimited in red.

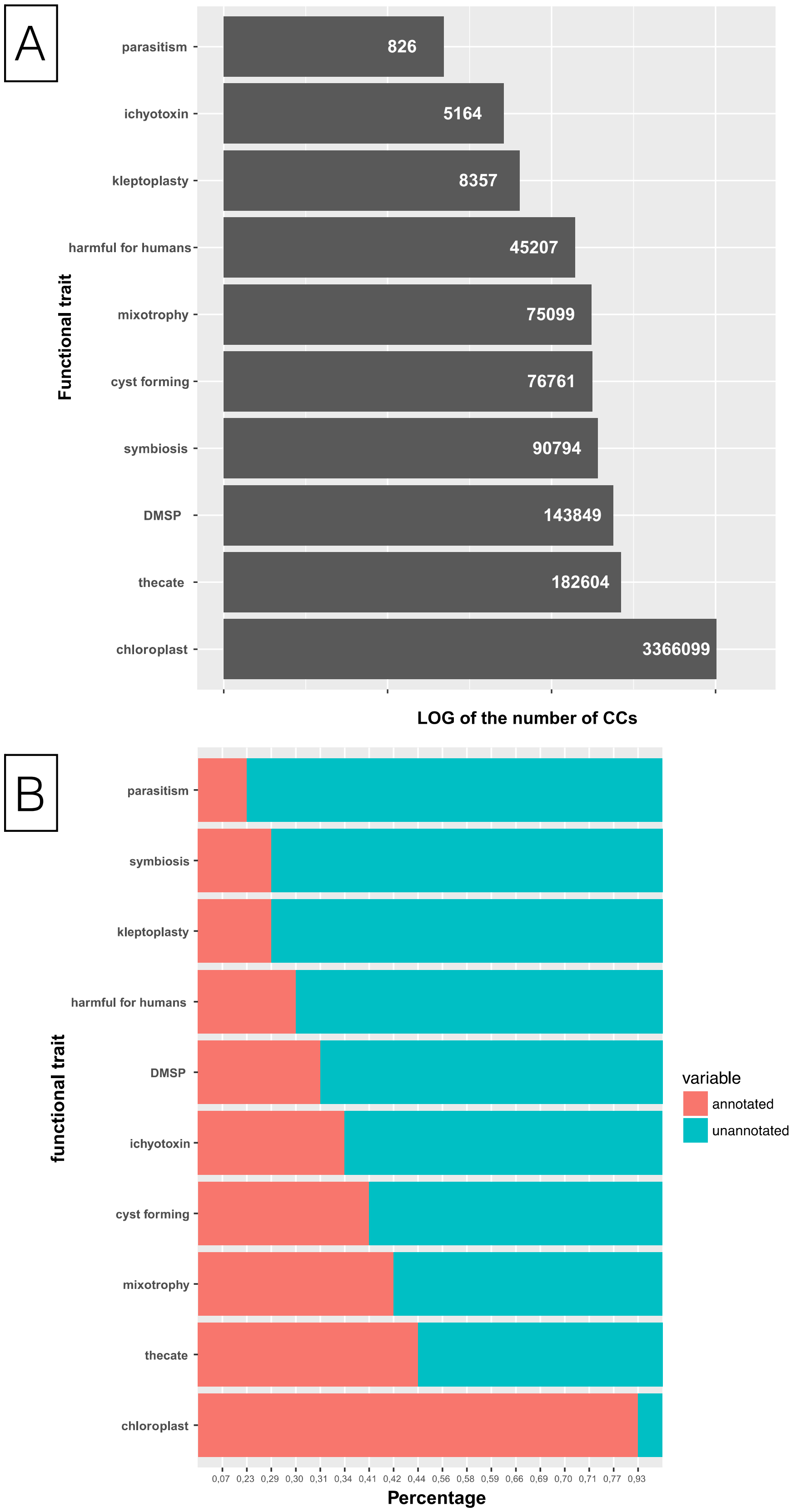


Fig. S 4: (A) Number of connected components for each functional trait. (B) Proportion of annotated sequences of connected components for each functional trait.

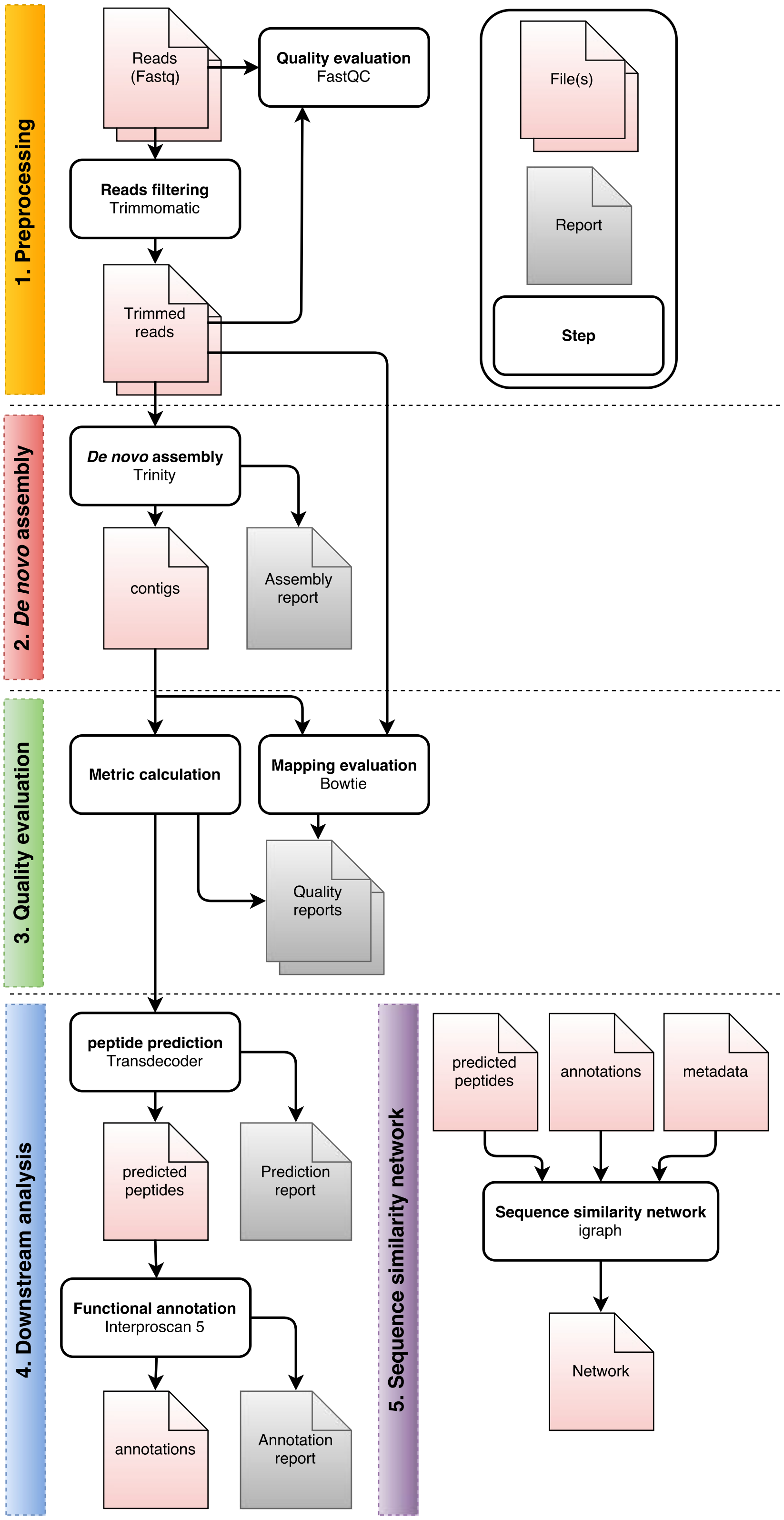


Fig. S 5: Pipeline diagram of our analysis composed of 5 distinct steps (for more details see Material & Methods): (1) Preprocessing step including read quality evaluation and filtering; (2) De novo assembly step in which assembled contigs were generated from cleaned reads with Trinity (ref. 57). (3) Quality evaluation of the previously assembled contigs. (4) Downstream analysis divided into two parts, with first detection of likely coding domains within contig sequences and then functional annotation of previously detected domains. (5) Construction of a sequence similarity network based on de novo assembly and downstream analysis results.

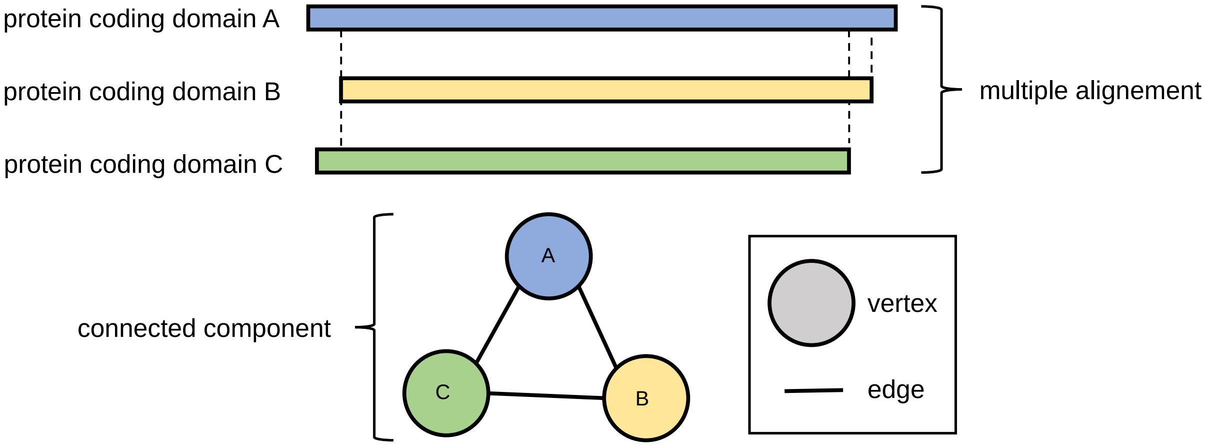


Fig. S 6: A connected component outline. At the top, a multiple alignment of 3 protein coding domain sequences A, B and C. The 3 alignments respect sequence identity threshold (>60%) and sequence coverage threshold (>80%). At the bottom, a sketch of the corresponding connected component where protein coding domain sequences are represented by vertices and each alignment between two sequences is represented by and edge.

*Fig. S 7: Table of the 60 datasets used in this study. It encompasses both MMETSP subsets with their ID and the 4 newly added datasets of our contribution (n°57-60 marked with *). We were not able to complete the assembly for n°22, 27 and 47. We defined low quality transcriptomes which show metrics that do not fit the quality threshold that we defined (see M&M): n°8, 16, 20, 26, 32, 37, 41, 42, 44, 56.*

ID	TAXONOMY								FUNCTIONAL TRAITS										
	order	family	genus	specie	strain	MMETSP ID	transcriptome quality	presence in SSN	chloroplast type	kleptoplasty	mixotrophy	toxicity potential	symbiont	ichyotoxicity	parasitism	DMSP	thecate	cyst forming
1	Dinophysiales	Dinophysiaceae	*Dinophysis*	*acuminata*	DAEP01	MMETSP0797	high	y	NC	y	y	DSP	n	n	n	y	y	n
2	Gonyaulacales	Ceratiaceae	*Ceratium*	*fusus*	PA161109	MMETSP1074 , MMETSP1075	high	y	P	n	y	n	n	n	n	n	n	n
3	Gonyaulacales	Crypthecodiniaceae	*Crypthecodinium*	*cohnii*	Seligo	MMETSP0323_2, MMETSP0324_2, MMETSP0325_2, MMETSP0326_2	high	y	R	n	n	n	n	n	n	y	y	n
4	Gonyualacales	Goniodomataceae	*Gambierdiscus*	*australes*	CAWD149	MMETSP0766_2	high	y	P	n	n	CFP	n	n	n	n	n	n
5	Gonyaulacales	Goniodomataceae	*Pyrodinium*	*bahamense*	pbaha01	MMETSP0796	high	y	P	n	n	PSP	n	n	n	n	n	y
6	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*andersonii*	CCMP2222	MMETSP1436	high	y	P	n	y	PSP	n	n	n	n	y	y
7	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*catenella*	OF101	MMETSP0790	high	y	P	n	y	PSP	n	n	n	n	y	y
8	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*fundyense*	CCMP1719	MMETSP0196, MMETSP0197, MMETSP0347	low	n	P	n	n	PSP	n	n	y	y	y	y
9	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*margalefi*	AMGDE01CS-322	MMETSP0661	high	y	P	n	n	n	n	n	n	n	y	y
10	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*minutum*	CCMP113	MMETSP0328	high	y	P	n	y	PSP	n	n	n	y	y	y
11	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*monilatum*	CCMP3105	MMETSP0093, MMETSP0095, MMETSP0096, MMETSP0097	high	y	P	n	y	PSP	n	n	n	n	y	y
12	Gonyaulacales	Gonyaulacaceae	*Alexandrium*	*tamarense*	CCMP1771	MMETSP0378, MMETSP0380, MMETSP0382, MMETSP0384	high	y	P	n	y	PSP	n	n	n	y	y	y
13	Gonyaulacales	Gonyaulacaceae	*Gonyaulax*	*spinifera*	CCMP409	MMETSP1439	high	y	P	n	y	n	n	n	n	y	y	y
14	Gonyaulacales	Gonyaulacaceae	*Lingulodinium*	*polyedra*	CCMP1738	MMETSP1032, MMETSP1033, MMETSP1034, MMETSP1035	high	y	P	n	y	DSP	n	n	n	y	y	y
15	Gonyaulacales	Gonyaulacaceae	*Protoceratium*	*reticulatum*	CCCM535=CCMP1889	MMETSP0228	high	y	P	n	n	DSP	n	n	n	n	n	y
16	Gymnodiniales	Gymnodiniaceae	*Akashiwo*	*sanguinea*	CCCM885	MMETSP0223_2	low	n	P	n	y	n	n	n	n	y	n	y
17	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*carterae*	CCMP1314	MMETSP0258, MMETSP0259, MMETSP0398, MMETSP0399	high	y	P	n	y	n	n	y	n	y	n	n
18	Gymnodiniales	Gymnodiniaceae	*Amphidinium*	*massartii*	CS-259	MMETSP0689_2	high	y	P	n	y	n	n	n	n	n	n	n
19	Gymnodiniales	Gymnodiniaceae	*Gymnodinium*	*catenatum*	GC744	MMETSP0784	high	y	P	n	y	PSP	n	n	n	n	n	y
20	Gymnodiniales	Gymnodiniaceae	*Gyrodinium*	*dominans*	SPMC103	MMETSP1148	low	n	R	n	n	n	n	n	n	n	n	n
21	Gymnodiniales	Gymnodiniaceae	*Togula*	*jolla*	CCCM725	MMETSP0224	high	y	P	n	NA	n	n	n	n	n	n	n
22	Gymnodiniales	Kareniaceae	*Karenia*	*brevis*	CCMP2229 Wilson SP3 SP1 Wilson	MMETSP0027, MMETSP0029, MMETSP0030, MMETSP0031 MMETSP0201, MMETSP0202 MMETSP0527_2, MMETSP0528_2 MMETSP0573, MMETSP0574 MMETSP0648_2, MMETSP0649_2	computational error	n	H	n	y	NSP	n	y	n	y	n	n
23	Gymnodiniales	Gymnodiniaceae	*Karlodinium*	*micrum*	CCMP2283	MMETSP1015, MMETSP1016, MMETSP1017	high	y	H	n	y	n	n	y	n	n	n	n
24	Noctilucales	Noctilucaceae	*Noctiluca*	*scintillans*	SPMC136	MMETSP0253	high	y	R	n	n	n	n	n	n	n	n	n
25	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	NA LB1974	MMETSP0468, MMETSP0469, MMETSP0470, MMETSP0471 MMETSP1424, MMETSP1425, MMETSP1426	high	y	R	n	n	n	n	n	n	n	n	y
26	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1788	MMETSP0044	low	n	R	n	n	n	n	n	n	n	n	y
27	Oxyrrhinales	Oxyrrhinaceae	*Oxyrrhis*	*marina*	CCMP1795	MMETSP0451_2, MMETSP0452_2	computational error	n	R	n	n	n	n	n	n	n	n	y
28	Peridiniales	Heterocapsaceae	*Heterocapsa*	*arctica*	CCMP445	MMETSP1441	high	y	P	n	NA	n	n	n	n	n	y	n
29	Peridiniales	Heterocapsaceae	*Heterocapsa*	*rotundata*	SCCAPK-0483	MMETSP0503	high	y	P	n	y	n	n	n	n	y	y	n
30	Peridiniales	Heterocapsaceae	*Heterocapsa*	*triquetra*	CCMP448	MMETSP0448	high	y	P	n	y	n	n	n	n	y	y	n
31	Peridiniales	*Amphidomataceae*	*Azadinium*	*spinosum*	3D9	MMETSP1036_2, MMETSP1037_2, MMETSP1038_2	high	y	P	n	NA	AZP	n	n	n	n	y	n
32	Peridiniales	Lessardiaceae	*Lessardia*	*elongata*	SPMC104	MMETSP1147	low	n	R	n	n	n	n	n	n	n	y	n
33	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3387	MMETSP1462	high	y	P	n	n	n	y	n	n	y	y	n
34	Peridiniales	Peridiniaceae	*Durinskia*	*baltica*	CSIRO_CS-38	MMETSP0116_2, MMETSP0117_2	high	y	D	n	NA	n	n	n	n	n	y	NA
35	Peridiniales	Peridiniaceae	*Glenodinium*	*foliaceum*	CCAP1116/3	MMETSP0118_2, MMETSP0119_2	high	y	P	n	NA	n	n	n	n	n	n	y
36	Peridiniales	Peridiniaceae	*Kryptoperidinium*	*foliaceum*	CCMP1326	MMETSP0120_2, MMETSP0121_2	high	y	D	n	NA	n	n	n	n	y	y	n
37	Peridiniales	Peridiniaceae	*Peridinium*	*aciculiferum*	PAER-2	MMETSP0370_2, MMETSP0371_2	low	n	P	n	NA	n	n	n	n	n	y	y
38	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei*	SHTV-5	MMETSP0359, MMETSP0360, MMETSP0361	high	y	P	n	n	n	n	n	n	n	y	y
39	Peridiniales	Peridiniaceae	*Scrippsiella*	*hangoei-like*	SHHI-4	MMETSP0367, MMETSP0368, MMETSP0369	low	n	P	n	n	n	n	n	n	n	y	y
40	Peridiniales	Peridiniaceae	*Scrippsiella*	*trochoidea*	CCMP3099	MMETSP0270, MMETSP0271, MMETSP0272	high	y	P	n	y	n	n	n	n	y	y	y
41	Prorocentrales	Prorocentraceae	*Prorocentrum*	*lima*	CCMP684	MMETSP0252	low	n	P	n	y	DSP	n	n	n	n	y	y
42	Prorocentrales	Prorocentraceae	*Prorocentrum*	*micans*	CCCM845	MMETSP0251_2	low	n	P	n	y	n	n	n	n	y	y	n
43	Prorocentrales	Prorocentraceae	*Prorocentrum*	*minimum*	CCMP1329 CCMP2233	MMETSP0053, MMETSP0055, MMETSP0056, MMETSP0057 MMETSP0267, MMETSP0268, MMETSP0269	low	y	P	n	y	DSP	n	n	n	n	y	n
44	Pyrocystales	Pyrocystaceae	*Pyrocystis*	*lunula*	CCCM517	MMETSP0229_2	low	n	P	n	y	n	n	n	n	y	n	n
45	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	MMETSP1338	high	y	P	n	n	n	y	n	n	y	n	n
46	Suessiales	Suessiaceae	*Polarella*	*glacialis*	CCMP1383	MMETSP0227	high	y	P	n	n	n	n	n	n	y	n	y
47	Suessiales	Symbiodiniaceae	*Symbiodinium*	*kawagutii*	CCMP2468	MMETSP0132_2, MMETSP0133_2, MMETSP0134_2, MMETSP0135_2	computational error	n	P	n	n	n	y	n	n	y	y	n
48	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	D1a	MMETSP1377	high	y	P	n	n	n	y	n	n	y	y	n
49	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP421	MMETSP1110	high	y	P	n	n	n	y	n	n	y	y	n
50	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C15	MMETSP1370, MMETSP1371	high	y	P	n	n	n	y	n	n	y	y	n
51	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	C1	MMETSP1367, MMETSP1369	high	y	P	n	n	n	y	n	n	y	y	n
52	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	CCMP2430	MMETSP1115, MMETSP1116, MMETSP1117	high	y	P	n	n	n	y	n	n	y	y	n
53	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	Mp	MMETSP1122, MMETSP1123, MMETSP1124, MMETSP1125	high	y	P	n	n	n	y	n	n	y	y	n
54	Suessiales	Symbiodiniaceae	*Symbiodinium*	*sp.*	cladeA	MMETSP1374	high	y	P	n	n	n	y	n	n	y	y	n
55	Syndiniales	Amoebophryaceae	*Amoebophrya*	*sp.*	Ameob2	MMETSP0795	high	y	R	n	n	n	n	n	y	n	n	n
56	Thoracosphaerales	Thoracosphaeraceae	*Thoracosphaera*	*heimii*	CCCM670=CCMP1069	MMETSP0225	low	n	P	n	n	n	n	n	n	y	n	n
57	Gymnodiniales	Gymnodiniaceae	*Gymnoxanthella*	*radiolariae*	RCC3507	/	high	y	P	n	n	n	y	n	n	y	n	n
58	Peridiniales	Heterocapsaceae	*Heterocapsa*	*sp.*	RCC1516	/	high	y	P	n	y	n	n	n	n	y	y	n
59	Peridiniales	Peridiniaceae	*Brandtodinium*	*nutricula*	RCC3468	/	high	y	P	n	n	n	y	n	n	y	y	n
60	Suessiales	Suessiaceae	*Pelagodinium*	*beii*	RCC1491	/	high	y	P	n	n	n	y	n	n	y	n	n

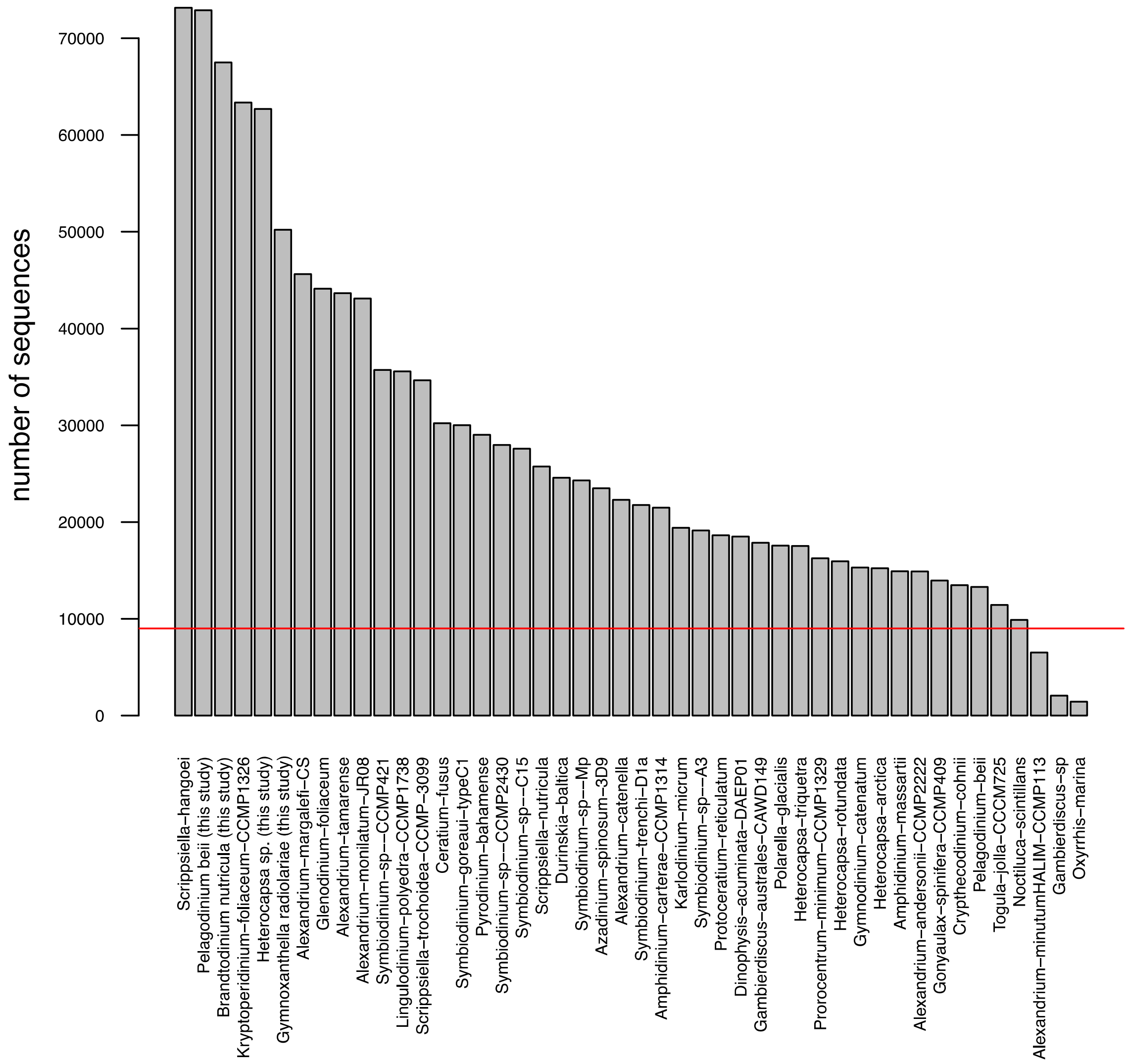


Fig. S 8: Number of peptide sequences per proteome derived from "high quality" transcriptomes. Red line represents the minimum number of sequences threshold (9,000 peptide sequences) required to perform core/accessory/pan proteome analysis.

SI Appendix Tab. S5 : kleptoplatic trait-CCs

Number of CC composed of kleptoplastic species sequences	Unannotated
6 995	4 493

Number of CC composed	
N	# CC
1	3 255