An "omic" approach to *Pyrocystis lunula*: New insights related with this bioluminescent dinoflagellate

Carlos Fajardo

Universidad de Cadiz

Francisco Amil-Ruiz

Universidad de Cordoba <https://orcid.org/0000-0002-3470-0860>

Carlos Fuentes Almagro

Universidad de Cordoba

Marcos De Donato

Instituto Tecnológico y de Estudios Superiores de Monterrey

Almudena Escobar Niño

Universidad de Cadiz

Rafael Carrasco

Universidad de Cadiz

Juan Miguel Mancera

Universidad de Cadiz

Francisco Javier Fernandez Acero (✉ franciscojavier.fernandez@uca.es)

Universidad de Cádiz <https://orcid.org/0000-0002-4052-6141>

Research article

Keywords: *P. lunula*, transcriptomics, proteomics, luciferin-binding protein, glutathione-S-transferase, bioluminescence

Posted Date: May 14th, 2019

DOI: <https://doi.org/10.21203/rs.2.9613/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Version of Record: A version of this preprint was published at Journal of Proteomics on October 1st, 2019. See the published version at <https://doi.org/10.1016/j.jprot.2019.103502>.

Abstract

Background: *Pyrocystis lunula* (Schutt) is a photoautotrophic dinoflagellate without armored, frequently found in marine environments. Today, there are several biotechnological applications derived from the bioluminescent system of this species. From a post-genomic perspective, in order to study the whole proteome of *P. lunula*, an "omic" approach (transcriptomic-proteomic analysis) was assessed using fresh microalgae samples. Results: A total of 80,874,825 raw reads were generated (11,292,087,505 pb; 55.82 % GC) by mRNA sequencing. Very high quality sequences were assembled into 414,295 contigs (219,203,407 pb; 55.38 % GC) by the Trinity software, generating a comprehensive reference transcriptome for this species. Then, a *P. lunula* proteome was predicted and further employed through the first proteomic analysis on this species. A total of 17,461 peptides were identified, yielding 3,182 protein identification hits. The identified proteins were further categorized according to functional description and gene ontology classification. Conclusions: The first comprehensive molecular analysis of the microalgae *P. lunula* using both, transcriptomic and proteomic approaches, has been reported. Proteomic results represent a valuable piece in the understanding of this microalgae regulation at molecular level, and shed light on the identification of important factors involved in gene expression regulation. Indeed, the presence of the luciferin-binding protein (LBP), which had not been described so far in the genus *Pyrocystis* has been highlighted.

Background

The dinoflagellates belong to a phylum of unicellular eukaryotes that live in marine and freshwater. In the ocean, they are the most important primary producers, only exceeded by diatoms, and have received a lot of attention, mostly because they are a source of valuable bioactive compounds. Additionally, dinoflagellates are related to various marine phenomena, such as bioluminescence, symbiosis and harmful algal blooms [1-7]. *Pyrocystis* is a marine photoautotrophic genus of the Dinophyceae, which has received attention as a model organism due to its bioluminescence capacity according to circadian rhythms [8-12]. Morphologically, *Pyrocystis* is particular among the Dinophyceae as it spends a considerable part of its life cycle as a non-mobile cell within a cellulose-covered shell [13-17].

In this genus, *Pyrocystis lunula* (Schutt) is an important species because there are several biotechnological applications derived from its bioluminescent system. The bioluminescence of *P. lunula* can be quantified by spectrophotometry as the normal bioluminescent output declining with increasing concentrations of aqueous toxicants. Examples of these current applications are the Lumitox and the QwikLite™ 200 bioassays, two portable test kits designed to measure toxicity of pollutants in marine environments. More recently, a bioassay to assess potential instances of groundwater contamination linked with natural gas extraction and unconventional drilling activity, has been developed. [18-20].

New perspectives derived from the study of bioluminescent systems and reactions have led to many biotechnology applications. A wide set of *in vivo* and *in vitro* tools have been developed based on different bioluminescent systems, including bio-imaging, immunoassays, gene expression assays, drug

screening, tests for different analytes, as well as the environmental monitoring [21]. New possibilities of additional applications are promising with the invention of luminescent reporters [22]. It is possible to use different types of luciferases (LCFs), for example dinoflagellate LCF coupled to the firefly or renilla LCFs, for dual and multiple reporter assays. Thus, this kind of assay can identify transcriptional activities of different genes at the same time [23]. Furthermore, these capacities can be improved through modifications of enzymes and substrates, originating from different bioluminescence organisms. Recently, it has been possible engineering a new bioluminescence system capable of more efficient light emission [24]. On the other hand, as a function of mechanical stimulation in the natural environment, the bioluminescent could potentially be used to track military targets [25 , 26]. Bioluminescence trends in science develop very quickly and the worldwide profits of these technologies are around \$5 billion/year [21].

A measurement of the DNA content of *P. lunula* has shown that the DNA synthesis is low. *P. lunula* vegetative cells are diploid and possess 2 of each of the 49 chromosomes. Although the genome size of *P. lunula* has not been determined yet, dinoflagellates present enormous genomes ranging from 2 to 200 pg/nucleus of DNA (about 1.5-245 Gbp) [16 , 17], representing 1 to 80 times more than a human haploid genome, and for this reason it has been difficult to sequence the complete genome of different members of this lineage [27], resulting in an additional limiting factor to carry out proteomic studies in these organisms. Nowadays, have been reported only a complete and a partial genome for *Symbiodinium kawagutii* and *S. minutum*. These two species have the smallest dinoflagellate genomes identified so far [28-30]. The real challenge relies on bioinformatic analysis to assemble reads into genes or chromosomes, and this is particular hard to achieve because dinoflagellate genomes are highly redundant [31].

Despite the development of proteomics platforms, not many approaches have been developed to study microalgae proteomes [6 , 32-54]. The main aim of most of them was oriented to improve biodiesel production, as well as other relevant aspects of their biology. These studies have used a variety of proteomic strategies, from the MALDI TOF/TOF analysis to the qualitative and quantitative mass spectrometry (LC-MS/MS) approaches [55 , 56]. Recently, new advances in the technology of massive sequencing have allowed the generation of transcriptomic databases that in the past would have been extremely difficult to achieve, not only in terms of economic expenses, but also of time spent. Thus, the quantity of transcriptomic data produced has grown very fast and can now be easily extend to a genome wide transcriptome [31]. These efforts are likely to provide valuable information for our understanding of the biology of this important group, as well as potentiate the impact of their use in biotechnological applications.

The synergistic combination between this tool and the modeling of the proteome by LC-MS/MS will allow obtaining a wealth of information that, to date is not available in the case of *P. lunula*. It is clear that *omics* provide information with important implications for systems biology approaches [57], and the experience has shown that such data can transform our understanding of the basic biology, at the same time that opens the door to new research directions. Today, the limited genomic information available is a

significant obstacle to develop global studies at the molecular level in the case of *P. lunula*. Driven by this challenge, our main objective was to generate a reference transcriptome and proteome that will support and complement future research lines in *P. lunula*.

Results

De novo transcriptome assembly

A total of 80,874,825 raw reads were generated (11,292,087,505 pb, 55.82 % GC, reads length media of 140 pb). The percentage of residual ribosomal RNA remaining in the sample after the enrichment step on messenger RNA was quantified as detailed in the material and methods section. Identified ribosomal RNAs were then categorized in an attempt to identify potential contamination of putative prokaryotes present in the sample. Thus, from the 11.39 % reads passing the threshold in the rRNA search (9,212,736 reads), 9.16 % and 1.87 % matched eukaryotic ribosomal 18s and 28s, respectively, while only 0.24 % and 0.12 % matched bacterial 16s and 23s, respectively. The raw reads generated for the *de novo* assembly were deposited at the NCBI SRA database, BioProject accession number PRJNA497104.

A total of 25 assemblies were generated and further assessed to select the most suited algorithm and parametrization for a successful *P. lunula* transcriptome assembly. Thus, an analysis of the assemblies structure, for contigs greater than 400 nt, was carried out. Attending to the three algorithms computational requirements [58], RAY was first employed to analyze the effect of setting kmer parameter, library size and reads quality (in terms of minimum read size and mean phred scores) in the algorithm performance. Results clearly show that increasing the kmer parameter have a tremendously deleterious effect on final number of contigs obtained, although, as it was expected, contigs' size distribution improves with higher kmer values, specially showing higher Nx for kmer over 41 (Figure 1). However, almost identical structure was obtained for all datasets when the kmer parameter was set within 21-41, suggesting that lower kmer values (within 21-31) could be appropriate in this case, avoiding an important loss of new contigs. Indeed, when each kmer combination is explored in detail to evaluate the effect of library size and reads quality filters, a slightly better structure is obtained when the Q20 (highest quality filter) is applied. With respect to the reads' length filters (L100 and L30), as both seems to behave similarly for the assemblies structure, we decide to further use the more strict filtered library (L100), because this is actually the more confident sequence collection (more strict quality and length filters applied). Moreover, its reduced size fits well within the Trinity and Mira pipelines, as these two algorithms present much higher computational requirements than Ray for large datasets [58]. Finally, comparing all 10 assemblies obtained from the Q20 L100 library (Figure 2), that one obtained with Trinity software, and k25_NC20 parameters, presented the best equilibrium in terms of structure and total number of contigs.

Then, to evaluate the representativeness of each assembly, a BLASTn search was carried out for each sequence in each assembly against each sequence in each assembly (all-against-all reciprocal search for similarity), including two additional datasets downloaded from repositories (*P. lunula*, strain CCCM 517,

<https://www.imicrobe.us/>; *S. minutum* [28]). A total of 3,645 BLASTn searches (27x27 assemblies at 5 e-value thresholds) were employed to retrieve the very best hit passing a specific e-value threshold [see Additional file 1]. The percentage of sequences of one assembly that find a hit in each other assembly was determined (Figure 3). To facilitate visualization, only results from reciprocal comparisons of Trinity_Q20 L100_k25_NC20 to all 27 assemblies were plotted. Results clearly indicate that while almost all sequences from most assemblies presented a hit, meaning that are also present in the Trinity evaluated assembly (red lines), the other assemblies clearly miss many of the contigs that are presented in the evaluated one (blue lines). Considering the two external datasets, results manifest that the *P. lunula* strain CCCM 517 dataset is very poor (only 9 sequences longest than 400 nt), and that the similarity at the nucleotide level of *P. lunula* with *S. minutum* is scarce.

A final evaluation of the generated assemblies was carried out by proteomics. Thus, inferred proteomes were employed as references in a proteomic assay for *P. lunula* protein samples. Searching the obtained spectra against Trinity_Q20 L100_k25_NC20 and Trinity_Q20 L100_k31_NC20 databases yielded the highest number of identified proteins. These results, together with representativeness and structure analyses commented above propose the Trinity_Q20 L100_k25_NC20 assembly (414,295 contigs (185,177 > 400 bases); 219,203,407 pb; 55.38 % GC) as the best suited to the identification profiles obtained by MS analysis [see Additional file 2].

Proteomic analysis

To infer a fine representative proteome from the transcript sequences assembled in Trinity_Q20 L100_k25_NC20, a BLASTx was performed against the “nr” database, which represents all the public, non-redundant, protein sequences. As a result, matching hits were found for around 100,000 sequences (threshold: e-value < e-15). This information was very useful for both, performing further functional annotation of these matching transcripts, as well as, to identify their reading frame, allowing the identification of a set of coding sequences of *P. lunula*. For the remaining 300,000 sequences that do not showed clear homology, all the possible ORFs of the 6 reading frames were identified, and the protein sequence encoded by the longer ORF was retained, together with the previously identified set of coding sequences, to be later searched within the proteomic assay. Although some of this newly inferred protein sequences could be artefactual (consequence of non-coding RNAs translations, or assembling/sequencing artifacts) introducing noise into the proteome database, we considered of great relevance analyzing putative novel proteins that could be present here that do not yet exist in public databases. This could be of special importance in species clearly underrepresented in databases, as occurs with the lack of molecular studies on this entire genus.

All the sequences generated in both groups, with or without homology against “nr”, as long as they were greater than 50 aa, were preserved, and the reference proteome was created. The Proteome Discoverer 2.1 software, with Sequest HT algorithm, was used to search all the spectra that were obtained in the protein sample in the MS/MS analysis (17,461 peptides) against that predicted proteome, and with a threshold of FDR < 0.01, 3,182 identifications were obtained. The bidimensional distribution of the identified

proteins thorough the proteomic assay within the total proteins present in the predicted proteome, according to their MW and IP was studied (Figure 4). The mass spectrometry data was deposited in the ProteomeXchange Consortium through the PRIDE [59] partner repository, dataset identifier PXD01155.

Functional annotation

Using BLAST2GO software, all the identified proteins were functionally annotated according to the GO categories, enzyme codes and InterPro domains. Thus, Figures 5 and 6 shows the main functions identified within the identified proteins according to the GO biological process (BP) and molecular function (MF) categories, respectively. A complete list of identified proteins, together with all de functional information associated is reported in the Additional file 3. Regarding the biological process category (BP), the first place is occupied by proteins related to oxide-reduction processes, followed by proteins related to the translation process and biogenesis of ribosomes. At the level of molecular function (MF), ATP binding molecules were found to be most abundant among the proteome followed by structural constituent of ribosome, protein and metal ion binding. All the enzymatic codes associated were further located into the KEGG database, using the KEGG-mapper tool, to identify the pathways in which the identified proteins could be acting [see Additional file 4]. Attending to the species contributing to the functional assignments, *S. microadriaticum* was the top-hit species, followed by *Vitrella brassicaformis*, *Perkinsus marinus* and *Heterocapsa triquetra* (Figure 7). Although this analysis may undergo a bias by the abundance of sequences from each species present within the public databases, our results seems consistent with the evolutionary relationship between Dinoflagellata, Chromerida, and Perkinzoa (alveolates) [9 , 12 , 31].

Discussion

Novel proteins and proteome mining

From the 3,182 proteins detected in *P. lunula*, 175 (5.49 %) turn out to be sequences that do not have any homology with any protein identified so far in the different databases. Within this group, 59 stand out (PSM $\geq$ 5) and should be considered with a high level of confidence as novel proteins [see Additional files 5 and 2]. This is not unexpected since the transcriptomics data sets of dinoflagellates has showed, among other things, that in some cases more than 50 % of the genes do not have homology with any gene documented so far [31]. Many genes appear to be unique in dinoflagellates. For example, although many of the redox-regulated genes identified in *P. lunula* are conserved, the majority, do not shown significant similarities with other genes in the databases as around 68 % were novel [9].

Evolutionarily, dinoflagellate nuclear genomes are very dynamic, because they have obtained plastid targeted genes through successive horizontal gene transfer from different kinds of sources (tertiary replacement plastid, peridinin plastid, bacteria, cyanobacteria, haptophyte, red and green algae), rising to a highly chimeric nuclear genome [60-63]. Furthermore, segmental genome duplications [64 , 65], and retro-transposition of cDNA to the genome [66], are likely contribute to the redundant of genes and expansion of these genomes [28]. Due to their immense proportions, only a few dinoflagellate genomes have been sequenced and the information of the genomic structure has mainly derived from the study of

individual genes. Tubulin, actin, polyketide synthase, rubisco, heat shock proteins, chlorophyll-binding protein, proliferating cell nuclear antigen, LCF, and luciferin-binding protein (LBP), are among those that have been investigated [31]. Interestingly, we have detected all these mentioned proteins within our study, and particularly relevant is the identification of LBP [see Additional file 3], which was thought absent from the bioluminescent system in the genus *Pyrocystis*.

Luciferin-binding protein (LBP)

The bioluminescence system in dinoflagellates is unique, from a cellular and molecular perspective. The production of light occurs in organelles named scintillons [67], which contain the substrate luciferin, the LCF enzyme, and in the most of the cases a LBP [68-71]. The light emission depends on LCF-catalyzed oxidation of the luciferin, generally protected from oxidation by a LBP that binds luciferin at physiological pH. However, this LBP has not received much attention as LCF, maybe because it has not been identified in all bioluminescent dinoflagellates, and therefore has not been considered to date as an essential component of this bioluminescence system [72]. *Lingulodinium polyedrum* (formerly *Gonyaulax polyedra*), the first organism from which this protein was isolated, cloned and fully sequenced, is the main model organism for studies of LBP [73-75]. In *L. polyedrum*, the LBP gene present two different types or isoforms (LBP*a* and LBP*b*) which share 86 % sequence identity and encode two protein that are expressed in equal amounts [74 , 75]. Additionally, each gene type has the typical dinoflagellate genomic organization and is presented in several non-identical tandem copies. Furthermore, in the *L. polyedrum* proteome, LBP was found to be very abundant, up to 1 % of the total proteins [75 , 76].

The LBP is well known in *L. polyedrum* [68 , 77], *Noctiluca scintillans* [78], and it has also been found in the genera *Gonyaulax*, *Ceratocorys*, *Protoceratium*, and *Alexandrium* [79-82]; therefore, these genes seems to be a fundamental component of the molecular bioluminescence system in dinoflagellates. Molecular studies have demonstrated a high variation among gene copies, revealing a very diverse gene family comprising multiple gene types in some cases [72]. In fact, *Pyrocystis* LBP have not been reported in protein extracts screened by a universal antibody for LBP [68 , 70]. This could not be taken as a conclusive proof of the absence of the LBP protein, given the high degree of variability reported for this gene. Currently is unknown how many of the bioluminescent dinoflagellate species utilize LBP in their bioluminescence system [11]; nevertheless, emerging information shows substantial evidence that LBP is an important component of dinoflagellate bioluminescence system [72].

It was reported that dinoflagellates LBP (DinoLBP) primers yielded PCR products from eleven species out of the eighteen tested, demonstrated the present of LBP in all members of the *Gonyaulacales* with the exception of the genera *Ceratium* and *Fragilidium*. It is important to note that, due the differences of LBP between species, the efficient amplification of LBP from *Ceratocorys horrida*, *Gonyaulax spinifer*, and *Alexandrium monilatum* only were possible after a reduction in the PCR stringency. Furthermore, the DinoLBP primers amplified LBP in all the species known to contain it, except *N. scintillans*, which present a highly divergent LBP sequence [78]. Published research have shown the diversity of LBP in several closely related dinoflagellates, with first reports in *P. reticulatum*, *C. horrida*, and *G. spinifera*, as well as

the first partial sequences from *Alexandrium affine*, *A. monilatum*, *A. tamarense*, and *A. fundyense* [72]. These results also agree with an absence of LBP in *Pyrocystis spp.*; however, the lack of detection in *N. scintillans* suggest that the negative PCR result in *P. lunula*, *Protoperdinium crassipes*, *Ceratium longipes*, and *Fragilidium cf. subglobosum*, are unlikely to be conclusive. Despite many efforts, the differences in LBP between species has precluded the design of universal primers for LBP, which has previously done in the case of LCF [83]. LBP sequences have shown to be a very diverse gene family [75 , 76]. Furthermore, phylogenetic studies on *Alexandrium spp.* also suggest the presence of more than one LBP type, similar to previous observations with their LCF sequences [83]. Multiple gene types in genes involved in bioluminescence seem to be common in both LBP and LCF for several species [75 , 83 , 84], creating more divergence between gene copies of LBP in these genomes.

Studies in *N. scintillans* also revealed that LBP is present in diverse forms, and as the result of important evolutionary events like gene fission or fusion. *N. scintillans* is unusual as LCF and LBP are found as two domains in one gene, whereas they are normally separate genes [72]. It has been suggested that the origin of the hybrid structure LCF/LBP was either by fusion of the two genes of photosynthetic species in *N. scintillans*, or by fission of the *N. scintillans* gene into those species. The hybrid LCF/LBP gene (2,396 bp) reported in *N. scintillans* [72 , 78] consist of part of the N-terminal region and the LCF domain followed by the LBP domain. A detailed analysis of our results has revealed a similar situation, being able to detect two different isoforms of LBP, one version that seems to correspond to the individual gene LBP, very similar (query length 422, query cover 95 %, ident. 75 %) to that reported for *A. tamarense* (GenBank AFN27006.1) [72] (Fig. 8) [see Additional file 6], and another that would represent the hybrid version between LCF/LBP, most similar to the sequences reported for *L. polyedra* (GenBank AAA29164.1 and AAA29163.1) (query length 1,116; query cover 96 %; ident. 55 %)[75] and *A. cantenella* (GenBank: ABY78836.1) (query length 1,116; query cover 95 %; ident. 54 %)[81] (Fig. 9) [see Additional file 7]. These results were corroborated by PCR, sequencing and BLASTx analyses [85]. In fact, a similar situation occurs in the case of the LCF gene in *P. lunula*, where up to three isoforms have been reported: LCFa (GenBank AF394059.1), LCFb (GenBank AF394060.1), and LCFc (GenBank AF394061.1) [84]. The characteristic element of the hybrid LCF/LBP is the presence of the LCF/LBP N-terminal domain (pfam05295, Luciferase_N), however, this part is not the catalytic domain of the protein. It has been suggested that this region could mediate an interaction between LBP and LCF or their relationship with the vacuolar membrane [72].

Additionally, in *N. scintillans*, both fragments present an identical N-terminal region of 108 bp. A study of the two cDNA sequences showed that in the shorter fragment, the N-terminal region led directly into the LBP domain in the absence of an LCF domain. Therefore, the combined LCF/LBP and the single separated LBP, which shared the N-terminal region and LBP domain, but not the LCF domain, corresponded to two different genes. It was also confirmed that *N. scintillans* did indeed express both of these genes [72]. It has been speculated that the second LBP in *N. scintillans* could have a role in binding luciferin in the scintillons or it could act to store it in the cytoplasm [78]. It has been suggested that if the LCF/LBP of *N. scintillans* is ancestral to the LCF and LBP of photosynthetic species [86], the single LBP of *N. scintillans* could have been formed by mRNA splicing of LCF/LBP and subsequent retro-

transposition on the genome [66 , 87]. The N-terminal region in both genes seems to support this hypothesis. Having separate LCF and LBP could enable differential regulation of each gene and this can be advantageous if LBP needs to be stoichiometrically proportional to luciferin [77], but LCF, which has a triple catalytic capacity in photosynthetic species, can be re-used [72].

Glutathione-S-transferase (GST) and the elusive luciferin

In our results, oxide-reduction processes are invoked as a key component, as evidenced by the GO and KO analyzes [see Additional files 3 and 4]. Oxidative stress is able to induce a transcriptional response. In *L. polyedrum*, the oxidative stress induced by metal increased the activity of the superoxide dismutase [88], being dependent on the type of metal, its time of exposure and concentration [89 , 90]. Furthermore, a study with 3,500 genes from *P. lunula* showed that up to 200 genes increased in abundance after treatment with sodium nitrite (1 mM) or paraquat (0.5 mM). It is also important to note that antioxidant enzyme GST (GenBank AAN85429.1) [9], whose presence has been detected in our results [see Additional file 3], also possesses the pfam05295 domain (Luciferase_N). In addition to the pfam05295 domain (N-terminal LCF/LBP), a GST-N-Sigma-like domain is also present, which belongs to the Thioredoxin-like superfamily. These proteins function as disulfide oxidoreductases (PDOs), altering the redox state of other proteins through the reversible oxidation of their dithiol active site. The thiol group of cysteine, on the reduced state, is able to donate a reduction equivalent (H^+ , $+e^-$) to other unstable molecules [9], such as reactive oxygen species like luciferin.

In *P. lunula*, the gene of the antioxidant enzyme GST has been shown to have an additional 255 bp extension at its end [9], encoding an N-terminal sequence previously reported in LCF and LBP from *L. polyedrum* [73 , 91]. The similarity between the N-termini of GST, LCF and LBP is around 45 %, compared with 50 % between *L. polyedrum* LCF and LBP [92]; exon recombination was suggested as a plausible explanation for this homology [93]. Although the function of this N-terminal region is still unknown, the presence of this conserved sequence in GST could indicate that its function is not restricted to bioluminescence and maybe related to other roles, like cellular processing or localization of these proteins. It is also important to note that in *P. lunula*, unlike other bioluminescent dinoflagellates, the amounts of luciferin and LCF are constant throughout the day and night [70]. Therefore, instead of daily *de novo* synthesis and destruction of all the components of the bioluminescent system, the rhythm involves changes in their intracellular localization [8 , 84], and it is likely that the GST protein may be involved in this process, however further studies are necessary in order to test this hypothesis.

The luciferin from *P. lunula* is a tetra pyrrole-type molecule, resembling to chlorophyll a and krill luciferin [94]. This luciferin presents photo-oxidation and it is extremely labile to O_2 , at high salt concentration, and low pH [95]. *P. lunula* contains luciferin in larger amounts than other dinoflagellate species, even 100 times more than *L. polyedrum* [94], and this luciferin it seems to be universal because LCF from any dinoflagellate can use it as a substrate to produce light [96]. Surprisingly, this luciferin could even cross-react with the bioluminescent system of the krill (*Euphasia superba*) [68 , 97]. It has been hypothesized that *P. lunula* luciferin is a photo-oxidation breakdown product of chlorophyll a [98]. According to this

paradigm, Liu and Hastings (2007) [78] hypothesized that heterotrophic dinoflagellates acquire luciferin from ingested prey, either directly or from the degradation of chlorophyll provide by the prey. However, the heterotrophic dinoflagellate *P. crassipes* can maintain its bioluminescence up to 1 year without the ingestion of chlorophyll or luciferin containing food [99], and therefore must contain a luciferin originating from a different precursor molecule. The presence of LBP in several photosynthetic species suggests that even within these species, there might be an alternative luciferin molecule that requires LBP for stabilization [74]. In any case, the hypothesis that the origin of luciferin is photo-oxidized chlorophyll a [98] would only be plausible for *P. lunula*, which maintains its luciferin throughout the light-dark cycle. *L. polyedrum* only contains luciferin at night when photo-oxidation is not possible [71], so its synthesis cannot be explained by the photo-oxidation mechanism. Therefore, it is likely that more than one mechanism is responsible of luciferin production even in closed related species [11].

In fact, previous reports has confirmed the intracellular production of luciferin in *P. lunula* [100]. In this regard, Fresneau and Arrio (1988) [101], argued that bioluminescence of *P. lunula* is controlled by the reduction state of the luciferin precursor. It was reported that that luciferin and its precursor P630, so called by his excitation wavelength peak (630 nm), possess the same peptide moiety. P630 is a chromopeptide more stable than luciferin at low temperature in methanolic solutions, and it is composed by a polypeptide (4.8 kDa) and a linear tetrapyrrole such as luciferin (600 Da). Cations could oxidize P630 or break the bond between the peptidic chain and the extended tetrapyrrole. Reduction of P630 is performed enzymatically by a NAD(P)H-dependent oxidoreductase or chemically by 2-mercaptoethanol or dithiothreitol. Reduced P630 has the same spectral characteristics as the purified luciferin, as LCF can oxidize this reduced molecule with a light emission at 480 nm. It is very important to note that purified luciferin spontaneously turned partly into P630, in methanol at -20 °C. All these observations suggest that the level of interconversion between P630 and luciferin could be the oxide-reduction equilibrium in this system [101]. These observations suggest that reduced P630 is a luciferin molecule, and the oxidized form seems, in these conditions, to be the precursor of luciferin [102]. According to Fresneau and Arrio (1988) [101], the bioluminescence is a complex mechanism controlled by at least two successive reactions. The first is the reduction of the luciferin precursor P630 by a NAD(P)H-dependent reductase [102], and the second is the LCF-luciferin reaction which generates the light emission. Since this molecule is reversibly reduced, it may be considered as an interchange point of reducing power involving a currently unknown electron transfer pathway including NAD(P)H. P630 seems to be a fundamental component involved in complex light-modulated reactions which are widespread in plants [103].

Fresneau and Arrio (1988) [101] also suggest that a recycling process of the luciferin would lead to the impossibility of reaching a level of zero light intensity with homogenates or with mixtures of the purified enzymes and substrates. In such a cyclic process, we would expect a steady state in the light emission, as in firefly extracts where ADP is converted into ATP by a kinase [104]. According with these authors [101], the bioluminescence could be considered as a metabolic process related to the regulation of excess intracellular reducing power produced through the photosynthesis and respiration. This assumption is consistent with the evidence of a specific chlororespiration reported by Bennoun (1982) [105] for higher

plant chloroplasts and other observations in cyanobacteria [106 , 107] of an alternative pathway of respiration in photosynthetic thylakoid membranes [108].

Furthermore, studies focused on the NADPH-dependent detoxification of reactive carbonyls, indicates that eukaryotic cells use different systems to deal with the harmful effects of this reactive molecules. The GSTs pathway, which is fueled by glutathione (GSH) or thioredoxin redox cycle, conjugates aldehydes with GSH, supplying an important mechanism of detoxification. In animals, the main detoxification pathway is GSH-dependent [109], and a GSH-dependent detoxification system is functional in plants as well. The GST proteins constitute a large group; some of these are localized in the chloroplasts and mitochondria. Because chloroplasts contain a milimolar order of GSH, the GSH-dependent detoxification system is seems to be very effective [110].

GSTs comprise a large family of eukaryotic and prokaryotic phase II metabolic isozymes that catalyze the conjugation of the reduced form of GSH to xenobiotic substrates for detoxification [111 , 112]. Could GST be the enzyme NAD(P)H-dependent reductase that controls the state of reduction of P630?. The enzyme GST is also associated with cytochrome P450 [see Additional file 3]. The presence of this GST with a characteristic domain pfam05295, which is part of the cytochrome P450 metabolic cycle, could indicate that the GST enzyme is involved in the synthesis process and/or storage of luciferin through an electron transfer system unknown until now. Also striking is the absence of cox1 (cytochrome c oxidase subunit I) and cox3 (cytochrome c oxidase subunit III) in our results(see section above mitochondrial and plastid genome), which could also be associated with this unknown electron transfer pathway. These are questions that open new lines of investigation in relation to the function of GST and the process of synthesis of luciferin in these organisms. The domain pfam05295 (Luciferase_N) is the common thread between LCF-LBP-GST in *P. lunula*, since all of these proteins contain it.

Nucleosome

The chromosomes of dinoflagellates, having a liquid crystalline structure [113] with bivalent cations acting as the stabilization of the matrix [114], are condensed permanently at all stages of the cell cycle. It was suggested that this DNA could play a role in the organization of the chromosome, maybe by a relation with a matrix protein [28 , 115]. A large proportion of this transcriptionally inactive DNA seems to be repeated sequences, and it is possible that this could contribute to the organization of the genome [82]. The matrix of the nucleus is a network of fibers that also plays an important role in the organization of the chromatin. Furthermore, studies in *Amphidinium carterae* showed the presence of two matrix proteins, topoisomerase II and lamins, similar to what is found in higher eukaryotes [116]. In our case, it was possible to detect the presence of topoisomerase II and III (3-alpha and 3-beta-1, respectively) [see Additional file 3]. Therefore, although this chromatin is organized differently than in other eukaryotes, its nuclear matrix is conserved [117].

A sequence-specific DNA binding protein is the dinoflagellate nuclear associated protein (Dinap1) [118]. Besides Dinap1, homologue of the Tubulin-like protein [119], a group transcription factors (membrane-tethered), related to signaling pathways [120], were found in *Alexandrium*. In our case, it was not

possible to detect the presence of Dinap1, Dip1, DapB, DapC or DapG. However, the presence of the DapA (DapA-like;Aldolase-type TIM barrel) and tubulin-like proteins (Tubulin alpha chain-like) are noteworthy [see Additional file 3]. A group of histone-like proteins (HLPs) [121] were first found in *C. cohnii* [122 , 123]. Nowadays, the presence of these proteins has been reported in other dinoflagellates [124], including our case, where HLPs were detected [see Additional file 3]. In the past, it was thought that dinoflagellates lacked histone proteins [125] and instead used HLPs for DNA organization [126]; nevertheless, recent studies have showed the presence of four core nucleosomal histones, H2A, H2B, H3, and H4 [127 , 128]. It is very important to note that three of the four histones that make up the core nucleosome in dinoflagellates have been found in our results (H2A, H3 and H4), and also two other histone proteins: Histone-lysine N-methyltransferase setd3, and Histone acetyltransferase [see Additional file 3]. Other basic protein identified in our results was DVNP (dinoflagellate/viral nucleoprotein) [see Additional file 3]. DVNP is found only in dinoflagellates [117]. This protein can bind DNA and also be post-translationally modified. In some dinoflagellates, DVNP seems to have displaced major histone functions [129]. It is evident that dinoflagellates show ancient conditions in the organization of chromatin, from which higher eukaryotes evolved the highly conserved nucleosome based organization.

Transcriptional regulation

Transcriptomic and genomic studies in dinoflagellates have revealed there is a paucity of transcription factors (sequence-specific), related with a constant transcription of most genes with some of them under transcriptional control [29 , 30 , 130]. In our results [see Additional file 3], we highlight the presence of putative mediator of RNA polymerase II transcription subunit 37c, DNA-directed RNA polymerase subunit beta, DNA-directed RNA polymerase, transcription factor TFIID complex (transcription initiation from RNA polymerase II promoter), RNA polymerase Rpb1, and DNA -directed RNA polymerase 2A. In addition to RNAP II, the functional eukaryotic transcriptional system requires other basal transcriptional factors (TF) [131 , 132]. TFIID performed the first step of promoter recognition, followed by the TATA binding protein (TBP) and TBP-associated factors (TAFs) [133 , 134]. TBP binding is considered to be the rate-limiting step in the transcription process [135]. Other important factors present in our results related to the transcription process were: transcription factor BTF3, transcription regulators (SARP family), pre-mRNA processing splicing factors (PRP8, 17, SPF27), and different splicing factors (U2AF, 3B subunit 1, 3B subunit 3-like, arginine / serine-rich 4) [see Additional file 3].

One of the most important post-transcriptional modifications is the removal of introns [136-138]. Dinoflagellates genes contain very few or lack introns completely [73 , 93 , 139]. However, in *P. lunula*, the *LCF_c* gene showed a 403 bp intron [84]. Additionally, studies on 17 dinoflagellates species, reported introns in only three: *Peridinium willei*, *Polarella glacialis* and *Thecadinium yashimaense* [117 , 140]. Nevertheless, it is evident that spliced leader (SL) trans-splicing is ubiquitous in dinoflagellates [141-143]. In this process, a unique and a highly conserved spliced leader sequence is transplanted to the 5' end of mRNA molecules. The RNA that donates the spliced leader was found to be mostly < 50 bp [142 , 144 , 145]. SL trans-splicing acts to transform polycistronic to monocistronic mRNA [117 , 141].

The length of the SL exon is different between species [146 , 147]. In dinoflagellates, is a 22 nt sequence 5'-DCCGUAGCCAUUUUGGCUCAAG-3' (D = U, A, or G) [141]. The dinoflagellate SL (DinoSL) sequence contains a Sm binding motif (AUUUUGG) in the exon, unlike all other SL RNAs where this conserved sequence is found in the intron [141]. Although SL trans-splicing is not present in organelle-encoded transcripts, recently a unique type of trans-splicing was found in the mitochondria of some dinoflagellates [148]. SL trans-splicing is evolutionarily ancient in dinoflagellates, and has nuclear-encoded transcripts with 5 different SL sequences, 3 of the SL are 22 nt long and similar to the core dinoflagellates (SL1) while the other two are truncated 21 nt SL with either A or G as the starting nucleotide (SL2) [117 , 149]. A careful analysis of our transcriptomic data [see Additional file 2] reveals the existence of up to thirteen transcripts that include the DinoSL sequence (Table 1), and from this information we could infer the detection of an equal number of introns in those transcripts. However, from the proteomic information we have identified that of the thirteen transcripts that showed the DinoSL sequence, only two can be identified at the proteome level, and the functional description assigned for both is hypothetical protein AN481_18150. This identification is derived from a metagenomic analysis performed with freshwater cyanobacteria belonging to the genus *Aphanizomenon* (GenBank: LJOY01000095.1).

Dinoflagellate introns generally lack the usual splice sites (GU-AG), as found in the case of the AT-TC intron reported in *P. lunula* LCFc [84], the G(C/A)-AG introns in rubisco of *Symbiodinium* [150] and the AG-AG intron in sxtG of *Alexandrium* [151]. Some of these new splice sites, such GC-AG, have been shown to function in other animal and plant genomes [140]. The paucity of introns, as well as the presence of the SL, has led to the proposal of a mRNA recycling system whereby mature mRNAs are inserted back into the genome through a recombination process [66 , 117].

Table 1: Spliced leader (SL) analysis.

Dinoflagellate nuclear DNA is widely methylated. The genes associated with the regulation of methylation are unclear but some potential candidates have been reported. The S-adenosylmethionine synthetase (SAM), which has been detected in our study [see Additional file 3], has also been recognize in other dinoflagellates in the context of saxitoxin synthesis which requires methyl transfer [152]. In addition, SAM itself is methylated and the inhibition of its methylation produce the arrest on cell cycle [153].

Other components related to transcriptional regulation found in our data were: i) ruvB-like 2, RNA-binding protein (RBP), ii) RRM (polyadenylate-binding protein 1, splicing factor U2AF family SnRNP auxiliary factor large subunit, iii) RRM domain-containing protein, and putative RRM domain and KH domain - SPAC30D11.14-like KH- protein), iv) Pumilio (Pumilio homolog 2 isoform X1 and Pumilio RNA-binding repeat), v) S1 (30S ribosomal protein S1), vi) KH (Tudor and KH domain-containing protein), vii) SAP (Heterogeneous nuclear ribonucleoprotein U), viii) LSM (LSM domain, eukaryotic/archaea-type), ix) ATP-dependent RNA helicase (OB NTP-binding), and x) La (Lupus La protein). RNA is attached to a number of RBP ribonucleoprotein complexes that influence splicing, transport to the cytoplasm, translation, stability and subcellular localization [154]. Recently, has been reported a list of RBP containing different RNA

binding domains on mammalian [155], and when these were used to compared with different dinoflagellate species, four RBPs were found to have the greatest level of sequence similarity: Pumilio, S1, the OB NTP fold, and RRM. All these domains have been detected in our results [see Additional file 3]. In general, the studies have revealed that although dinoflagellates make use of some transcriptional regulation, post-transcriptional control is the dominant process of the regulation of gene expression in these organisms [30].

Translational regulation

The basic mechanism of protein synthesis can be divided into four stages: i) initiation, ii) elongation, iii) termination, and iv) recycling. Although a modulation in the translation can take place on all the stages, the kinetics studies have showed that the fundamental regulatory step is the initiation [156 , 157]. The initiation alludes to the disposition of ribosomes (translation-competent), where the initiator tRNA (Met-tRNA_i) in the ribosomal P-site is base-paired to the mRNA start codon. This comprise orthologs to the bacterial/eukaryotic factors IF1/eIF1A and IF2/eIF5B, which are widely conserved in all the domains of life [30]. In our results, we highlight the presence of eIF3F, translation initiation factor 1A (eIF1A), IF1, IF2, eIF4A, eIF4E, eIF4G, putative translation initiation factor eIF5A, and translation elongation factors (Tu, Ts, 1-alpha, 1-beta, 2, 3, P, and G) [see Additional file 3].

Additionally, a study of dinoflagellates transcriptomic datasets and environmental cDNA, revealed the presence of 79 ribosomal proteins that frequently are present in eukaryotes [145], some of which have been reported to be highly represented in dinoflagellates transcriptomes [31]. Currently, little is known about dinoflagellate ribosomes. Nevertheless, it was reported that *Lingulodinium* has ribosomal proteins comparable to those of higher plant and mammalian [28 , 29 , 130]. In our case, a plenty of ribosomal proteins were detected [see Additional file 8]. The importance of translational control is emphasize in *L. polyedra* by the scarce agreement between the quantity of a protein and the level of its transcript [158].

Mitochondrial and plastid genome

Dinoflagellates have an extremely reduced mitochondrial genome, which only contain three proteins: i) cob (cytochrome b), ii) cox1 (cytochrome c oxidase subunit I), and iii) cox3 (cytochrome c oxidase subunit III), and two fragmented rRNAs. Cox2 has apparently being relocated to the nuclear genome [31]. In the case of the proteome of *P. lunula*, evidence of the presence of the cob protein was found, but not of cox1 or cox3 [see Additional file 3]. In dinoflagellates, the mitochondrial genome is also extremely duplicated and recombined. All the genes are present in multiple copies, generally interrupted by fragments of other genes [159]. The stop codon TGA requires either its diversion or RNA editing (for example to TGG) to codons like serine or tryptophan [160]. These genes are transcribed, some of which are polyadenylated [161], and requires trans-splicing to be translatable [31 , 159].

In dinoflagellate containing peridinin plastids, the genome is separated into plasmid-like mini-circles, in most of the cases each containing 1-2 genes. The extension of the mini-circles is different between species. Nevertheless, all the mini-circles share a conserved core in the non-coding region that could

compress transcription initiation signals [162]. The genes encoded in the plastid genome are linked to photosynthesis, including the core subunits of the photosystem, cytochrome b6f, and ATP synthase complex (atpA, atpB, petB, petD, psaA, psaB, psbA-E, psbI) as well as four other genes (ycf16, ycf24, rpl28, and rpl23) [31]. In the case of the proteome of *P. lunula*, evidence of expression of the proteins ATP synthase complex (petB, psaA, psaB) and cytochrome b6f was found. It is noteworthy the presence in our results of important proteins associated with plastids such as: i) ferredoxin, ii) TPT (Phosphotransferase KptA / Tpt1), iii) light harvesting protein, and iv) photosystem II (D2 subunit) [see Additional file 3]. The light harvesting protein and the photosystem II subunits seems to be related with the stabilization and protection of the photosystem, while ferredoxin and TPT play an important role in the process of exporting the products of photosynthesis from the plastid [31 , 163]. In addition, other proteins present in our results, related to plastids, were detected: coproporphyrinogen oxidase isoform 1, glyceraldehyde 3-phosphate dehydrogenase, fructose 1,6-bisphosphatase, glutamate 1-semialdehyde 2,1-aminotransferase, transketolase, NAP50, fructose 1,6 -bisphosphatase isoform 2, and uroporphyrinogen decarboxylase isoform 1 [see Additional file 3]. Studies carried out on non-photosynthetic dinoflagellates (*Noctiluca*, *Oxyrrhis*, and *Dinophysis*) revealed that contain cryptic plastid metabolisms and lack alternative cytosolic pathways, suggesting that all free-living dinoflagellates are metabolically dependent on plastids. These finding led to a hypothesis about the dependency on plastid organelles in eukaryotes that have lost photosynthesis. Furthermore, it is also been proposed that the evolutionary origin of bioluminescence in non-photosynthetic dinoflagellates may be linked to plastid tetrapyrrole biosynthesis [12]. However, further studies are necessary to test this hypothesis.

Proton-pump rhodopsin

The discovery of the bacterio-rhodopsin, which contains all-trans retinal as a chromophore, change the notion that biological conversion of light energy to ATP is only possible through photosynthesis [164]. In dinoflagellates, this rhodopsin was first detected in *P. lunula*, in which was reported to be expressed more actively at early light phase than at early dark phase, suggesting its implication in circadian photoreception and phase shifting. This gene has been also reported in other dinoflagellates (*O. marina*, *Polarella antartica*, *K. veneficum*, *A. cantenella*, *P. lunula*) as well as in natural dinoflagellate assemblage. These results suggest that rhodopsin is ubiquitous in dinoflagellates [145 , 165 , 166]. In fact, in our transcriptomic data (>TRINITY_DN107399_c0_g1_i6 len=1,117) [see Additional file 2] was also detect the presence of rhodopsin (GenBank: AAO14677.1). Furthermore, this rhodopsin, closely related to xanthorhodopsin, is unique in the sense that harvests solar energy with carotenoid molecules as the antenna [167]. Even more, the dinoflagellate rhodopsin sequences contain the characteristic proton donor site. All these facts lead the proposal that dinoflagellate rhodopsin is a proton-pumping, instead of a sensory rhodopsin. On mixotrophic species such as *K. veneficum*, it is remarkable that light stimulate heterotrophic growth [168], where rhodopsin-generated energy may be used to improve ingestion and digestion of preys [145].

Conclusions

The major contribution of the present work is making available a reference transcriptome and proteome of *P. lunula*, that is now accessible for the research community, and a functional description of the 3,182 proteins identified by proteomics and transcriptomic, including 175 novel proteins, which have already been deposited in the ProteomeXchange and NCBI SRA databases, respectively. In addition to this, a series of important factors related to the regulation of gene expression were identified. This annotated transcriptome and proteome should help to accelerate functional genomics in this species and perhaps others commercially and environmentally important microalgae, while full genome sequencing projects are in progress. The transcriptomic and proteomic data should not be seen as a substitute for the sequencing of the genome. These approaches have different strengths and weaknesses and should be considered as complementary and not mutually exclusive. However, until more genomes are available, the *omic* approaches could provide a valuable reference point. The presence of the LBP protein stands out in our results, which previously had been considered absent in the bioluminescent system of the *Pyrocystis* genus. Furthermore, the presence of two isoforms of this gene in *P. lunula*, similar to the reports in other dinoflagellates species, highlights the importance of this protein in the bioluminescence system of *P. lunula*. It is also important to highlight the presence of the GST protein, which could be involved in the process of synthesis and storage (controlling the oxidative degradation) of luciferin in *P. lunula*.

Materials And Methods

Due to the lack of genomic information on *P. lunula*, a refined reference transcriptome was generated through massive sequencing and whose assembly was made through the use of the three main bioinformatics programs currently used for this purpose (RAY, Trinity and MIRA). In parallel, purification of total protein extracts, subsequently analyzed by LC-MS, was carried out. Then, to assess the quality of different assemblies obtained, a multiple proteomic analysis was carried out, using the inferred proteomes from all the *P. lunula* assembled transcriptomes as searched database. Consequently, the best suited assembly was selected and refined to finally produce a comprehensive *P. lunula* reference proteome. A fine-tuned proteomic analysis was carried out against this last, yielding very relevant information on the *P. lunula* proteome composition. Furthermore, a functional analysis of the identified proteins based on gene ontology terms (GO), enzyme codes and pathways (KEGG) associations, was made to obtain a broad picture of the cell functions represented by the molecular components here analyzed. Finally, a process of curing information and data mining was carried out to highlight some of the most relevant molecular components identified within the *P. lunula* proteome.

Culture conditions

P. lunula samples (CCAP 1131/1) were cultured in F2 medium [169] under a photoperiod of 12 h of darkness and 12 h of light at a temperature of 18-20 °C, and a salinity of 31 g/L.

Isolation of total RNA

ARN was extracted using the NucleoSpin RNA® kit (Macherey-Nagel), following the indications of the manufacturer, and quantified by UV spectrophotometry, as well as by electrophoresis in agarose gels. The quality of the isolated RNA was assessed using the 2100 Bioanalyzer System (Agilent Technologies), with the RNA 6000 Nano® kit, following the indications of the manufacturer. Only high-quality RNA (RIN = 8.3 and A260:A280 ratio near 2.0) was used for subsequent experiments.

Transcriptome sequencing

The cDNA library was built using 2 ug of total RNA, in an Ion Chef System (Thermo Fisher Scientific) and the Ion Total RNA-Seq Kit v2 for whole transcriptome libraries (Life Technologies Corporation, California, USA), according to manufacturer instructions. The next generation sequencing was carried out using the semiconductor technology through an Ion S5 System (Thermo Fisher Scientific) using an Ion 540 sequencing chip.

De novo assembly and analysis of high throughput RNA sequencing data

The raw reads obtained from the sequencing platform were pre-processed in order to retain only high-quality sequences to be subsequently used for the assemblies (Cutadapt version 1.9, BBDuk version 35.43). Sequencing adapters were first clipped, and low-quality bases (with phred score below a threshold) were trimmed in raw sequences. Two different phred score values were selected as thresholds (15 and 20), producing two sequence datasets. Then, these datasets were further processed into two each by filtering out reads shorter than 30 or 100 nt, for a total of four sequence datasets labeled according to the quality and length filters they passed (Q15 L30, Q15 L100, Q20 L30 and Q20 L100). Reads quality assessment was carried out using FastQC software (version 0.11.3) to evaluate the effect of every step of this process. All subsequent analyses with RAY software were conducted using these four high quality datasets to evaluate the effect of library size and quality threshold in the tested assembly algorithm throughput. The remaining ribosomal RNA was detected by SortMeRNA software (version 2.0). Thus, adapters clipped reads were mapped to Sortmerna prepackaged databases id98 (from Silva v.119, and Rfam) with default parameters. Three different softwares were employed to assemble the *P. lunula* transcriptome through a *de novo* approach, considering there is no reference genome available to date, which were further assessed to contrast the obtained assemblies by a Proteomic analysis. The four pre-processed sequence datasets described above were assembled into contigs using 5 different parametrizations of kmer (21, 31, 41, 51, 71) with RAY 2.3.1. assembler [170], yielding a total of 20 assemblies. RAY was used to assemble read pairs that overlapped unambiguously, as well as those unmerged paired ends as a pool, since it uses de Bruijn graphs that uses defined seeds that can be extended with the algorithm. The extension process is carried out with a heuristic approach that can stop if the readings family does not clearly indicate the direction of the extension [170]. Additionally, in order to generate more confidence in the assemblies, due to the lack of a reference genome, Trinity and Mira software were also used to generate *de novo* assemblies. Because the RAY assembler results were best with the Q20 L100 sequence dataset, and due to the tremendous computational resources required by Trinity and Mira with the increase of the input reads file, only that

sequence collection fitting the highest quality requirements (Q20 L100) was employed with Trinity and MIRA assemblers. Trinity 2.5.1, which performs a *de novo* assembly using an algorithm based also on de Bruijn graphs [171], was used for the Q20 L100 sequence dataset with the following parameters: “–max_memory 1000G–CPU 20 –SS_lib_type F–bflyCalculateCPU–normalize_max_read_cov 20”, with “–KMER_SIZE 25” or “31”, also with and without “–min_kmer_cov 2”, yielding a total of 4 assemblies. MIRA 4.9.6 software [172], unlike Trinity and RAY, is based on the strategy known as Overlap/Layout/Consensus, and was employed to produce one assembly with the Q20L100 sequence collection and the following parameters: “job = denovo,est,accurate; COMMON_SETTINGS - GENERAL:number_of_threads=12 -KMERSTATISTICS:lossless_digital_normalisation=yes; IONTOR_SETTINGS -ALIGN:min_relative_score=70 -ASSEMBLY:minimum_read_length=100; - CLIPPING:quality_clip=no -CLIPPING:qc_minimum_quality=15 -CLIPPING:qc_window_length=20; - CLIPPING:clip_polyat=yes -CLIPPING:cp_min_sequence_len=12; technology = iontor”. The evaluation of the structure of the 25 generated assemblies was performed by an in-house R script. In addition, a BLAST search was carried out to compare all assemblies against each other (NCBI BLAST 2.2.31+) [173] to assess their representativeness.

Proteomics analysis

The mechanical lysis of the samples was made with a Minilys® homogenizer. The protein precipitation was carried out using a previously described phenol-based procedure [174]. The assay was performed in triplicate. Once the protein extracts from *P. lunula* were obtained, the protein concentration was estimated by fluorometric measurement using Qubit System® (Thermo Fisher, Inc.). Proteomic analysis was made by LC-MS. All samples were analyzed as described in [175]. Briefly, Nano-LC were performed in a Dionex Ultimate 3000 nano UPLC (Thermo Scientific, San Jose, CA) with a C18 75 µm ´ 50 Acclaim Pepmap column (Thermo Scientific) at 300 nl/min and 40 °C. The peptide mix was previously desalted and cleaned up on a 300 µm ´ 5 mm Acclaim Pepmap cartridge (Thermo Scientific) in 2% acetonitrile/0.05% formic acid for 3 min at 5 µl/min. Buffer A (0.1% formic acid) and Buffer B (80% acetonitrile, 0.1% formic acid) was used as Mobile phase in chromatographic separations with the next conditions: 4-35%B for 60 min; 35-55% B for 3 min; 55-90% B for 3 min followed by 8 min wash at 90% B and a 12 min re-equilibration at 4% B. Eluting positive peptides were analyzed in positive mode on a trihybrid Thermo Orbitrap Fusion (Thermo Scientific) with nano-electrospray ionization source operating in Top30 Data Dependent Acquisition mode with maximum cycle time of 3 s. MS1 full scans were acquired in a 400 to 1500 m/z mass range at 120,000 resolution (at 200 m/z). Only those precursor ions with charge state 2–5 were sampled for MS/MS. Monoisotopic ions, previously isolated in the quadrupole at 1.2 Da, were CID-fragmented in the ion trap with a 35% normalized collision energy. The automatic gain control was 2 ´ 10³ and the maximum injection time was 300 ms. A dynamic exclusion time was set to 15 s with a 10 ppm tolerance around the selected precursor and its isotopes.

Gene ontology and KEGG pathway analysis

Functional annotation of the gene ontology (GO) terms, Enzyme codes, InterPro and KEGG pathways associations was done using the BLAST2GO program [176, 177]. Associated Enzyme codes were mapped to reference KEGG pathways (KO) to better illustrate representative pathway components identified in *P. lunula* (KEGG mapper tool).

Verification of the presence of LBP and LCF/LBP by PCR

The purification of the total RNA was carried out following the indications of the commercial kit NucleoSpin RNA® (Macherey-Nagel). The synthesis of cDNA and amplification of the fragments by RT-PCR was carried out using qScript® kit (Quanta Biosciences), Phusion Flash High-Fidelity PCR Master Mix (Thermo Scientific) and Q5® High-Fidelity DNA Polymerase (New England Biolabs Inc.), following the recommendations of the manufacturers. The conditions of the reaction to amplify LBP were as follows: 2X Phusion Flash PCR Master Mix 10 µL, 10 µM LBP F: 5'-TGT GGC GGA CCA TGG TGG A-3' 1 µL, 10 µM LBP R 5'-GGT CTT TGA TGG AGA GCT CCA-3' 1 µL, cDNA (65 ng/µL) 1 µL and water 7 µL, for a total reaction volume of 20 µL. With the LBP F and LBP R primers a fragment of 453 pb is obtained. The thermocycler program was as follows: initial denaturation at 98°C for 10 sec, followed by 35 cycles at 98 °C for 1 sec, 66 °C for 5 sec, 72 °C for 8 sec, with a final extension cycle at 72 °C for 1 min and stop at 4°C. In the case of the hybrid isoform LCF/LBP, the conditions were as follows: 5X Q5 reaction buffer 5 µL, 10 µM LCF/LBP F: 5'-GAG GAC TAC TCG AAC GGA TTG ATC C-3' 1.25 µL, 10 µM LCF/LBP R: 5'-GTC CAT GAT GCG CCA GGG CT-3' 1.25 µL, 10 mM dNTPs 0.5 µL, cDNA (65 ng/µL) 1 µL, Q5 High-Fidelity DNA polymerase 0.25 µL, and water 15.75 µL, for a total reaction volume of 25 µL. With the primers LCF/LBP F and LCF/LBP R a fragment of 1,395 pb is obtained. The thermocycler program was as follows: initial denaturation at 98 °C for 30 sec, followed by 35 cycles at 98 °C for 10 sec, 69 °C for 20 sec, 72 °C for 45 sec, with a final extension cycle at 72 °C for 2 min and stop at 4 °C. For the purification of the fragments LBP and LCF/LBP, the instructions of the GeneJET Gel Extraction Kit (Thermo Scientific) were followed. The fragments were visualized by electrophoresis (1% agarose, stained with Gel-Red) run for 1 h at 110 V. The results were corroborated by sequencing and BLASTx analysis [85].

List Of Abbreviations

Dinap: Dinoflagellate nuclear associated protein; DinoSL: Dinoflagellate spliced leader; DVNP: Dinoflagellate/viral nucleoprotein; GO: Gene ontology, GSH: Glutathione (reduced); GST: Glutathione-S-transferase; HLP: Histone-like protein; KEGG: Kyoto encyclopedia of genes and genomes; KO: KEGG orthology; LBP: Luciferin-binding protein; LCF/LBP: Hybrid isoform LCF/LBP; LCF: Luciferase; RBP: RNA-binding protein; RNAP: RNA polymerase; SAM: S-adenosylmethionine synthetase; SL: Spliced leader; TAF: TBP-associated factor; TBP: TATA binding protein; TF: Transcriptional factors.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets generated and/or analyzed during the current study are available in the [NCBI SRA], [\[https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA497104\]](https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA497104); and [ProteomeXchange] repositories, [\[http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD011555\]](http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD011555).

Competing interests

The authors declare that don't have competing interests.

Funding

This research has been financed by the INNOVAALGA project (Exp. 14/775) from CTA and A4HW project (RTC-2016-4860-2) financed by the Ministry of Economy, Industry and Competitiveness under the Collaboration Challenges program. This aid is co-financed by the European Union through FEDER funds. The funding body declares that don't have influence in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

CF performed the extraction procedure of RNA and total proteins, design of primers, PCR analysis, compilation of bibliographic information, interpretation of results, and was the major contributor in writing the manuscript; FAR was responsible for the bioinformatics analyzes and contributed to the drafting of the manuscript; CFA was responsible for the MS analyzes and contributed to the drafting of the manuscript; MDD contributed to the drafting of the manuscript; AEN contributed to the drafting of the manuscript; RC contributes as lab manager and material optimization; JMM contributed to the drafting of the manuscript; FJFA was the coordinator of the project, designed and supervised the experimental processes and contributed to the drafting of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Bioinformatics analyses were performed at the Bioinformatics Units, Central Facilities for Research Support (SCAI), University of Cordoba (UCO), Rabanales, Spain. The procedure of verification of the presence of LBP and LCF/LBP by PCR was carrying out on the Department of Marine Biology and Aquaculture (ICMAN), Spanish National Research Council (CSIC), Puerto Real, Spain. This research was partly sponsored through scholarships granted by the International Relations Office of the University of Cadiz (ORI-UCA) and the Ibero-American Postgraduate University Association (AUIP).

Additional Files Legend

File name: Additional file 1.

File format: .pdf

Title of data: Similarity analysis of all datasets by BLASTn.

Description of data: Summary of the 3,645 BLASTn searches comparing all 25 assemblies and two additional public datasets.

File name: Additional file 2.

File format: .pdf

Title of data: Assembly generated with Trinity software (Trinity_Q20 L100_k25_NC20).

Description of data: Transcriptome assembly that best suited to the identification profiles obtained by MS analysis.

File name: Additional file 3.

File format: .xlsx

Title of data: Final result of the proteome annotation.

Description of data: File that includes the result of all identification rounds through BLAST2GO.

File name: Additional file 4.

File format: .pdf

Title of data: KEGG analysis.

Description of data: All the enzymatic codes and pathways identified in the proteome of *P. lunula*.

File name: Additional file 5.

File format: .xlsx

Title of data: Novel proteins in *P. lunula*.

Description of data: Sequences that do not have any homology with any protein identified so far in the different databases.

File name: Additional file 6.

File format: .docx

Title of data: LBP sequence.

Description of data: Sequencing analyses of the LBP fragment obtain by PCR.

File name: Additional file 7.

File format: .docx

Title of data: LCF/LBP sequence.

Description of data: Sequencing analyses of the LCF/LBP fragment obtain by PCR.

File name: Additional file 8.

File format: .xlsx

Title of data: Ribosomal proteins detected in the proteome of *P. lunula*.

Description of data: All the ribosome related proteins identified in the proteome of *P. lunula* according to BLAST2GO.

Table

Table 1: Spliced leader (SL) analysis.

<i>Transcriptomic data</i> <i>(Assembly Trinity Q20 L100 k25 NC20)</i>	<i>Type of SL present in the sequence †</i> <i>‡ §</i>	<i>Proteomic data</i> <i>(Functional description)</i>
>TRINITY_DN13698_c0_g1_i1 len=428	SL1	No detected
>TRINITY_DN82698_c0_g1_i1 len=437	SL2	No detected
>TRINITY_DN105556_c1_g1_i2 len=2597	N/A	No detected
>TRINITY_DN105556_c1_g1_i3 len=1761	N/A	No detected
>TRINITY_DN105556_c1_g1_i8 len=2511	N/A	No detected
>TRINITY_DN105556_c1_g1_i9 len=1194	N/A	No detected
>TRINITY_DN105556_c1_g1_i15 len=2037	N/A	No detected
>TRINITY_DN103446_c2_g1_i4 len=292	SL1	No detected
>TRINITY_DN103446_c2_g1_i5 len=874	SL1	Hypothetical protein AN481_18150
>TRINITY_DN103446_c2_g1_i6 len=873	SL1	No detected
>TRINITY_DN103446_c2_g1_i7 len=788	SL1	Hypothetical protein AN481_18150
>TRINITY_DN107555_c3_g1_i12 len=334	SL1	No detected
>TRINITY_DN106580_c3_g1_i1 len=279	SL1	No detected

† SL1: Nucleotide of the beginning of the sequence T.

‡ SL2: Nucleotide of the beginning of the sequence A or G (A in this case).

- N/A: There are no more available reads at the beginning of the sequence.

Figures

Assemblies Structure

Figure 1

Structural analysis of 20 assemblies generated by RAY

Figure 2

Structural analysis of 10 assemblies generated by all three algorithms (RAY, Trinity, MIRA) from the Q20 L100 dataset.

Figure 3

Analysis of representativeness by sequence similarity for Trinity_Q20 L100_k25_NC20 against all 25 assemblies and two more public datasets

Figure 4

Page 26/31

Distribution of the predicted and identified proteins according to their MW and IP calculated by the R library Peptides.

Figure 5

Functional description according to biological process (BP) classification.

Figure 6

Functional description according to molecular function (MF) classification.

Top-Hit Species Distribution [ids_clean_bxmapann_ipr_merge_annex_bx2_dcmbn]

Figure 7

Top-hit species identification according to BLAST2GO.

Figure 8

M: Molecular weight ladder; 1) LBP fragment amplified by PCR (453 pb) and subjected to sequencing and BLASTx analyses.

Figure 9

M: Molecular weight ladder; 1) LBP/LCF fragment amplified by PCR (1,395 pb) and subjected to sequencing and BLASTx analyses.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplement1.docx](#)
- [supplement2.docx](#)
- [supplement3.pdf](#)
- [supplement4.xlsx](#)
- [supplement5.xlsx](#)
- [supplement6.pdf](#)
- [supplement7.xlsx](#)
- [supplement8.pdf](#)