Symplectin evolved from multiple duplications in bioluminescent squid

Warren R. Francis[1,2], Lynne M. Christianson[1] and Steven H.D. Haddock[1]

[1] Monterey Bay Aquarium Research Institute, Moss Landing, CA, United States of America
[2] Department of Biology, University of Southern Denmark, Odense, Denmark

ABSTRACT

The squid *Sthenoteuthis oualaniensis*, formerly *Symplectoteuthis oualaniensis*, generates light using the luciferin coelenterazine and a unique enzyme, symplectin. Genetic information is limited for bioluminescent cephalopod species, so many proteins, including symplectin, occur in public databases only as sequence isolates with few identifiable homologs. As the distribution of the symplectin/pantetheinase protein family in Metazoa remains mostly unexplored, we have sequenced the transcriptomes of four additional luminous squid, and make use of publicly available but unanalyzed data of other cephalopods, to examine the occurrence and evolution of this protein family. While the majority of spiralians have one or two copies of this protein family, four well-supported groups of proteins are found in cephalopods, one of which corresponds to symplectin. A cysteine that is critical for symplectin functioning is conserved across essentially all members of the protein family, even those unlikely to be used for bioluminescence. Conversely, active site residues involved in pantetheinase catalysis are also conserved across essentially all of these proteins, suggesting that symplectin may have multiple functions including hydrolase activity, and that the evolution of the luminous phenotype required other changes in the protein outside of the main binding pocket.

Subjects Biochemistry, Evolutionary Studies, Genomics, Marine Biology

Keywords Luciferase, Neofunctionalization, Coelenterazine, Squid, Gene duplication, Symplectin, Evolution, Bioluminescence

Submitted 22 May 2017
Accepted 11 July 2017
Published 31 July 2017

Corresponding author
Steven H.D. Haddock,
haddock@mbari.org

Academic editor
Robert Toonen

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.3633**

INTRODUCTION

Luminous cephalopods (squids and octopods) generate light through mechanisms both native to the animal and through control of symbiotic bacteria (*Haddock, Moline & Case, 2010*). Although two independent evolutionary events of bacterial bioluminescence have been identified in squid (*Lindgren et al., 2012*), native bioluminescence (also called autogenic bioluminescence) is by far more common. The bioluminescence mechanisms of many of these autogenic species are unknown, although it is known that most of the species make use of the same luciferin, coelenterazine, which is the consumable substrate for the reaction. Coelenterazine is the most widely occurring luciferin in marine bioluminescence, its use being report in at least eleven phyla (*Haddock, Moline & Case, 2010*). Some species have been found to use forms of coelenterazine with additional modifications: the firefly squid *Watasenia scintillans* utilizes sulfated coelenterazine (*Inoue et al., 1975*), while the squid *Sthenoteuthis oualaniensis* (*Takahashi & Isobe, 1994*) and the

clam *Pholas dactylus* (*Tanaka, Kuse & Nishikawa, 2009*) use a version that is dehydrated at the 2-position, called dehydrocoelenterazine (hereafter abbreviated as dhCtz).

Much of the work on coelenterazine-using bioluminescent systems has focused on cnidarians (both hydromedusae like *Aequorea victoria* or *Obelia* spp. and the octocoral *Renilla reniformis*) and crustaceans like *Gaussia princeps* and *Oplophorus*. Nonetheless, some key protein components have been identified in two squid species. The luciferase of the firefly squid *Watasenia scintillans* was recently identified (*Gimenez et al., 2016*), belonging to the same protein superfamily as firefly luciferases. The other protein comes from the purpleback flying squid *Sthenoteuthis oualaniensis*, where a 500-amino acid photoprotein has been cloned and characterized (*Isobe et al., 2008*) and was named symplectin. While the family of firefly luciferases and marine coelenterazine luciferases has been well studied, little attention has been given to the distribution and origins of symplectins.

Previous work highlighted the sequence similarity of symplectin to members of the biotinidase/pantetheinase family (*Fujii et al., 2002*), part of the superfamily of carbon-nitrogen hydrolases. Such proteins have been characterized in mammals, and in *Drosophila* (*Swango & Wolf, 2001*) and have roles in recycling of the enzymatic cofactors biotin and pantothenic acid. However, information about this protein family is limited outside of model organisms. Thus, very little insight could be offered to explain how a biotinidase evolved into a photoprotein, and how this apparently happened independently of other mechanisms of bioluminescence in squid.

Transcriptomics has proven helpful in identifying new fluorescent proteins (*Hunt et al., 2012*) and photoproteins (*Powers et al., 2013*) in organisms where traditional cloning approaches have failed. Based on this, we sought to identify orthologs of symplectin in transcriptomes of other luminous squid. We present transcriptomes from light-producing tissues of four luminous squid, *Chiroteuthis calyx*, *Octopoteuthis deletron*, *Vampyroteuthis infernalis*, *Pterygioteuthis hoylei*, and one luminous nudibranch *Phylliroe bucephala*. We found homologs of symplectin in all five species, and then further compared them to homologs identified in other animal phyla. Many important catalytic residues are conserved across the entire protein family, including among cephalopod proteins. As this protein family has undergone many independent expansions in several other groups since the last common ancestor of animals, it may be that large expansions of protein families facilitate evolution of novel phenotypes, including bioluminescence.

MATERIALS AND METHODS

Specimens and sequencing

Specimens of *Chiroteuthis calyx*, *Octopoteuthis deletron*, and *Vampyroteuthis infernalis* were collected around the Monterey Bay, off the coast of California using detritus samplers on ROVs (remotely-operated vehicles) from the Monterey Bay Aquarium Research Institute (MBARI). *Pterygioteuthis hoylei* and *Phylliroe bucephala* were caught in the Gulf of California by trawls and blue-water diving, respectively. ROVs were operated during daytime hours, between 07:00 and 19:00 h. Operations were conducted under permit SC-4029 issued to SHD Haddock by the California Department of Fish and Wildlife.

Table 1 Transcriptomic data sources.

Species	Tissue	Luminous	Bases (Gb)	Assembled transcripts	Accession	Reference
Chiroteuthis calyx	Arm	Yes	6.3	78,445	SRR5527417	This study
Octopoteuthis deletron	Arm tips	Yes	6.1	122,672	SRR5527415	This study
Vampyroteuthis infernalis	Arm tips	Yes	6.4	149,961	SRR5527416	This study
Pterygioteuthis hoylei	Photophore	Yes	6.8	93,201	SRR5527418	This study
Dosidicus gigas	Photophore	Yes	6.0	94,197	SRR5152122	*Francis et al. (2013)*
Watasenia scintillans	Arm/mantle	Yes	30.0	216,307	GEDZ00000000.1	*Gimenez et al. (2016)*
Uroteuthis edulis	Photophore	Bacterial	8.8	119,033	PRJNA257113	*Pankey et al. (2014)*
Euprymna scolopes	Multiple	Bacterial	45.6	280,433	PRJNA257113	*Pankey et al. (2014)*
Octopus vulgaris	Nerve/brain	No	1.2	59,859	JR435555	*Zhang et al. (2012)*
Sepia pharaonis	Whole animal	No	4.6	131,176	SRR3011300	*Wen et al. (2016)*
Loligo vulgaris	Sucker	No	3.1	43,951	SRR3472303	*Jung et al. (2016)*
Doryteuthis pealeii	Multiple	No	15.2	212,516	PRJNA255916	*Alon et al. (2015)*
Phylliroe bucephala	Whole animal	Yes	6.1	66,535	SRR5527414	This study

Species used are unprotected and unregulated, and no vertebrates or octopus were used, so the International and NIH ethics guidelines are not invoked, although organisms were treated humanely. All samples were flash-frozen in liquid nitrogen. Total RNA was extracted with the Qiagen RNA-easy kit following manufacturer's instructions. All samples were sequenced at the University of Utah using the Illumina HiSeq2000 platform. Libraries were generated using the Illumina TruSeq kit, with oligo-dT selection, and were run with six samples per lane. Assemblies were generated as previously described (*Francis et al., 2013*). NCBI SRA numbers are given in Table 1. All assemblies can be downloaded at https://bitbucket.org/wrf/squid-transcriptomes.

Data acquisition

We made use of public data from a number of previous studies (Table 1). When available, assembled transcriptomes were used. Otherwise, data were downloaded from NCBI SRA and assembled with Trinity v2.2.0 (*Grabherr et al., 2011*), with the options `--normalize_reads` and `--trimmomatic`.

Gene trees

Homologs of symplectin were identified by BLAST alignment using blastp or tblastn, using symplectin (NCBI accession: C6KYS2.2) as the query and an *e*-value threshold of 10^{-10}. All BLAST searches were done using the NCBI BLAST 2.2.29+ package (*Camacho et al., 2009*). As the query was 501 amino acids, partial proteins under 100 amino acids were unlikely to provide useful comparisons and were excluded. Alignments for protein sequences were created using MAFFT v7.029b, with L-INS-i parameters for accurate alignments (*Katoh & Standley, 2013*). Phylogenetic trees were generated using either FastTree (*Price, Dehal & Arkin, 2010*) with default parameters or RAxML-HPC-PTHREADS v8.2.10 (*Stamatakis, 2014*), using the PROTGAMMALG model for proteins and 100 bootstrap replicates with the "rapid bootstrap" (-f a) algorithm.

Structure modelling

A model structure of symplectin was generated using the HHPred webserver (*Alva et al., 2016*), with human vanin-1 (PDB accession: 4CYG) as the template protein, using default parameters. Vanin-1 was the highest scoring model available (probability of 100, *e*-value of $1.4 * 10^{-71}$). Model evaluations can be downloaded at https://bitbucket.org/wrf/squid-transcriptomes.

RESULTS

Symplectin-like proteins in other cephalopods

We generated approximately 6 Gb of pair-end RNAseq data for 4 squid and one luminous nudibranch (Table 1). Reads for each species were assembled de novo using Trinity (*Grabherr et al., 2011*). With symplectin as the query, we used BLAST (*Camacho et al., 2009*) to identify homologs in the transcriptomes and genomes of other animals. In general, cephalopods have four copies of this protein sharing a single origin (Fig. 1), while the precise number varies between species, likely subject to coverage limits of transcriptome sequencing or tissue-specific expression. We found full-length proteins of each of the four groups from only three species, *Watasenia scintillans*, *Sepia pharaonis* (non-luminous), and *Pterygioteuthis hoylei*. The symplectin group and group 1 proteins were not found in any octopodiform (represented by only *V. infernalis* and two *Octopus* species) (*Zhang et al., 2012*; *Albertin et al., 2015*).

Of the four protein groups, homologs of symplectin were found in two luminous species where the mechanisms of light production are unknown, *D. gigas* and *P. hoylei*. These homologs could potentially be the active protein in the bioluminescence of these species. However, the firefly squid *W. scintillans* has five proteins in total, one of each of the four groups, as well as an additional duplicate of the symplectin group. We were also able to find symplectin homologs in three species that are not luminous, *Sepia pharaonis*, *Loligo vulgaris*, and *Doryteuthis pealei*, indicating that the tree position alone is not enough to predict bioluminescence of the organism, or whether the enzyme could act as a photoprotein; it could be that all four cephalopod protein groups can act as photoproteins, though it could also be that symplectin is the only enzyme that acts as a photoprotein in this entire protein family. Additionally, two species that have luminous systems which do not involve coelenterazine still have members of this protein family: the squid *Euprymna scolopes* and *Uroteuthis edulis* both generate light from interactions with symbiotic, luminous bacteria, yet we still found homologs from protein groups 1, 2 and 3.

Symplectin-like proteins across metazoans

To better understand the relative importance of the cephalopod duplications, we examined homologs of symplectin across Metazoa (Fig. 2). At the sequence level, symplectin is similar to two annotated proteins in human (*Fujii et al., 2002*), vanin-1 (pantetheinase, 30% identity) and biotinidase (btd1, 31% identity), both of which are hydrolyzing amidases and do not require other cofactors. Homologs of symplectin/biotinidase were found in most animal groups, including vertebrates, arthropods, cnidarians, sponges, placozoans, as well as choanoflagellates (single-celled eukaryotes that are sister-group to animals). We were

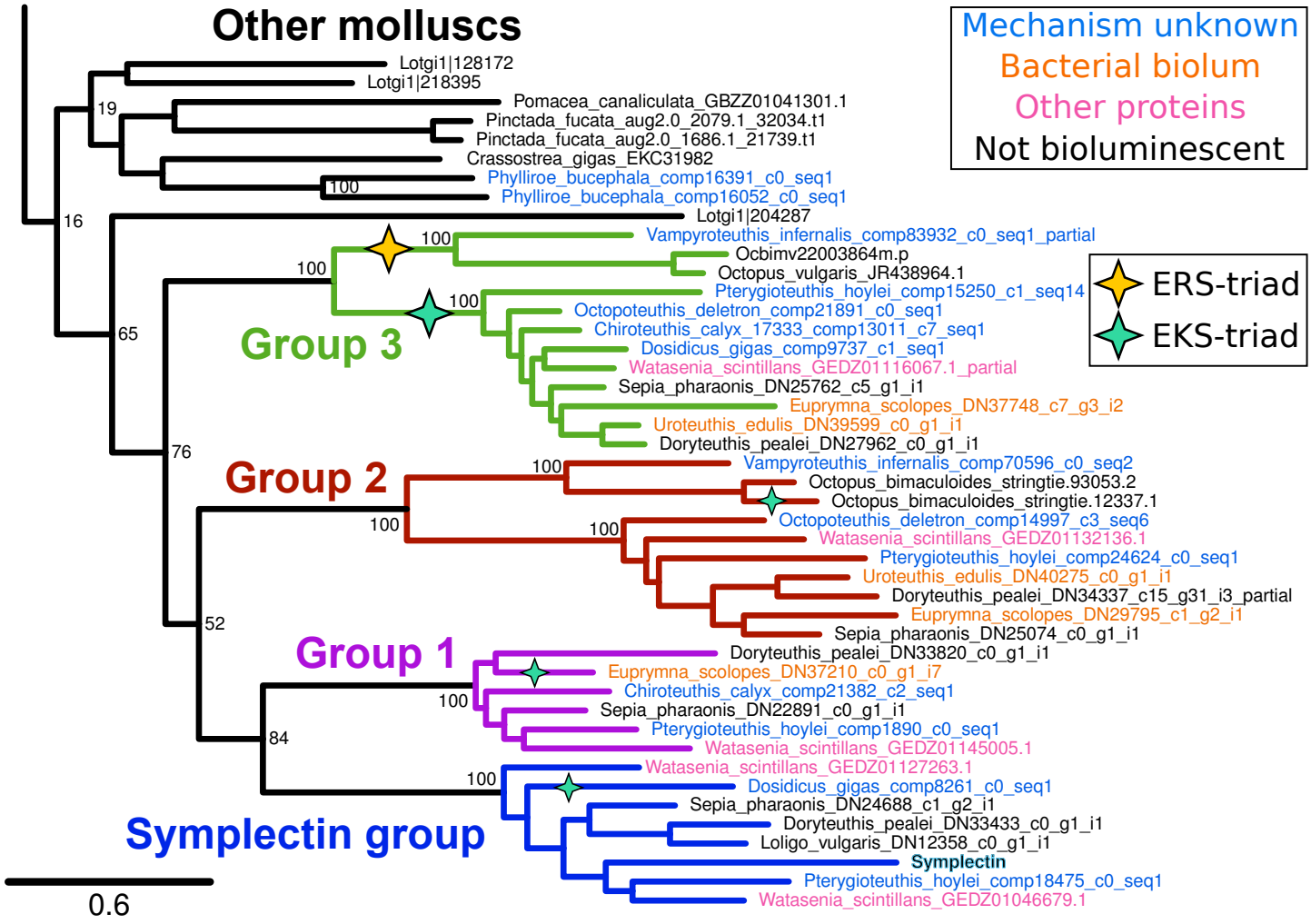

Figure 1 Phylogenetic tree of symplectin and homologs from other molluscs. Four strongly supported groups are identified in cephalopods. Species that are luminous but have unknown mechanisms are high-lighted in blue, while those with bacterial bioluminescence or other known proteins are highlighted in orange and pink, respectively. Black labels indicate non-luminous species. Clades/proteins with a catalytic triad of E-K-S have a green star, and those with E-R-S triad have a yellow star; all others have the conserved E-K-C triad. Internal bootstrap values are removed for clarity. Complete version of the same tree containing homologs from across metazoans is shown in Fig. 2.

unable to find any homologs in hemichordates (based on the genomes of *Ptychodera flava* and *Saccoglossus kowalevskii*) nor any ctenophores (based on the genome of *Mnemiopsis leidyi* and transcriptomes from 11 species). Because of the presence in choanoflagellates, absence of this protein family in hemichordates and ctenophores is likely a secondary loss. However, hemichordates are only represented by two species so sequencing of additional species, or deeper sequencing of the studied species, may reveal homologs in this clade.

Comparison to annotated proteins

It is clear from the alignment that the original symplectin sequence is not the complete CDS (Fig. 3), as it does not start with methionine, and around 30 residues are missing compared to symplectin homologs in other cephalopods. At present, no transcriptomic

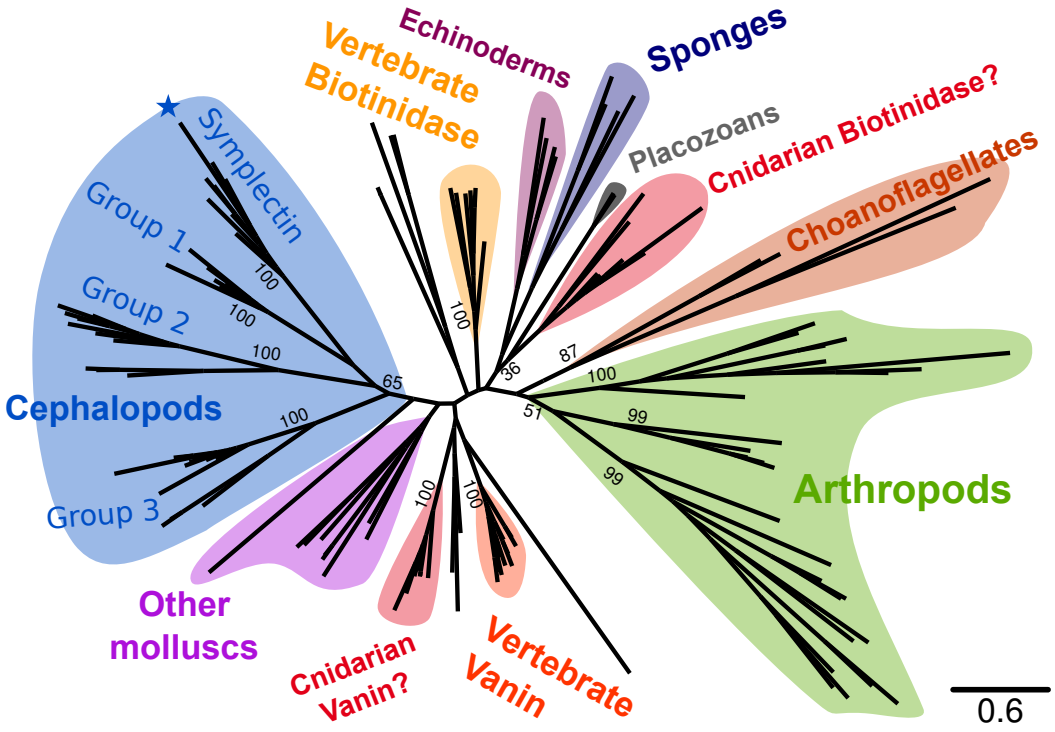

Figure 2 Phylogenetic tree of known symplectin and homologs from all animals. Cephalopod-specific groups are indicated as in Fig. 1. Symplectin is indicated by the blue star. Most species have only a single homolog and the proteins form a clade by phylum. This is not true of chordates (yellow and orange for biotinidase and vanin-1/pantetheinase, respectively) and cnidarians (red). Internal bootstrap values are removed for clarity, though many support values for backbone nodes are low. Detailed version of the same tree showing only molluscs is shown in Fig. 1.

or genomic data is available for *S. oualaniensis*, thus we could not examine the protein completeness or copy number in that species.

While symplectin does not have an available crystal structure, the structure of the human protein vanin-1 has been determined (*Boersma et al., 2014*), which is the only member of this family to have a crystal structure. The vanin-1 protein structure is divided into two domains, the catalytic "nitrilase" domain, and the base domain, which has an unknown function in vanin-1 (*Boersma et al., 2014*). The catalytic triad residues of human vanin-1 include a glutamate (E79), lysine (K178) and cysteine (C211) (*Boersma et al., 2014*). Two of the residues, K178 and E79, are conserved in essentially all taxa, including cephalopods (though this cannot be evaluated in partial sequences). The cysteine is also conserved in most proteins, except some cephalopods sequences have a serine instead of cysteine, although serine could still function as a nucleophile for hydrolysis. All Group 3 proteins are serine-containing, but otherwise the serine proteins do not form a clade, indicating that this mutation has occurred multiple times independently. The cysteine is followed by a phenylalanine (F212 in vanin-1) in nearly all proteins, while most of the cephalopod serine proteins are followed by tyrosine, suggesting some role of this position in the specificity or reactivity of the enzyme.

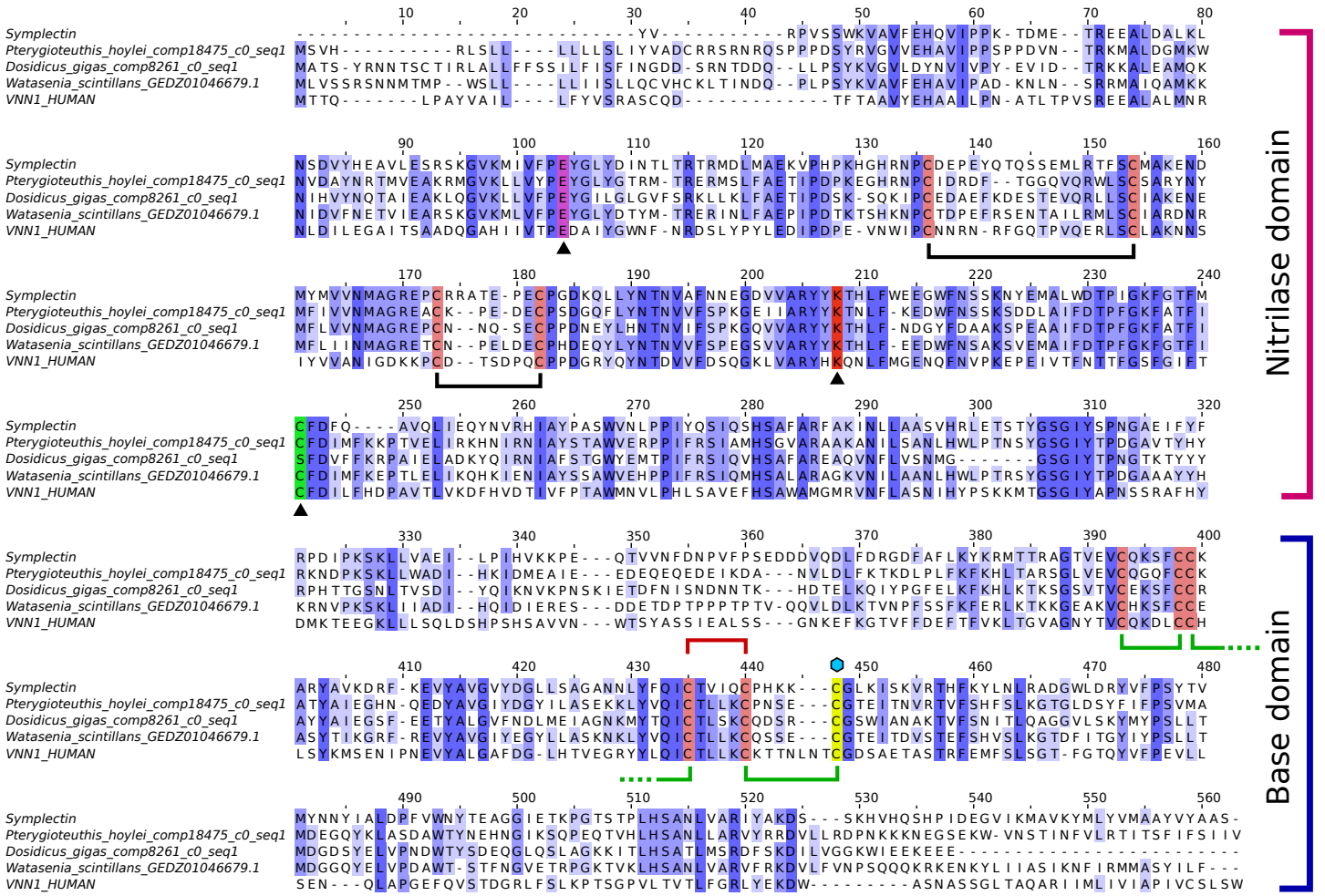

Figure 3 Alignment of symplectin and vanin-1. Multiple sequence alignment of symplectin and vanin-1 with top hits from *D. gigas*, *P. hoylei*, and *W. scintillans*. Intensity of blue color shows conservation. Catalytic residues (E, K, C) identified in vanin-1 are indicated by triangles beneath, and the catalytic cysteine is shown in green, though this position is a serine for *D. gigas*. The dhCtz-binding cysteine is shown in yellow, indicated by the blue hexagon above. Disulfide bridges found in both symplectin and vanin-1 are shown in black. Those found only in vanin-1 are shown in green, while the one remaining disulfide bond found in symplectin is shown in red. For Group 2 cephalopod proteins (Fig. 1), the conserved cysteine at alignment position 440 is substituted and a neighboring residue (I383 in symplectin) is instead a cysteine.

Conservation and role of cysteines

Instead of using the more-common coelenterazine molecule, symplectin requires dehydrocoelenterazine (dhCtz). Unlike cnidarian or ctenophore photoproteins, which hold coelenterazine through a peroxide bond to tyrosine (*Head et al., 2000*), dhCtz is linked to symplectin by a thioether to a free cysteine residue (C390 in Symplectin) (*Isobe et al., 2008*). Symplectin has 11 cysteines (Fig. 3), initially suggesting that the odd number allowed for a free cysteine to bind coelenterazine. However, mass spectrometry data indicate that only three pairs are in the form of disulfide bridges (*Kongjinda et al., 2011*), though perhaps the remaining two pairs are not effectively captured. If the conserved cysteine (C196) is actually catalytic in the binding pocket of the nitrilase domain of the symplectin

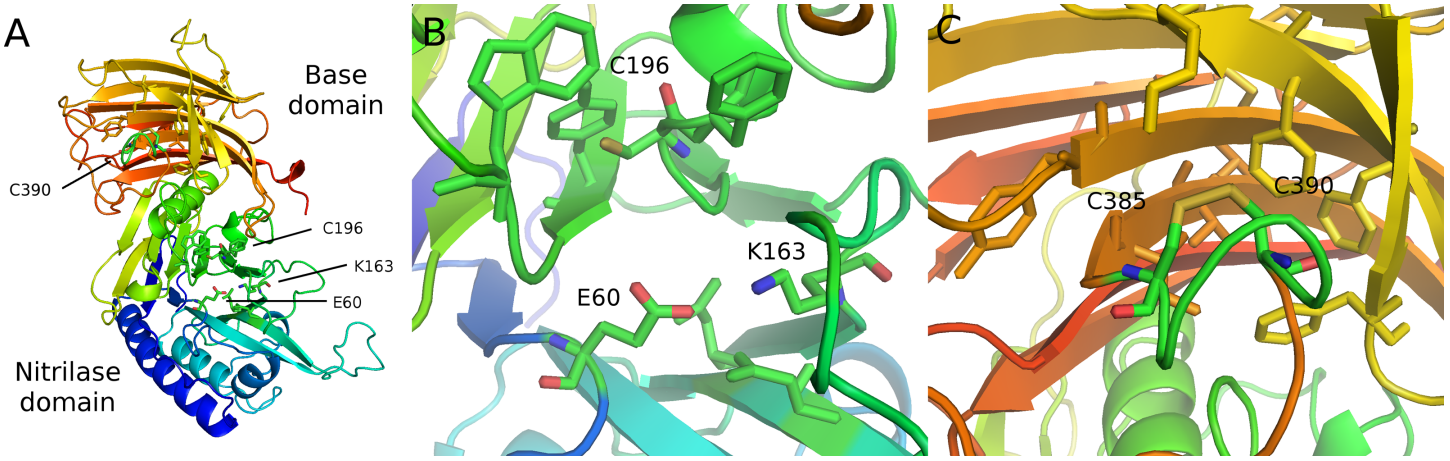

Figure 4 Modeled structure of symplectin. PDB format structure based on the structure of vanin-1 (4CYF). (A) Overview of the two domains, named after those defined in vanin-1. Residues 1–290 (blue to green) compose the nitrilase domain while residues 291–465 compose the base domain (yellow to red). (B) Close-up view of the catalytic triad E60-K163-C196 (C) Putative disulfide bridge of C390 and C385, while overall these residues are located close to a number of hydrophobic residues, potentially involved in dehydrocoelenterzine binding.

and does not form a bridge, and three pairs are already identified, then this leaves one additional pair of cysteines for disulfide bonding.

C390 is conserved in essentially all other proteins of this family, including biotinidase (C471) and vanin-1 (C411). In vanin-1, C411 is in close proximity to C403 and forms a disulfide bond, though apparently this cannot occur in symplectin between C385 and C390, since C390 would no longer be available for thioether bonding to dhCtz (*Isobe et al., 2008*). Mass spectrometry data suggest that C385 instead forms a bridge with C380 (*Kongjinda et al., 2011*). However, when symplectin is modeled based on the structure of vanin-1 (Fig. 4), C380 would be on another beta strand and point away from the binding pocket to bridge with C345, while the adjacent cysteine (C344) bridges further away to C339 in a beta-hairpin motif. The cysteine pair for C345/C380 is found in all proteins suggesting an important structural role of C380, and arguing against the formation of the C380/C385 bond. By comparison, the C344/C339 pair is not found in sponges, polychaetes, or choanoflagellates, suggesting it is dispensable.

This discrepancy may be reconciled three ways: (1) the structures of symplectin and vanin-1 differ enough that vanin-1 cannot reliably be used to model disulfide bridges in symplectin; (2) C385/C390 natively form a disulfide bond, though when digested, other changes in the redox state disrupt this bond and the portion containing C380 and C385 dynamically forms a disulfide bond when fragmented. However, this presents an additional problem, as even if C385 and C390 do indeed form a disulfide bridge, this bond must break in order for the thioether to form, and C385 is then a free cysteine; (3) the fragmentation data are incorrect and do not reliably capture the disulfide bridges of the native protein, and the free cysteine in the catalytic core of the nitrilase domain (C196) is actually responsible for the photoprotein activity.

DISCUSSION

Catalytic structure predictions

Assuming the overall structure is similar between vanin-1 and symplectin, the conserved catalytic triad is located in the nitrilase domain of symplectin (E60, K163, C196), while the cysteine for the thioether linkage is found in the base domain. Therefore, catalytic binding pocket of nitrilase domain is not responsible for the bioluminescence activity of symplectin, which is instead likely carried out by residues in close spatial proximity to C390. Although two other luciferases have solved crystal structures (*Loening, Fenn & Gambhir, 2007*; *Tomabechi et al., 2016*), these were unbound forms so mechanistic generalizations cannot be made.

Nonetheless, three phenylalanines (F316, F321, and F323) and one tyrosine (Y359) are predicted to be close to C390, and may have a role in coordinating the binding of coelenterazine, although only the tyrosine is well-conserved outside of symplectin. Besides the non-polar residues, several other residues predicted to be in proximity to C390 may have a role in the catalytic activity, including D314, K325, and E357. All three are conserved in one protein each from *P. hoylei* and *W. scintillans*, which are the two closest proteins to symplectin in the tree. However, given that the overall conservation is low, even proteins evolutionarily close to symplectin may not have any luminescence activity.

It was also noted that the native vanin-1 structure is a homodimer (*Boersma et al., 2014*). Head-to-tail arrangement of the two domains within a homodimer would allow another alternative, where the nitrilase domain of one monomer catalyzes the oxidation of dhCtz bound to the other monomer. In this case, there may be no need for any residues directly surrounding C390 to have a role in the enzymatic activity.

Evolution of the protein families

Interpreting the phylogenetic tree and determining the ancestral number of genes is challenging given the copy number in some animal lineages relative to the tree position. For most cases, such as the majority of arthropods, lineage-specific duplications have created multiple copies, the most being eight homologs in the spider *Parasteatoda tepidariorum*. The same is seen for the two polychaetes and other non-cephalopod molluscs. However, cnidarians (only anemones and corals represented) and chordates have two separate protein groups. For chordates, pantetheinase and biotinidase each have their own group, and although the copy number relative to the cephalochordate outgroup *Branchiostoma floridae* suggests a duplication specific to vertebrates, the tree does not indicate this arrangement with strong support. For cnidarians, both groups have unknown functions and are not monophyletic. Clearly a duplication must have occurred to give rise to the two groups, although the copy number in other groups (such as Porifera or Placozoa) is inconsistent with a duplication in the common ancestor of all animals. Two scenarios may be considered. The two cnidarian groups resulted from a duplication specific to cnidarians even though the tree (Fig. 2) does not indicate a single origin, suggesting that this protein family is poorly resolved by existing phylogenetics programs. The same would therefore also be true for the two vertebrate proteins. Alternatively, multiple bilaterian phyla (molluscs,

annelids, arthropods, echinoderms) must have lost at least one of the ancient copies that was otherwise retained by chordates.

Given the ubitquity of both biotin and pantothenic acid, metabolism of one or both of these molecules is likely to be the ancestral function. Biotinidase activity was detected in *Drosophila melanogaster* (*Swango & Wolf, 2001*), providing evidence that the monophyletic group of arthropod proteins are biotinidases, and possibly all bilaterian members of this family. This would therefore suggest several points about the evolution of this protein family. First, one or many of the cephalopod proteins may still have biotinidase activity or act as a hydrolase in other contexts via the nitrilase domain. Since symplectin is therefore predicted to have two separate functional domains, it is possible that symplectin performs multiple functions, and may even still have a role in biotin metabolism. Secondly, if the ancestral function is a biotinidase, then the pantetheinase activity of vanin-1 and related proteins found only in vertebrates therefore is a derived function.

Among the four cephalopod protein groups, homologs of the symplectin group were still found in several non-luminous species, *Sepia pharaonis*, *Loligo vulgaris*, and *Doryteuthis pealei*. For this reason, it could be that only a very small set of symplectin homologs are luminous because of key amino acid changes in the protein. While we were not able to find all four homologs in the transcriptomes of most species, we found five in *W. scintillans*, where two were in the symplectin group. Of these, one was highly similar to symplectin, including a symplectin-specific indel of two amino acids found in only one other protein (in *P. hoylei*). Candidate luciferases were already identified from *W. scintillans* (*Gimenez et al., 2016*) belonging to another protein family, so it is unclear what the role would be of the symplectin homologs identified here, or whether they have photoprotein activity with dhCtz.

Bioluminescence in cephalopods

Phylogenetic analysis of cephalopods suggests a complex pattern of gains and losses of bioluminescence (*Lindgren et al., 2012*), where a luminous phenotype appears to have been acquired five times, and pelagic species were significantly more likely to display autogenic bioluminescence. In addition to the two independent gains of bacterial bioluminescence, there are a total of seven inventions of bioluminescence in this class. As the protein mechanisms are only known for two species, *W. scintillans* and *S. oualaniensis*, is it possible that several other luciferases have evolved in this family.

Convergent evolution of bioluminescence is comparably commonplace within other animal groups. For instance, cnidarians are likely to have at least five separate evolutionary events of bioluminescence, in octocorals, deep-sea anemones (*Johnsen et al., 2012*), coronate and semaeostome scyphomedusae (*Shimomura et al., 2001*) and hydromedusae, though the proteins responsible have only been identified in octocorals (*Renilla*-type luciferases) and hydromedusae (calcium-activated photoproteins). Thus, even within the same clade, multiple separate origins of bioluminescence making use of the same luciferin with different proteins has already happened at least once in metazoans.

For the proteins themselves, three inventions of luciferases have been found in family of adenylating enzymes: for fireflies, the New Zealand glow worm *Arachnocampa luminosa*

(*Sharpe et al., 2015*), and the squid *W. scintillans* (*Gimenez et al., 2016*). This may suggest that members of this protein family has some propensity for becoming luciferases. Because many luciferins are hydrophobic molecules, enzymes that already catalyze reactions on other hydrophobic molecules (such as fatty acids) may have some innate affinity for luciferins, and then mutations that allow binding of oxygen ultimately change the function of these enzymes into luciferases.

Future directions

The case of convergent evolution of luciferases from adenylating enzymes was essentially unpredictable. For this reason, it is just as plausible to consider that none of the symplectin homologs here can function as luciferases as it is to consider that all of them function this way. Cloning and characterizing all symplectin-group homologs *in vitro* may resolve this, but does not effectively account for the possibility where luciferase activity independently evolved twice from the same protein family. Thus, if it were necessary to clone all members of a protein family to check for luciferase activity, this may be prohibitively time consuming. For this reason, a more directed approach of proteomic investigations of bioluminescent material from certain species may prove to be more useful, particularly for a species like the vampire squid, which has a secreted bioluminescence (*Robison et al., 2003*) and would be comparably easy to acquire concentrated luminous material. Such studies typically require a reference genome or transcriptome, so potentially the transcriptomes presented here could serve as a reference for these or closely related species. Further investigation of the mechanisms of bioluminescence in this animal clade may reveal more general principles of protein evolution, and how a few amino acid changes may have dramatic effects on the phenotype and biology of an organism.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by NIH grant NIGMS-5-R01-GM087198 to S.H.D.H. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
NIH: NIGMS-5-R01-GM087198.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Warren R. Francis conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables.
- Lynne M. Christianson performed the experiments.

- Steven H.D. Haddock conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables.

Field Study Permissions

The following information was supplied relating to field study approvals (i.e., approving body and any reference numbers):

Operations were conducted under permit SC-4029 issued to SHD Haddock by the California Department of Fish and Wildlife.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Reads for all five transcriptomes are at NCBI SRA with accessions SRR5527414–SRR5527418.

Data Availability

The following information was supplied regarding data availability:

All assemblies, alignments, and intermediate files are deposited at BitBucket: https://bitbucket.org/wrf/squid-transcriptomes.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3633#supplemental-information.

REFERENCES

Albertin CB, Simakov O, Mitros T, Wang ZY, Pungor JR, Edsingergonzales E, Brenner S, Ragsdale CW, Rokhsar DS. 2015. The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524(7564)**:220–224.

Alon S, Garrett SC, Levanon EY, Olson S, Graveley BR, Rosenthal JJC, Eisenberg E. 2015. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. *eLife* **4**:1–17 DOI 10.7554/eLife.05198.

Alva V, Nam S-Z, Söding J, Lupas AN. 2016. The MPI bioinformatics toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Research* **44(W1)**:W410–W415 DOI 10.1093/nar/gkw348.

Boersma YL, Newman J, Adams TE, Cowieson N, Krippner G, Bozaoglu K, Peat TS. 2014. The structure of vanin 1: a key enzyme linking metabolic disease and inflammation. *Acta Crystallographica Section D* **70(12)**:3320–3329 DOI 10.1107/S1399004714022767.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**:421 DOI 10.1186/1471-2105-10-421.

Francis WR, Christianson LM, Kiko R, Powers ML, Shaner NC, Haddock SHD. 2013. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* **14(1)**:167 DOI 10.1186/1471-2164-14-167.

Fujii T, Ahn JY, Kuse M, Mori H, Matsuda T, Isobe M. 2002. A novel photoprotein from oceanic squid (*Symplectoteuthis oualaniensis*) with sequence similarity to mammalian carbon-nitrogen hydrolase domains. *Biochemical and Biophysical Research Communications* **293(2)**:874–879 DOI 10.1016/S0006-291X(02)00296-6.

Gimenez G, Metcalf P, Paterson NG, Sharpe ML. 2016. Mass spectrometry analysis and transcriptome sequencing reveal glowing squid crystal proteins are in the same superfamily as firefly luciferase. *Scientific Reports* **6**:27638 DOI 10.1038/srep27638.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, Di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29(7)**:644–652 DOI 10.1038/nbt.1883.

Haddock SH, Moline MA, Case JF. 2010. Bioluminescence in the sea. *Annual Review of Marine Science* **2(1)**:443–493 DOI 10.1146/annurev-marine-120308-081028.

Head JF, Inouye S, Teranishi K, Shimomura O. 2000. The crystal structure of the photoprotein aequorin at 2.3 A resolution. *Nature* **405(6784)**:372–376 DOI 10.1038/35012659.

Hunt ME, Modi CK, Aglyamova GV, Ravikant DVS, Meyer E, Matz MV. 2012. Multi-domain GFP-like proteins from two species of marine hydrozoans. *Photochemical & Photobiological Sciences* **11(4)**:637–644 DOI 10.1039/c1pp05238a.

Inoue S, Sugiura S, Kakoi H, Hasizume K, Goto T, Iio H. 1975. Squid bioluminescence II. Isolation from Watasenia scintillans and synthesis of 2-(p-hydroxybenzyl)-6-(p-hydroxyphenyl)-3, 7-dihydroimidazo [1, 2-a] pyrazin-3-one. *Chemistry Letters* **4(2)**:141–144 DOI 10.1246/cl.1975.141.

Isobe M, Kuse M, Tani N, Fujii T, Sobe BMI, Use MK, Ani ÃNT. 2008. Cysteine-390 is the binding site of luminous substance with symplectin, a photoprotein from Okinawan squid, *Symplectoteuthis oualaniensis*. *Proceedings of the Japan* **84**:386–392 DOI 10.2183/pjab/84.386.

Johnsen S, Frank TM, Haddock SHD, Widder EA, Messing CG. 2012. Light and vision in the deep-sea benthos: I. Bioluminescence at 500–1,000 m depth in the Bahamian Islands. *Journal of Experimental Biology* **215(19)**:3335–3343 DOI 10.1242/jeb.072009.

Jung H, Pena-Francesch A, Saadat A, Sebastian A, Kim DH, Hamilton RF, Albert I, Allen BD, Demirel MC. 2016. Molecular tandem repeat strategy for elucidating mechanical properties of high-strength proteins. *Proceedings of the National Academy of Sciences of the United States of America* **113(23)**:6478–6483 DOI 10.1073/pnas.1521645113.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30(4)**:772–780 DOI 10.1093/molbev/mst010.

Kongjinda V, Nakashima Y, Tani N, Kuse M, Nishikawa T, Yu CH, Harada N, Isobe M. 2011. Dynamic chirality determines critical roles for bioluminescence in symplectin-dehydrocoelenterazine system. *Chemistry-An Asian Journal* **6(8)**:2080–2091 DOI 10.1002/asia.201100089.

Lindgren AR, Pankey MS, Hochberg FG, Oakley TH. 2012. A multi-gene phylogeny of Cephalopoda supports convergent morphological evolution in association with multiple habitat shifts in the marine environment. *BMC Evolutionary Biology* **12(1)**:129 DOI 10.1186/1471-2148-12-129.

Loening AM, Fenn TD, Gambhir SS. 2007. Crystal structures of the luciferase and green fluorescent protein from Renilla reniformis. *Journal of Molecular Biology* **374(4)**:1017–1028 DOI 10.1016/j.jmb.2007.09.078.

Pankey MS, Minin VN, Imholte GC, Suchard MA, Oakley TH. 2014. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proceedings of the National Academy of Sciences of the United States of America* **111(44)**:E4736–E4742 DOI 10.1073/pnas.1416574111.

Powers ML, McDermott AG, Shaner NC, Haddock SHD. 2013. Expression and characterization of the calcium-activated photoprotein from the ctenophore Bathocyroefosteri: insights into light-sensitive photoproteins. *Biochemical and Biophysical Research Communications* **431(2)**:360–366 DOI 10.1016/j.bbrc.2012.12.026.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5(3)**:e9490 DOI 10.1371/journal.pone.0009490.

Robison BH, Reisenbichler KR, Hunt JC, Haddock SHD. 2003. Light production by the arm tips of the deep-sea cephalopod Vampyroteuthis infernalis. *The Biological Bulletin* **205(2)**:102–109 DOI 10.2307/1543231.

Sharpe ML, Dearden PK, Gimenez G, Krause KL. 2015. Comparative RNA seq analysis of the New Zealand glowworm Arachnocampa luminosa reveals bioluminescence-related genes. *BMC Genomics* **16**:825 DOI 10.1186/s12864-015-2006-2.

Shimomura O, Flood PR, Inouye S, Bryan B, Shimomura A. 2001. Isolation and properties of the luciferase stored in the ovary of the scyphozoan medusa Periphylla periphylla. *Biological Bulletin* **201(3)**:339–347 DOI 10.2307/1543612.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9)**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Swango KL, Wolf B. 2001. Conservation of biotindase in mammals and identification of the putative biotinidase gene in drosophila melanogaster. *Molecular Genetics and Metabolism* **74(4)**:492–499 DOI 10.1006/mgme.2001.3244.

Takahashi H, Isobe M. 1994. Photoprotein of luminous squid, *symplectoteuthis oualaniensis* and reconstruction of the luminous system. *Chemistry Letters* **23(5)**:843–846 DOI 10.1246/cl.1994.843.

Tanaka E, Kuse M, Nishikawa T. 2009. Dehydrocoelenterazine is the organic substance constituting the prosthetic group of pholasin. *ChemBioChem* **10(Scheme 2)**:2725–2729 DOI 10.1002/cbic.200900503.

Tomabechi Y, Hosoya T, Ehara H, Sekine SI, Shirouzu M, Inouye S. 2016. Crystal structure of nanoKAZ: the mutated 19 kDa component of Oplophorus luciferase catalyzing the bioluminescent reaction with coelenterazine. *Biochemical and Biophysical Research Communications* **470(1)**:88–93 DOI 10.1016/j.bbrc.2015.12.123.

Wen J, Zhong H, Xiao J, Zhou Y, Chen Z, Zeng L, Chen D, Sun Y, Zhao J, Wang F. 2016. A transcriptome resource for pharaoh cuttlefish (Sepia pharaonis) after ink ejection by brief pressing. *Marine Genomics* **28**:53–56 DOI 10.1016/j.margen.2016.05.005.

Zhang X, Mao Y, Huang Z, Qu M, Chen J, Ding S, Hong J, Sun T. 2012. Transcriptome analysis of the Octopus vulgaris central nervous system. *PLOS ONE* **7(6)**:1–11 DOI 10.1371/journal.pone.0040320.