UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DISCOVERIES ON THE CHEMICAL AND GENETIC BASES OF BIOLUMINESCENCE IN GELATINOUS ZOOPLANKTON

A dissertation submitted in partial satisfaction of the
requirements for the degree of

Doctor of Philosophy

in

OCEAN SCIENCES

by

Warren Russell Francis

June 2014

The Dissertation of Warren Russell Francis
is approved:

Doctor Steven H. D. Haddock, Chair

Professor Jonathan P. Zehr

Professor Casey W. Dunn

Dean Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Discoveries on the chemical and genetic bases of bioluminescence in gelatinous zooplankton

by

Warren Russell Francis

In this thesis I will discuss three projects aimed to explore different aspects of bioluminescence and genetics in marine animals. The first project describes a series of experiments on the chemistry in the novel bioluminescent system of the marine worms of the genus *Tomopteris*. These luminous worms release glowing exudate when agitated, and this exudate was rich in a fluorescent pigment. The structure was determined to be an anthraquinone and the possible origins and chemical roles are discussed. The second part examines some technical aspects of transcriptome assembly and analysis for invertebrates, including many bioluminescent species. Because information content is theoretically finite yet noise from sequencing errors is introduced continuously, the optimal balance of sequencing depth is experimentally addressed and described with analysis strategies. The final part presents a detailed gene analysis of a group of putative oxidase genes which are strongly conserved across a group of luminous ctenophores and absent in genomes and transcriptomes of non-luminous ctenophores. This class of oxidases is known for its functional diversity in bacteria and fungi, and their occurrence and roles in ctenophores are discussed with relevance to bioluminescence.

Acknowledgments

The text of this thesis includes reprints of the following published material:

Francis, W.R., Powers, M.L.P, S.H.D. Haddock (2014) Characterization of an anthraquinone fluor from the bioluminescent, pelagic polychaete *Tomopteris*. *Luminescence* (online early-edition); WRF and SHDH designed experiments and analyzed data. WRF, MLP, and SHDH caught animals. MLP and SHDH acquired the bioluminescence spectrum. WRF did the experiments. WRF wrote the paper with corrections from the other authors.

Francis, W.R., L.M. Christianson, R. Kiko, M.L. Powers, N.C. Shaner, and S.H.D. Haddock (2013) A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* 14:167. WRF, RK and SHDH designed experiments. LMC, RK, MLP and SHDH caught animals. LMC, RK, MLP and NCS processed animals and extracted RNA. WRF assembled transcriptomes. WRF, RK and SHDH analyzed data. WRF wrote the paper.

I would like to thank J. Maitin-Shepard for help with Python optimization, R. Linington for helpful discussions and techincal advice on natural product discovery, Meghan Powers, Nathan Shaner and Lynne Christianson for help and expert advice with the molecular biology, Joe Ryan and Christine Schnitzler for suggestions with the ctenophore transcriptomes, and Casey Dunn and Mark Howison for preliminary versions of the Agalma package.

Chapter 1

Introduction

The study of bioluminescence is complex process, bringing together aspects of biology, chemistry and ecology. There are several examples of luminous organisms in terrestrial environments, including fireflies, glowworms, millipedes, other luminous insects and even mushrooms. While these cases are uncommon on land, in the marine environment bioluminescence is the rule rather than the exception. Reports from scientists and mariners dating back centuries have noted light-producing animals in the water. Luminous animals are ubiquitous in the oceans, and the ability to create light occurs across many phyla of unrelated organisms.

The modern body of knowledge on the chemistry of bioluminescence started with experiments from over 100 years ago. Toward the end of the 19th century, Raphaël Dubois had studied many luminous animals and is credited with the discovery of the classical luciferase-luciferin reaction, an enzyme-substrate reaction, as well as inventing those terms which are still in use. Following this, substantial work was done by Edmund

Newton Harvey, both on the chemistry and the classification of luminous animals, and is known for his discovery that many animals use different luciferins. However, even while such an observation may seem obvious now, researchers are continually surprised when unrelated animals are found to use the same luciferin. Finally, in the latter half of the 20th century, Osamu Shimomura became an important figure in the study of the chemistry of bioluminescence. His work included purification of both *Cypridina* luciferin and coelenterazine, two structurally-related marine luciferins, as well as detailed studies on the *Aequorea* luminescence system which led to the discoveries of both calcium-activated photoproteins and the world-famous green fluorescent protein.

Although there are still many questions to be answered, a few of them have been addressed in the present work. When examining a new bioluminescence system, years can pass between the discovery of a new luciferin and the determination of its structure. In this respect, Chapter 2 will discuss the chemical characterization of a fluorescent compound from the pelagic worms in the genus *Tomopteris*, as published previously [26]. With only cursory chemical investigations during the past century, the mechanisms of the bioluminescence in the genus *Tomopteris* remain a mystery. One report states that homogenates from the animal did not produce light with *Cypridina* luciferin, at the time suggesting the possibility that a novel luciferin is used. Work by E. N. Harvey noted that there was a bright yellow pigment visible in the parapodia. Following that, a commonly discussed connection between the light emitters in fluorescence and bioluminescence had therein prompted B. Terio to examine two fluorescent compounds in parapodia of the *Tomopteris*, one appearing yellow-green, the

other yellow-orange. His detailed observation under the microscope revealed that the yellow-orange fluorescent material was located near the photocytes (light-emitting cells) and had a fluorescence emission maximum between 550 and 570nm when excited from ultraviolet light. It was speculated that this compound might be involved in the luminous reaction, though was never characterized further. In this project, I purified and identified the fluorescent yellow-orange compound from whole animals and determine the structure, which is the anthraquinone aloe-emodin.

Chapter 3 will discuss some fundamental qualities of *de novo* transcriptome assembly, as published previously [25]. Advances in sequencing have enabled routine acquisition of enormous quantities of sequencing data from mRNA, called RNA-seq. This is useful for studying both changes in expression of genes, but also for acquisition of a minimum set of genes from rare organisms, such as many of those from the deep sea. For studies on organisms with sequenced genomes, numerous programs have been created to resolve individual transcript sequences by mapping the reads onto the genomes. Without a reference genome, the typically-short sequences need to be stitched together to form "contigs", or long, contiguous sequences, in a process called *de novo* assembly. In this project, I discussed the limitations of *de novo* assembly as measured by transcript number, bulk statistics on length, and gene content, to ultimately advise on the optimal sequencing depth for these types of RNA-seq projects.

Lastly, Chapter 4 will discuss the comparative gene content among groups of luminous and non-luminous ctenophores to identify candidate genes involved in the biosynthesis of the luciferin coelenterazine. Coelenterazine is the most widely used sub-

strate for bioluminescence in the marine environment, however its origins are unknown as many species that use it appear to get the molecule from their diets. Several scientific reports through the ages have pointed to ctenophores as a likely candidate for producing the molecule because they are luminous for their entire life cycle. Examining the genomes of 2 ctenophores and the transcriptomes of 22 other species, the most promising candidates are a group of non-heme oxidases that are highly conserved across the phylum. These genes are members of a superfamily of proteins known for their heterocyclic chemistries, and, most unusually, the protein itself contains a motif which is expected to cyclize into the luciferin coelenterazine. A thorough search revealed that the proteins with the hypothetical pro-luciferin motif only occur in luminous species of ctenophores. Detailed examination of the transcriptomes and genome of the two non-luminous control species also indicates that they are missing photoproteins, suggesting that one reason they are non-luminous is the lack of photoproteins which may be connected to losses of other genes involved in luminescence.

Chapter 2

Characterization of an anthraquinone fluor from the bioluminescent, pelagic polychaete *Tomopteris*

2.1 Abstract

Tomopteris is a cosmopolitan genus of polychaetes. Many species produce yellow luminescence in the parapodia when stimulated. Yellow bioluminescence is rare in the ocean and the components of this luminescent reaction have not been identified. Only a brief description half a century ago noted a fluorescence in the parapodia with a remarkably similar spectrum to the bioluminescence, which suggested that it may be the luciferin or terminal light-emitter. Here we report the isolation of the fluorescent yellow-orange pigment found in the luminous exudate and in the body of the animals. LCMS revealed the mass to be 270m/z with a molecular formula of $C_{15}H_{10}O_5$, which

ultimately was shown to be aloe-emodin, an anthraquinone previously found in plants. We speculate that aloe-emodin could be a factor for resonant-energy transfer or the oxyluciferin for *Tomopteris* bioluminescence.

2.2 Background

The ocean is rife with luminous animals, most of which emit blue light [34,40]. An exception is the annelid worms in the genus *Tomopteris*, a group of pelagic polychaetes of which several species are reported to produce yellow bioluminescence [37]. When agitated, these animals can release glowing material into the water that persists for several seconds. The yellow luminescence of *Tomopteris* has been known for some time [18, 36], yet is unstudied when compared to bacterial, beetle, or cnidarian systems. It was reported by Harvey that homogenates from the polychaete did not show a luciferin-luciferase type reaction nor did they produce light with the ostracod luciferin (*Cypridina* luciferin), suggesting the possibility that a previously uncharacterized luciferin is used [37]. Shimomura [87] also performed some preliminary investigations into the yellow bioluminescence.

The connection between oxyluciferin fluorescence and the bioluminescence has been described for several systems including cnidarians and ctenophores, the firefly, and luminous bacteria [40]. For cnidarians, notably *Aequorea*, the bioluminescence spectrum was identical to the fluorescence spectrum of the photoprotein following the bioluminescence reaction, that is, coelenteramide bound by the photoprotein [89]. In

the case of the firefly, the bioluminescence matches the fluorescence of the oxyluciferin, the oxidized product of the consumable substrate [95,96]. Similarly, in bacterial systems the bioluminescence spectra also matches the fluorescence of a flavin cation, which is oxidized in the reaction and later regenerated [21, 29, 107].

With this in mind, Terio had examined two fluorophores in *Tomopteris nationalis* specimens, one appearing yellow-green with ultraviolet excitation, the other yellow-orange [100, 101]. His detailed observation under the microscope revealed the yellow-orange fluorescent material was located near the photocytes (light-emitting cells), indicating a likely involvement in the bioluminescence. The material had a fluorescence emission maximum between 550 and 570nm, and appeared similar to the bioluminescence emission. The fluorescence was unchanged in non-polar solvents suggesting the compound was non-polar. Finally, Terio had speculated that this compound might be involved in the luminous reaction, possibly as the luciferin, but it was never characterized further. While few luciferins have been isolated, it is thought that bioluminescence evolved many times and novel chemistries may still be found [39]. Fewer than ten luciferins have been identified and the discovery and characterization of a novel luciferin would be a substantial advancement in the study of bioluminescence [34].

Here we report the isolation and characterization of the fluorescent yellow-orange material from whole Tomopterid specimens. We were able to obtain an accurate mass of the compound as well as the molecular formula. Through a comparison of literature spectra and by LCMS, we identified the compound as aloe-emodin (Figure 2.1), a polyhydroxyl-substituted anthraquinone. Finally we speculate on possible roles

based on known redox properties and chemiluminescence from other anthraquinones.

2.3 Results

2.3.1 Acquisition of raw material

Specimens were caught at depth by remotely operated vehicles (ROVs). This often enabled careful capture of very large specimens which could be returned to the lab in excellent condition. When agitated, luminescence begins in the parapodia and nearly all of our specimens released glowing material from their parapodia. To our knowledge, there is no mention in the literature of these animals releasing luminescent particles. We consider this may due to the majority, possibly all, of specimens in the literature being agitated or injured during capture by the plankton nets.

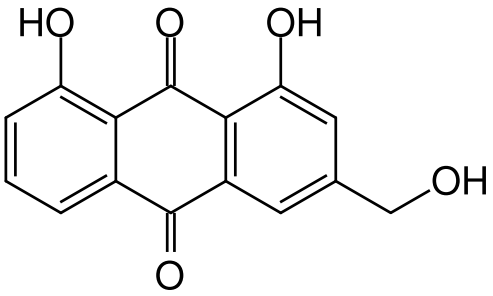

Figure 2.1: Structure of aloe-emodin

We acquired the bioluminescence spectra of the luminous exudate (shown in Figure 2.2A, λ_{max}: 565nm), which is in good agreement with the Atlantic species *Tomopteris nisseni* measured by Latz [53]. The bioluminescence spectrum also matches

perfectly with the digitized fluorescence spectrum of the yellow-orange fluor measured by Terio [101], to the extent that the image may be converted to a spectrum.

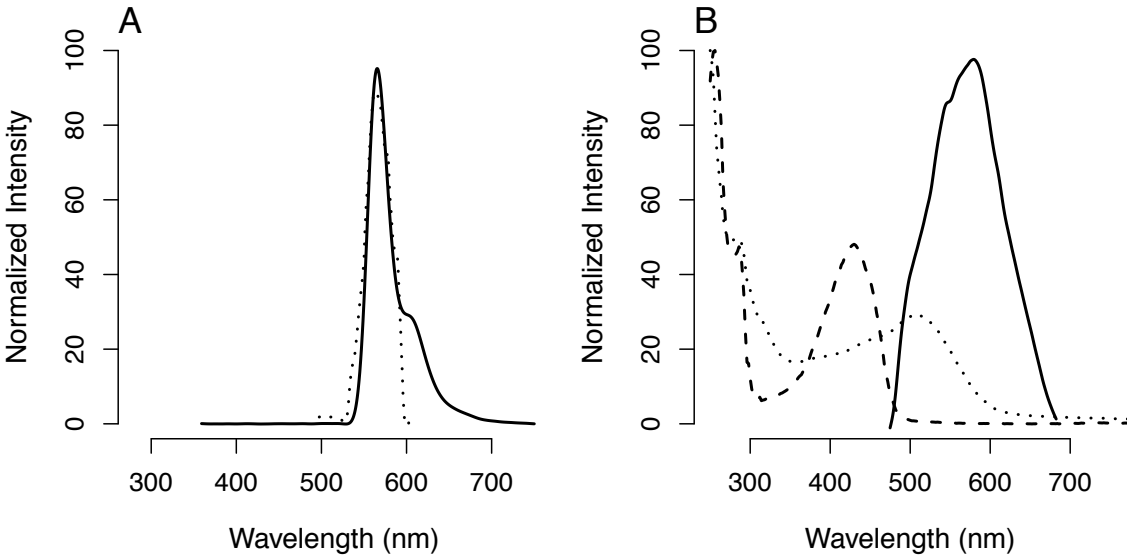

Figure 2.2: Absorption and bioluminescence spectra

(A) The *in vivo* bioluminescence spectra of *Tomopteris* (solid line) and the digitized *in vivo* fluorescence data from Terio [101] (dotted line). (B) Absorption spectra of the fluorescent pigment in methanol (dashed line) and with a drop of NaOH (dotted line), as well as the fluorescence emission spectrum in chloroform (solid line).

Because live specimens release glowing material, we reasoned that the light emitter could be isolated from the exudate. Luminous exudate has a bright yellow-orange fluorescence under blue light, however the quantity obtained was insufficient for further analysis. Whole animals displayed a bright yellow-green fluorescence around the coelom in the parapodia even when fixed or frozen (Figure 2.3A). This material was clearly visible as a bright yellow pigment in the parapodia for frozen specimens and was

seen even in specimens frozen for over 10 years. Due to the irregularity of acquiring new specimens at sea and collecting exudate, we instead extracted material from frozen specimens (see Methods).

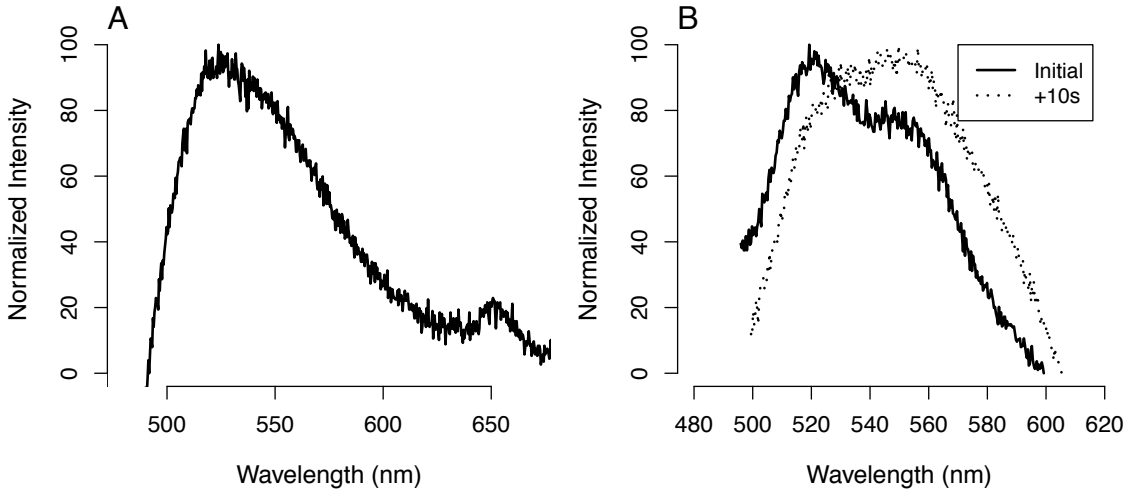

Figure 2.3: Additional fluorescence measurements

(A) Fluorescence of the parapodia and (B) Fluorescence emission from the MeOAc layer immediately after extraction and after 10 seconds of exposure to blue light (peak 460nm).

2.3.2 Non-polar extractions

Frozen specimens were homogenized, and methyl acetate was added to the homogenate. After centrifugation, nearly all of the fluorescent material was in the non-polar phase and appeared pale-yellow. The absorption spectrum of the non-polar phase showed a large peak at 364nm (Figure 2.4). The aqueous layer was dimly fluorescent green, likely due to riboflavin or a similar compound. Often, the fluorescence of the

methyl acetate layer appeared bright yellow-green immediately after extraction (λ_{max}: 519nm). When exposed to blue light, this changed to the characteristic yellow-orange color in seconds (Figure 2.3B). This effect was attenuated in the presence of ascorbic acid, suggesting that oxygen or reactive oxygen species could be involved in this transition.

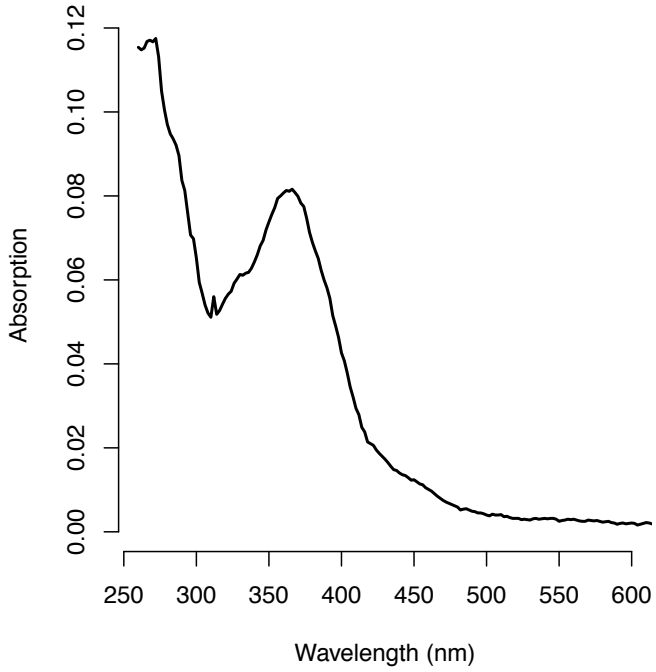

Figure 2.4: Absorption of methyl acetate extract

2.3.3 Purification of the yellow-orange compound by HPLC

This crude organic extract was separated by reversed-phase HPLC to isolate the fluorescent yellow-orange pigment (Figure 2.5A). Very large absorption peaks at

254nm and 430nm of a yellow material with a bright fluorescence peak were observed around 5.6 minutes (Figure 2.5B). This single peak was collected over multiple injections. The absorption (λ_{max}: 286, 430nm) and fluorescence emission (λ_{max}: 580, 548nm(shoulder)) were acquired for the purified compound (Figure 2.5B). The absorption peak of the purified compound is 430nm, however this does not appear to be abundant enough in the unpurified extract to show a distinct peak (Figure 2.4). Instead, it likely that some other pigment accounts for the peak at 364nm in the original methyl acetate extract. Although the fluorescence emission does not perfectly match the digitized spectrum reported by Terio or the bioluminescence (Figure 2.2B), [101] this may be due to the solvent or that the spectrum changes when bound by a protein, as seen for coelenterazine [43, 90, 92].

2.3.4 Mass determination and molecular formula

Knowing the absorption spectrum of the compound permitted easy mass determination of the compound with LCMS. The same methyl acetate extract was analyzed by LCMS, where the yellow compound was identified at 337m/z with the major fragment at 269m/z (Figure 2.6) which corresponded to a mass difference of 68m/z. To find the molecular formula and identities of the fragments, the accurate mass was determined for the purified compound at 337.0331m/z, corresponding to $C_{15}H_9O_5$+$NaCHO_2$ (M-H). It was then determined that the major fragment was actually the molecular ion, at 269.0455m/z which indicated the loss of the sodium formate adduct and the uncharged molecular formula of $C_{15}H_{10}O_5$ (Figures 4.7-4.9).

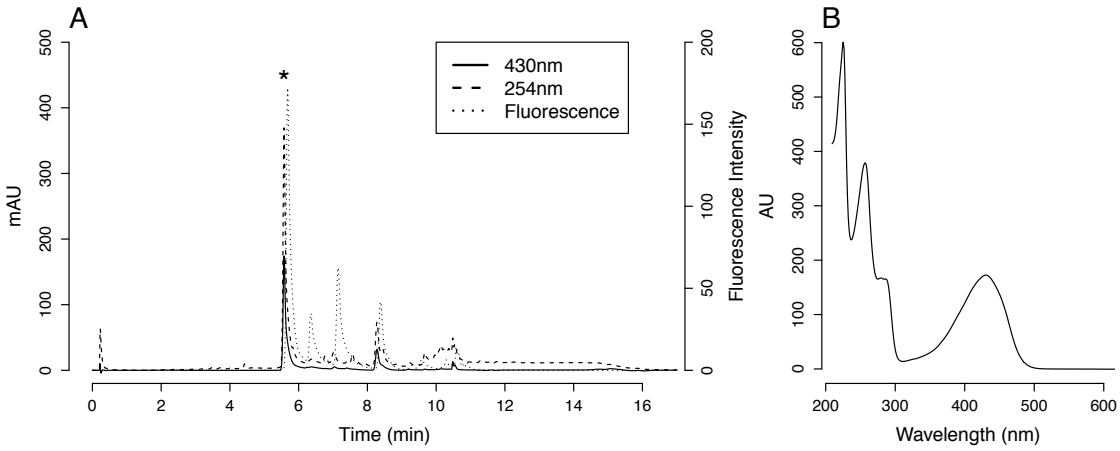

Figure 2.5: HPLC chromatogram of the MeOAc extract

(A) The UV-vis absorption (254 and 430nm) and fluorescence chromatograms show a large peak of the fluorescent yellow-orange compound at 5.6 minutes, indicated by the star. (B) The corresponding absorption spectrum at 5.6 minutes clearly showing the characteristic peak at 430nm.

2.3.5 Confirmation of the identity as aloe-emodin

The fluorescent material in methanol undergoes a bathochromic shift from yellow to red upon addition of saturated NaOH solution (Figure 2.2B, dotted line, λ_{max}: 510nm, also in Figure 4.10). The spectra and this transition are thought to be a property of 1,8-dihydroxy-9,10-anthraquinones [45]. After a comparison of our spectrum with 20 published UV-vis spectra of anthraquinones with the same molecular formula [104, 105], we noticed that our spectrum is remarkably close to the reported spectrum of aloe-emodin (structure in Figure 2.1) [108]. Aloe-emodin was purchased (Sigma-Aldrich) and was found to have an identical absorption spectrum as the yellow-

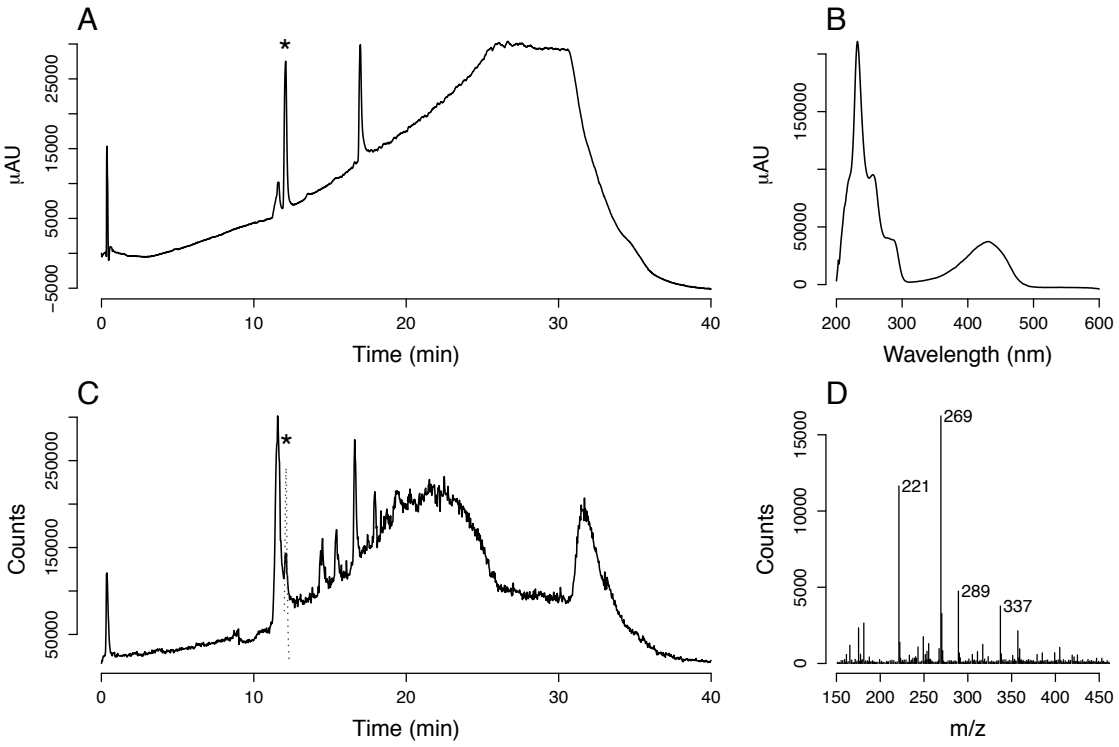

Figure 2.6: LCMS chromatogram

(A) The UV-vis absorption (254nm) chromatogram and (B) the corresponding absorption spectrum as indicated by the star in part (A). (C) Full MS-MS relative abundance chromatogram and the corresponding mass spectrum (D) of the same peak. The abundance of the 269m/z ion for the MS-MS is indicated by the dotted line in (C) and only occurs for the peak seen in parts (A) and (B); the values are multiplied by 40 to be visible on the graph.

orange fluor (Figure 4.10).

The product ion mass spectrum (MS-MS) is sometimes used to confirm the presence of rare metabolites for cases where NMR cannot be used to deduce the structure [54]. To ultimately confirm the identity of the compound, the HPLC-purified sample

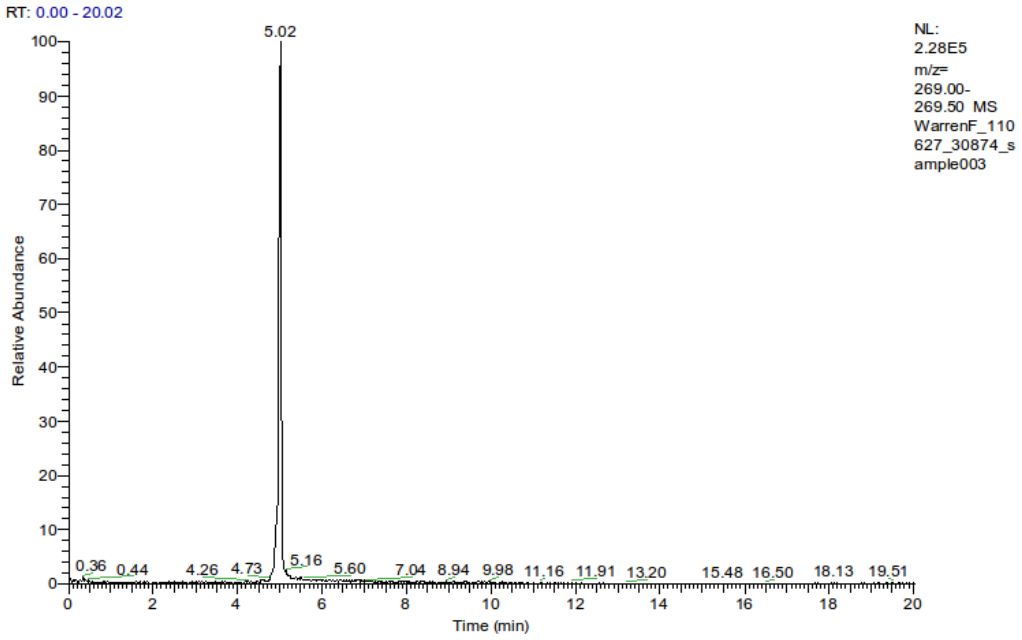

Figure 2.7: HR-LCMS chromatogram

Chromatogram of the relative abundance for the range of 269.00 to 269.50m/z. There is only one peak which corresponds to the yellow-orange fluor.

and a standard of aloe-emodin were sent out for analysis by LCMS. The retention time, the calculated and measured m/z ratios, and the product ion spectra were all identical matches, consistent with the hypothesis that the yellow-orange fluor is indeed aloe-emodin (Figures 4.11-4.13).

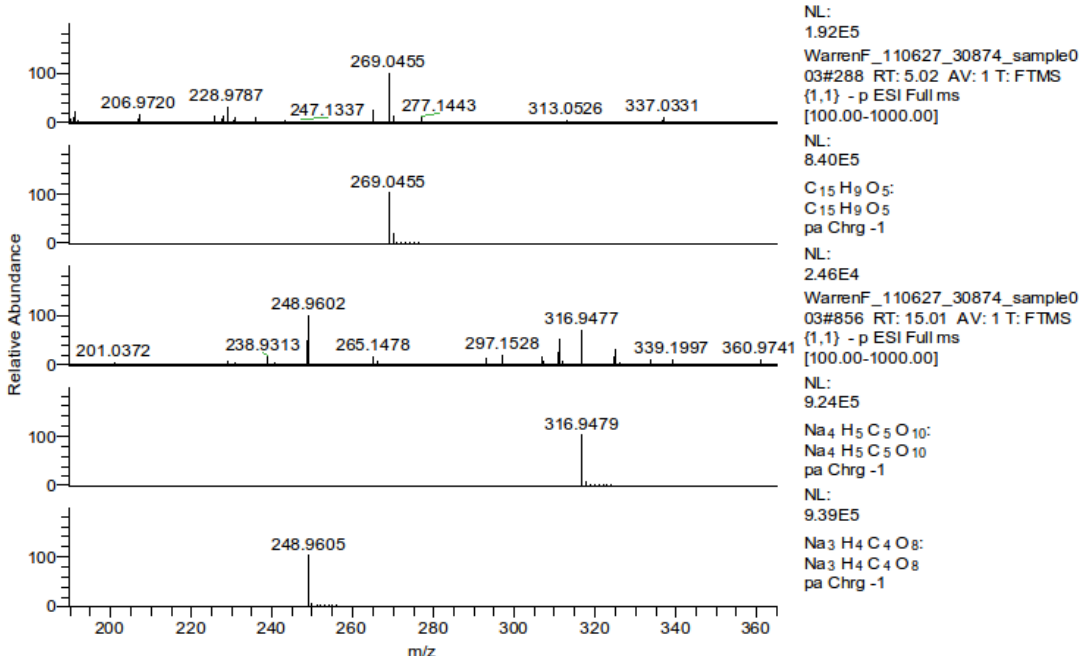

Figure 2.8: HR-Mass spectra

The panels from top to bottom indicate: the measured mass for the sample of yellow-orange fluor, model for $C_{15}H_9O_5$, measured sodium formate clusters, model for $Na_4(CHO_2)_5$, and the model for $Na_3(CHO_2)_4$. The models show that the peak at 337.0331m/z is not the molecular ion, but rather a cluster with sodium formate.

2.4 Discussion

2.4.1 Extraction yield

Here we described the extraction and identification of the yellow-orange fluor in the *Tomopteris* which was first noted over 50 years ago. As the mass and structure were only determined towards the end of our experiments, some questions related to extraction yields were unaddressed. However, estimated from the published extinction

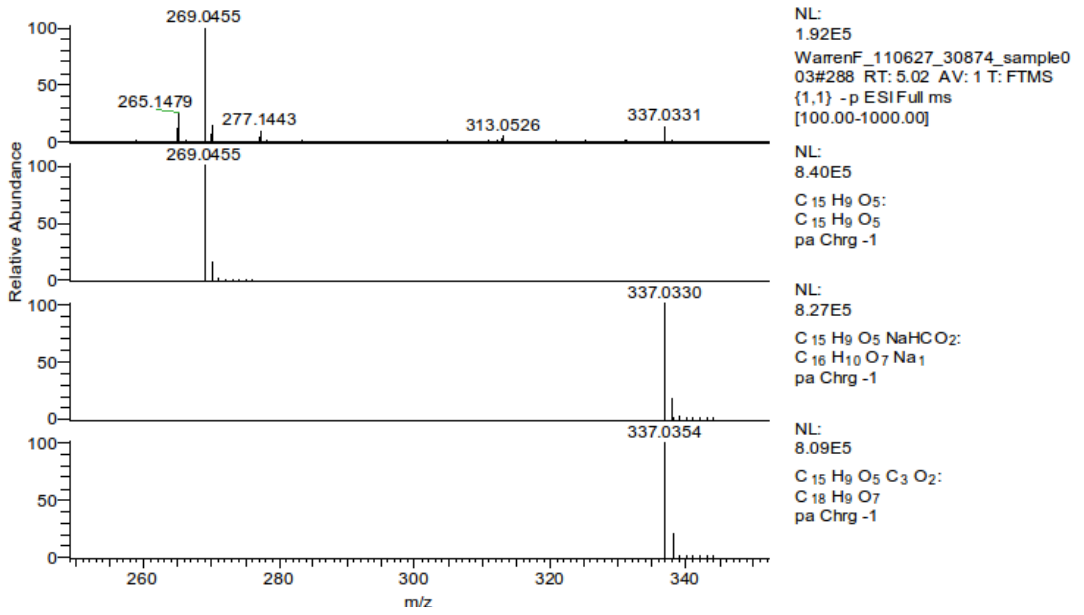

Figure 2.9: 269m/z compared with the model

The top panel shows the measured sample while the lower panel shows the model for $C_{15}H_9O_5$. The masses are identical, indicating that $C_{15}H_9O_5$ is the correct molecular formula.

coefficients of aloe-emodin, the HPLC data (from Figure 3) suggest that the single injection of 10μL contains on the order of 35μg of aloe-emodin. Because multiple HPLC runs were necessary to separate all the material and not saturate the column, we estimate that even a relatively small worm (3-5cm, estimated to be 200-500mg) could contain 200μg of aloe-emodin. Measurements of other *Tomopteris* specimens suggest dry material accounts for around 15% of the mass [14]. For a 500mg worm this means that dry mass accounts for 75mg, where 200μg of aloe-emodin is almost a third of a percent of the dry mass.

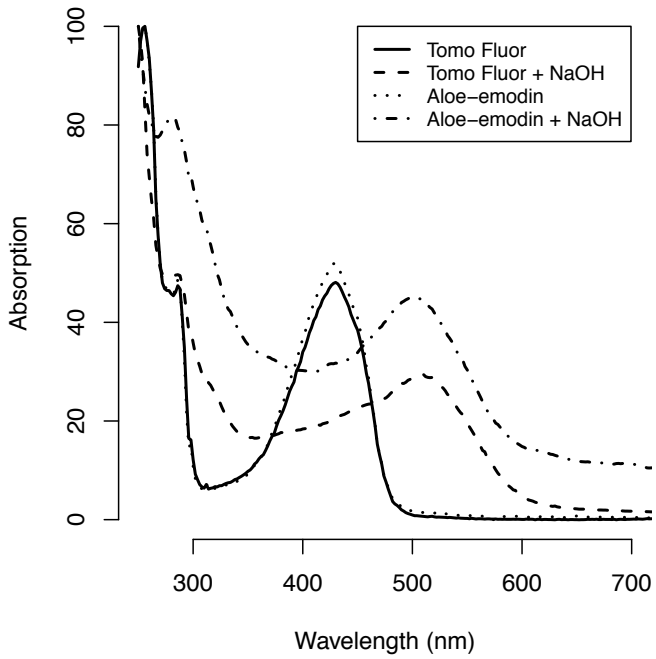

Figure 2.10: Spectra of aloe-emodin and the yellow-orange fluor

2.4.2 Functions of quinones

We have ultimately confirmed the compound to be aloe-emodin, but we do not know the function of aloe-emodin for this marine animal. Given that aloe-emodin is an anthraquinone it is logical that it is used similarly as other anthraquinones. There are a number of cases for insects where quinones and anthraquinones have been suggested to have various defensive roles, possibly as toxins [20, 45, 112]. Quinones also are known to participate in redox reactions, such as in the electron transport chain. Since all known bioluminescence reactions involve an oxidation [39], quinones are well suited for this type of chemistry. Aloe-emodin has been discussed in literature for both antioxidant

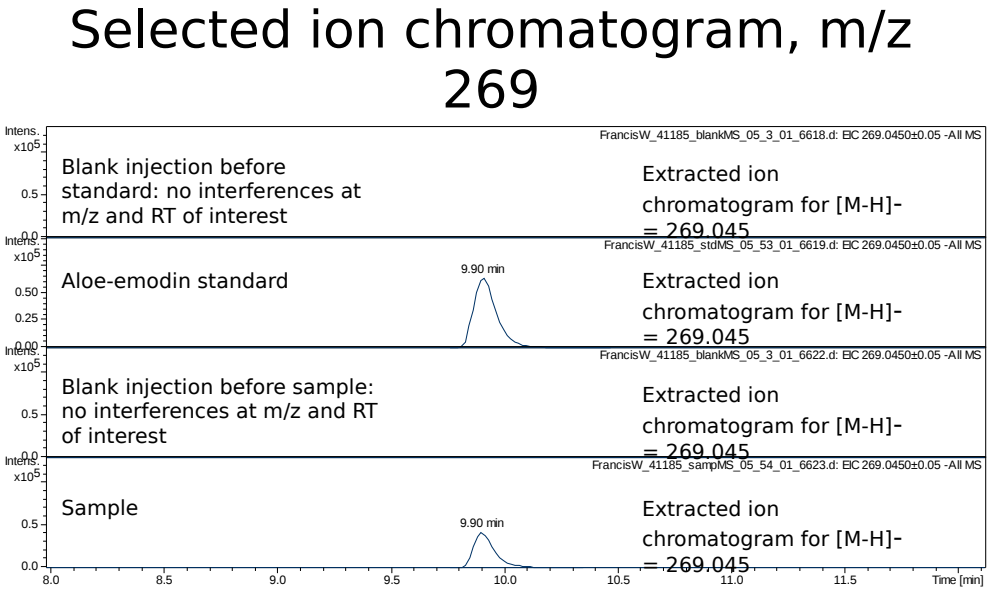

Figure 2.11: Selected ion chromatogram at 296m/z

The panels from top to bottom show: a blank injection prior to the standard without peaks at the relevant mass range or retention time, the aloe-emodin standard eluting at 9.90 minutes, a second blank before the sample without peaks for the mass range or time, the yellow-orange fluor sample with the identical retention time as the aloe-emodin standard.

and prooxidant properties, making a strong case for its role in this regard [58,106,116].

2.4.3 Quinones in other bioluminescent systems

Furthermore, there is a precedent of a quinone in bioluminescence from an unusual polybrominated benzoquinone that is used in the luminous system of the acorn

Full scan mass spectra

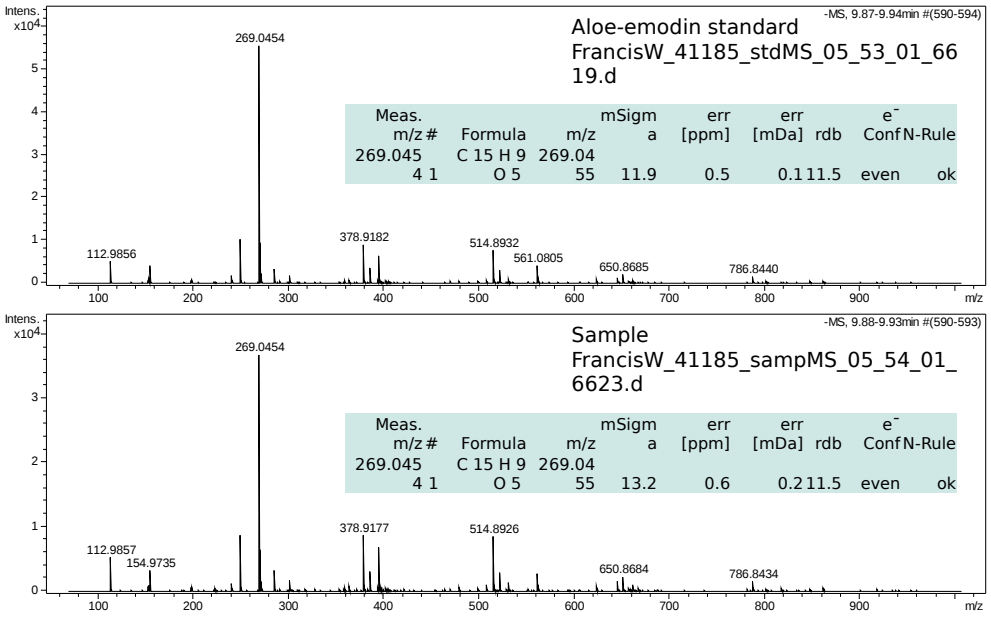

Figure 2.12: Full scan mass spectra

The upper panel shows the aloe-emodin standard for retention time range of 9.87-9.94 minutes. The lower panel shows the yellow-orange fluor sample over the retention time range of 9.88-9.93 minutes. The standard and the sample have identical peaks for the molecular ion at 269.0454m/z.

worm, *Ptychodera flava*, which also requires riboflavin [46, 47]. Given that the green color of the light of the acorn worm closely matches the fluorescence of riboflavin, it is possible that riboflavin is the light emitter and this benzoquinone serves as an electron carrier for the oxidation of riboflavin. Alternatively, the authors of that work had demonstrated that polybrominated quinones themselves were chemiluminescent,

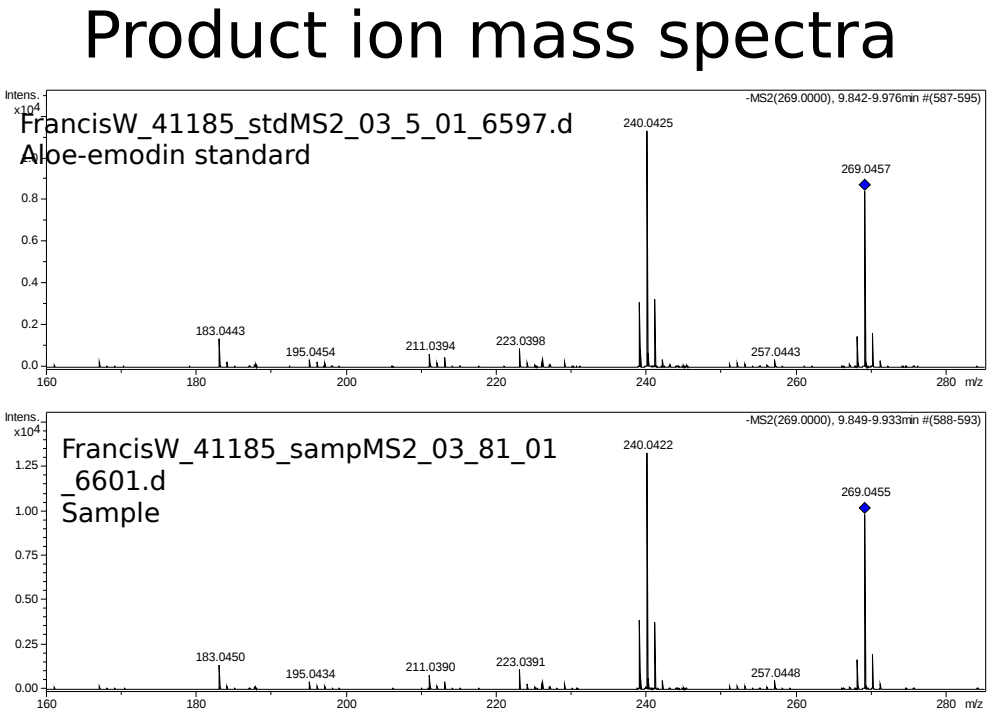

Figure 2.13: Product ion mass spectra

The upper and lower panels display the MS2 spectra for the 269m/z ion from the aloe-emodin standard and the yellow-orange fluor, respectively.

suggesting that perhaps riboflavin is only present as a fluor for resonant energy transfer to change the color of the emitted light.

2.4.4 Chemiluminescence of anthraquinones

Other anthraquinones have been shown to be chemiluminescent (λ_{max}: 568nm) when reduced to the hydroquinone or semiquinone and reacted with molecular oxygen [94]. Additionally, it was also shown that a semiquinone form was chemiluminescent (or

fluorescent) in yellow-green (λ_{max}: 515nm) [94]. We hypothesize that aloe-emodin, a substituted anthraquinone, would have very similar properties. In fact, our observations of a fluorescent yellow-green compound which transitions to aloe-emodin (where it is fluorescent yellow-orange) suggest the possibility that the yellow-green compound is a reduced form of aloe-emodin, possibly the anthrone which would be very susceptible to oxidation [23, 42]. If aloe-emodin were the oxyluciferin in this context, then plausibly the fluorescent yellow-green compound, the anthrone or a similar compound, could be the luciferin.

2.4.5 Past work on *Tomopteris* bioluminescence

The only modern characterization of *Tomopteris* luminescence suggested that chemiluminescence could be elicited from homogenate with superoxide ions [87], as seen for several other polychaetes [65, 88]. A large amount of Triton-X (2%) was needed to solubilize the light-emitter, suggesting that the enzyme may be a membrane-bound photoprotein [87]. However, we consider it is unlikely that the *in vivo* mechanism of light emission requires superoxide. For example, coelenterazine is chemiluminescent with superoxide yet the light output was an order of magnitude lower than the same quantity of coelenterazine bound to obelin and activated with calcium ions [57].

2.4.6 Theories of origins of aloe-emodin

It was surprising to find this compound in a deep-sea animal as the compound was discovered from several *Aloe* species. It is unknown whether the *Tomopteris*

synthesizes aloe-emodin or acquires it elsewhere, perhaps through its diet or from a symbiont. Many anthraquinones are biosynthesized through a convergent mechanism using polyketide synthases [6, 7], a mode that is conserved across prokaryotes, fungi and plants, thus any of those modes of acquisition may be possible. A dietary link from land plants would be preposterous; however, there are other cases of anthraquinones from marine organisms, [113] including a marine fungus which lives commensally with a green alga and appears to produce several anthraquinones and an isomer of aloe-emodin [41]. Another possibility is that a symbiont is generating the compound and there is some precedent of this scenario in metazoans. It was thought that some insects may synthesize their own polyketides [7], though one study had shown that the compounds were made by an uncultured bacterial symbiont [75]. To our knowledge there has not been a confirmed case of polyketide synthesis by metazoans. Although this does not rule out such a possibility, it suggests that the aloe-emodin from the *Tomopteris* may ultimately derive from another organism or involve biosynthetic mechanisms other than polyketide synthases.

2.5 Conclusions

From our detailed purification and LCMS, we have shown that the fluorescent yellow-orange compound is aloe-emodin. Evidence from the overlap of the fluorescence and bioluminescence spectra is very compelling to suggest that aloe-emodin is the final light-emitter for *Tomopteris* bioluminescence. While evidence from related systems

favors the interpretation that aloe-emodin is the oxyluciferin, this does not exclude the possibility that aloe-emodin is an acceptor for resonant energy transfer from another molecule. Detailed chemical studies are needed to discern these two cases. Ultimately, full characterization of the *Tomopteris* luminous system may lead to a new generation of bioluminescent sensors or reporters, particularly for plants or fungi where many anthraquinones are endogenous.

2.6 Methods

2.6.1 Samples

Tomopteris specimens were collected in the Monterey Bay using ROVs (remotely-operated vehicles) from 1999 to 2011. Many were caught previously and frozen in liquid nitrogen. The specimens were found between depths of 269m and 1316m, typically around 400m. Specimens varied considerably in size, from 3cm (∼0.5g, wet) to over 40cm (∼50g, wet). Polychaete taxonomists recognize that there are several undescribed species in these waters (E.V. Thuesen and K.J. Osborn, pers. comm.) and all tested species had the same luminescent properties, so no attempt was made to discern species. Condition and amount of extractable material was also variable, due to specimens often releasing luminous material prior to being caught or being damaged by the sampling apparatus.

2.6.2 Chemicals

Water for HPLC was purified by reverse-osmosis. All other solvents were HPLC grade and were purchased from Fisher Scientific. Aloe-emodin was purchased from Sigma-Aldrich.

2.6.3 Extractions

Luminous material was collected when released from live animals in a tube with gentle agitation. Frozen specimens were homogenized using a tissue grinder. The homogenate was divided into microfuge tubes and an equal volume of MeOAc was added to each tube, typically 1mL. This formed emulsions. The tubes were briefly vortexed and then centrifuged for 2 minutes at 16,000 xg. This separated the emulsion into three layers: aqueous, lipids and debris, and organic. The MeOAc layers (organic) were pooled and dried under vacuum at ambient temperature. The sample was reconstituted with three extractions of 20μL MeOH and transferred to a HPLC vial for injection.

2.6.4 Purification

HPLC was done using a Shimadzu Nexera system with a Hypersil Gold C18 column (50mm x 2.1mm, 1.9μm Thermo). Run parameters were: 1mL/min flow rate; binary gradient of H_2O:MeOH + 0.1% formic acid from 95:5 to 5:95 over 10 minutes; 60° C column temperature; 450nm fluorescence excitation; 548nm fluorescence detection; photo-diode array scans from 210nm to 800nm at 250 scans per minute.

2.6.5 Spectra

The fluorescence and *in vivo* bioluminescence spectra were acquired using a Ocean Optics QE65000 spectrometer with attached fiber optic. The associated Ocean Optics program SpectraSuite was used to collect spectra. The absorption spectra were measured in a 1mL cuvette on a Tecan Infinite 200 running Tecan i-control software. The digitized data from Terio (1960, 1964) were captured with ImageJ using the "Measure" and "Plot Profile" commands to generate a graph of the intensity across the photograph from the original papers.

2.6.6 Mass Analysis

Low-resolution mass was determined by LCMS using a Thermo Finnigan LC/MS (LTQ) electrospray ionization (ESI) mass spectrometer (Thermo, San Jose, CA). For the LC, a Hypersil Gold C18 column (50mm x 2.1mm, 1.9μm, Thermo) was used, and run parameters were: 0.5mL/min flow rate; binary gradient of H_2O:MeOH + 0.1% formic acid from 95:5 to 5:95 over 28 minutes; 60° C column temperature. For the MS: negative ionization mode (M-H); source voltage 5.0kV; mass range from 150.0 to 1000.0 m/z; photo-diode array range from 200nm to 600nm; normalized collision energy of 35% for MS/MS.

For accurate mass determination, the sample was dried under vacuum and sent out to the Vincent Coates Foundation Mass Spectrometry Laboratory at Stanford University (http://mass-spec.stanford.edu). The sample was reconstituted in 100μL of 1:1 H_2O:MeOH and sonicated for 10 minutes immediately prior to analysis.

For mass profile determination, another dried sample of the HPLC purified compound and a standard of aloe-emodin were sent for LCMS to the Vincent Coates Foundation Mass Spectrometry Laboratory at Stanford University. The sample and standard were reconstituted in 50μL 1:1 H_2O:MeOH, vortexed for 30 seconds then sonicated for 10 minutes. A portion was diluted 1:10 with H_2O:MeOH and transferred into an HPLC vial. For the LC, a Agilent C18 column (50x2.1 mm, 1.8μm) was used, with parameters: 0.2mL/min flow rate; binary gradient of H_2O:acetonitrile + 0.1% formic acid from 90:10 to 0:100 over 10 minutes. The mass was analyzed with a Bruker MicroTOF-QII Quadrupole Time of Flight Mass Spectrometer in negative ESI mode.

Chapter 3

A comparison across non-model animals suggests an optimal sequencing depth for *de novo* transcriptome assembly

3.1 Abstract

Background

The lack of genomic resources can present challenges for studies of non-model organisms. Transcriptome sequencing offers an attractive method to gather information about genes and gene expression without the need for a reference genome. However, it is unclear what level of sequencing depth is adequate to assemble the transcriptome *de novo* for these purposes.

Results

We assembled transcriptomes of animals from six different phyla (Annelids, Arthropods, Chordates, Cnidarians, Ctenophores, and Molluscs) at regular increments of reads using Velvet/Oases and Trinity to determine how read count affects the assembly. This included an assembly of mouse heart reads because we could compare those against the reference genome that is available. We found qualitative differences in the assemblies of whole-animals versus tissues. With increasing reads, whole-animal assemblies show rapid increase of transcripts and complete assembly of conserved genes, while single-tissue assemblies show a lower rate of assembly of conserved genes though the assembled transcripts were often longer. A deeper examination of the mouse assemblies shows that with more reads, assembly errors become more frequent but such errors can be mitigated with more stringent assembly parameters.

Conclusions

These assembly trends suggest that representative assemblies are generated with as few as 20 million reads for tissue samples and 30 million reads for whole-animals for RNA-level coverage. These depths provide a good balance between coverage and noise. Beyond 60 million reads, the discovery of new genes is low and sequencing errors of highly-expressed genes are likely to accumulate. Finally, siphonophores (polymorphic Cnidarians) transcriptomes are an exception and possibly require alternate assembly strategies.

3.2 Background

RNA-seq has provided a powerful tool for analysis of transcriptomes. For non-model organisms with limited genomic information, transcriptome sequencing provides a cost-saving tool by only sequencing functional and protein coding RNAs, thus providing direct information about the genes [110]. There are many benefits of sequencing a genome, but for relatively large genomes such as human and mouse, protein coding regions account for under 5%, thus most of the sequencing effort would go to sequencing either regulatory regions or repetitive elements [80]. Smaller genomes could be sequenced and assembled to complement the transcriptomes, though this is not a tractable approach if a genome is quite large. Moreover, *de novo* genome assembly can produce errors by itself [82].

Despite its advantage, transcriptome assembly does present additional challenges when compared to genome assembly. Unlike genomes where most sequences should be approximately equally represented, coverage of any given sequence in a transcriptome can vary over several orders of magnitude due to expression differences [8]. Because coverage can vary, there is also a question of sequencing depth. Theoretically, there is a sequencing depth beyond which addition of more reads does not provide new information, known as the saturation depth. Several studies have used approaches which map reads onto reference genomes and these have suggested saturation depths at 95% gene coverage ranging from 1.2 million reads to 50 million for mRNA level coverage,

and up to 700 million for splice variants [4, 15, 55]. However, these studies all made use of short reads around 36bp and were not assembling the transcriptomes *de novo*.

Several recent studies have already made use of next-generation sequencing technologies for *de novo* transcriptome assembly [2, 17, 22, 28, 60, 86, 109, 114]. The number of reads used for assembly in these studies varies widely, ranging from 2.6 million reads up to 106 million reads [28, 114]. The assembly strategies are equally varied, but share the initial step of removing low-quality reads and adapters whereupon all remaining reads are assembled. The assembly quality estimates vary as well with the most common measure of quality based on BLAST hits to public databases like Uniprot, though it was noted that under-representation of many taxa in public databases limits this approach [22].

While many parameters must be optimized for the specific assembly, it is both inconvenient and costly to acquire more reads by resequencing. Presently, there is no clear consensus of what sequencing depth is optimal or what factors would contribute to the adequate depth. The problems of omitted genes or variants are obvious with too few reads. On the other hand, it was suggested that greater depth may create errors in differential expression analyses, cost more, and take longer to assemble [98]. Thus, here we use the same assembly strategy across a diverse set of organisms to isolate the effects of read count on assembly quality to attain a general estimate of optimal read count. We compare trends from *de novo* assemblies across six phyla. These animals include

the mouse (used as a control for the non-model samples), the Humboldt squid *Dosidicus gigas*, the scaleworm *Harmothoe imbricata*, the decapod *Sergestes similis*, the copepod *Pleuromamma robusta*, the ctenophore *Hormiphora californensis*, and the siphonophore *Chuniphyes multidentata*. To our knowledge, this is the first study to suggest an optimal number of reads for *de novo* assembly for the purposes of mRNA level analysis. These results are applicable to studies of organisms with limited genomic resources.

3.3 Results and Discussion

3.3.1 De novo assembly of transcriptomes

3.3.1.1 Assembly of mouse heart transcriptome

Raw mouse-transcriptome reads from the ENCODE project were downloaded from NCBI short-read archive. Sample SRR453174 (mouse heart RNA-seq) consisted of 82,886,668 x76bp reads as paired-ends. Filtration (see Methods) removed 11.7% of the reads, almost 95% of which were due to low quality scores. In order to examine the effect of number of reads on the assembly, we computationally sub-sampled randomized sets from the original library. It is suggested that sequencing of very small numbers of reads can be subject to biases and that cDNA normalization can improve the uniformity of the library at low numbers of reads. [35] Such an approach might be quite costly, and the computational sub-sampling approach has the advantage of drawing from the largest pool of reads and avoid biases which could occur at low numbers

of reads. Subsets of the filtered library were generated containing 1,5,10,20,30,40,50,60, and 70 million reads. Reads from each set were included in the next largest set, thus all of the reads in the 1 million set are included in the 5 million read set, and so forth. These sets were assembled with Velvet/Oases [85, 117] and Trinity [30] (For a detailed comparison of assemblers, see [119]).

Schulz *et al.* reported reliable parameters for Oases which produced high-quality assemblies of mouse and human cell cultures, using 64 million and 30 million reads, respectively [85]. This included use of a broad k-mer range with a low starting k-mer of 19 or 21 up to a k-mer of 33 or 35. Accordingly we used k-mers from 21 to 33. Also, a minimum k-mer coverage is required by Oases to retain any given node during the assembly process; by default this is 3 in Oases, that is, any node must have at least three-fold coverage for that node to be used. Some differences were observed in the output when this parameter was changed, and so the same data were assembled with coverage cutoff of 3 (referred to hereafter as C3) and a stricter cutoff of 10 (C10).

The number of transcripts (Oases terminology for contigs) increases steadily for all assemblies (Figure 3.1A). C10 also had substantially fewer transcripts and accordingly much higher mean and median lengths (Figure 3.1B-D). The pattern of increase for median and N50 (length for which half of the total bases are in contigs of this length or longer) tracked the mean for the C10 assembly, but not the C3 assembly which did not have a clear qualitative pattern. The mean, median and N50 were all lower for the

Trinity assembly than the C3 despite having far fewer contigs.

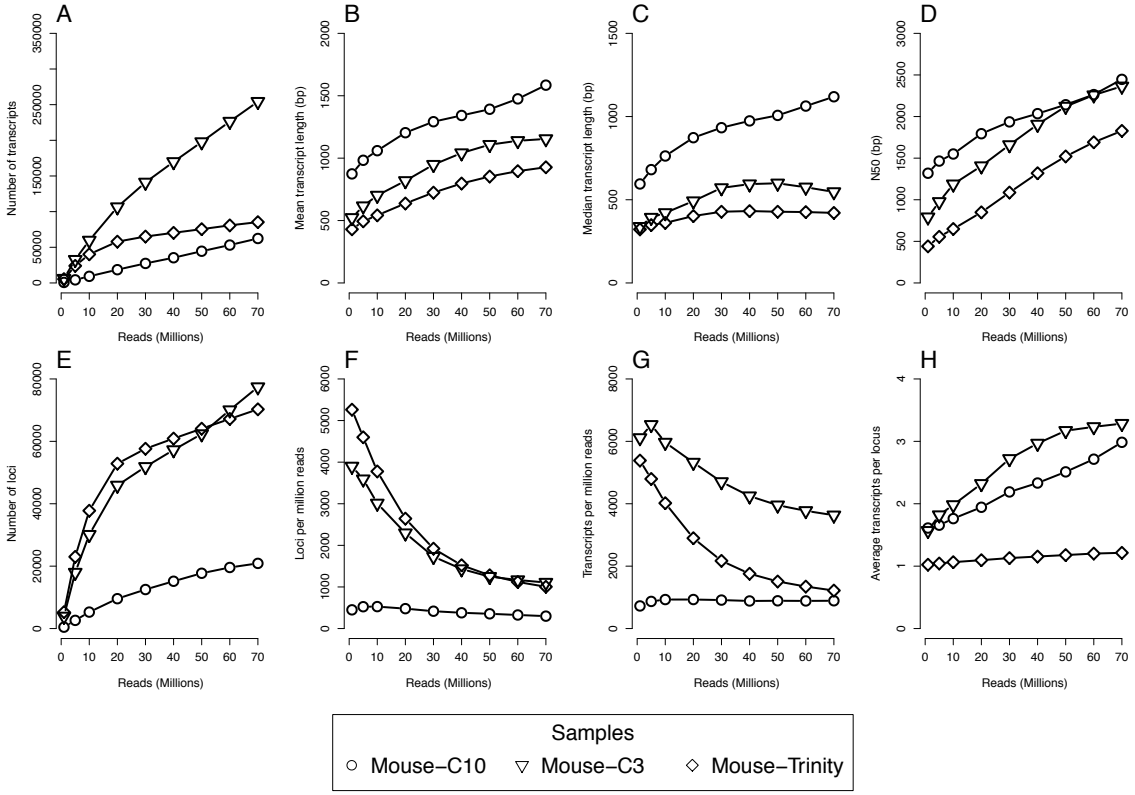

Figure 3.1: Assembly metrics for mouse heart transcriptome

Assorted size metrics for the mouse heart transcriptome showing (A) number of transcripts; (B) mean length; (C) median length; (D) N50 of the assembly; (E) number of loci; (F) loci per million reads; (G) transcripts per million reads; (H) transcripts per locus.

Oases generates transcript "loci", which is Oases terminology for the de-Brujin graph clusters meant to represent genes and their splice variants or highly-similar par-

alogs. Both curves approach a plateau for locus counts (Figure 3.1E-F). The greatest increase in loci was between using 10 million to 20 million reads for both C3 and C10. Similarly, the C3 assembly shows a decrease in the number of transcripts per read (Figure 3.1G), while the C10 assembly shows an almost constant number of transcripts per read. The number of transcripts increases while the number of loci tend to level off and this means the number of transcripts per locus always increases with more reads (Figure 3.1H). That is, on average, more variants will be generated with more reads even though some of these are likely due to noise. While the Trinity assembly more closely matches the trends for transcripts per read of the C3, the "components" (closest obvious parallel of loci) remain close to a unit ratio, suggesting that most components have only one associated sequence.

3.3.1.2 Assembly of invertebrate transcriptomes

Transcriptomes across a broad range of taxa were assembled as with the mouse and statistics of the largest assemblies are presented in Table 1. The stated GC content of the mouse genome is 42% while a subset of conserved genes showed a much higher value of 51.24%. [78, 111] Interestingly, for all assemblies except for mouse, the average GC content of the assembled contigs was lower than that of the raw reads (Figure 3.2), suggesting either that certain genes contribute much more to the overall GC content of the library or that biases can be introduced from the assembly.

For three of six samples (*D.gigas*, *H.imbricata* and *S.similis*), only select tis-

Table 3.1: Summary statistics of the largest transcriptome assembly for each organism

Table 1 - Assembly Statistics									
Organism	Mouse cov-cutoff-3	Mouse cov-cutoff-10	Mouse-Trinity	*Chuniphyes multiden-tata*	*Sergestes similis*	*Pleuro-mamma robusta*	*Dosidicus gigas*	*Hormiphora califor-nensis*	*Harmothoe imbricata*
Phylum	Chordata			Cnidaria	Arthropoda	Arthropoda	Mollusca	Ctenophora	Annelida
Tissue	Heart			Whole body	Legs	Whole body	Mantle	Whole body	Scale
Raw Reads (Millions)	82.88			103.41	93.59	64.11	60.66	64.67	75.60
Raw GC (%)	51.90			42.29	50.74	48.86	39.89	53.71	41.52
Filtered Reads (Millions)	73.18			102.36	92.42	63.86	56.26	57.58	70,34
Assembled Reads (Millions)	70			80	80	63.86	56.26	57.58	70.34
Transcripts	254,215	62,353	85,294	338,254	107,082	196,104	86,897	175,701	191,290
Total Length (Mbp)	293.55	98.84	79.12	314.99	159.59	240.05	143.09	272.23	216.66
Mean (bp)	1,154	1,585	927	931	1,490	1,224	1,646	1,549	1,132
Median (bp)	547	1,119	421	421	837	855	1,026	1,153	689
N50 (bp)	2,364	2,447	1,828	1,854	2,803	1,993	2,876	2,373	1,949
Oases Loci	77,411	20,889	70272	49,831	18,139	22,385	14,227	17,960	21,914
GC (%)	54.08	53.95	53.46	31.24	44.66	45.78	36.55	51.66	40.53

sues were used for RNA extraction while the rest were whole body (*C.multidentata*, *H.californensis* and *P.robusta*. It should be noted that the *C.multidentata* sample combined sequences from the two major tissues, siphosome and nectophore and that the *P.robusta* sample was a combination of multiple individuals. This decision was based on size of the animals since very small organisms are difficult to dissect. Assembly trends analogous to Figure 3.1 for the six animals are shown in Figure 3.3. Mouse C10 data from Figure 3.1 are shown in gray as reference. Three main trends emerged. Whole-body samples were characterized by a rapid gain of transcripts and increases in transcript size through 40 million reads, while all other parameters level off after 40 million reads. Single tissue samples showed a slow gain of relatively long transcripts across fewer loci. Lastly, the whole-body siphonophore showed continuous gain of both

short transcripts and loci without reaching an asymptote at the maximum number of reads assembled.

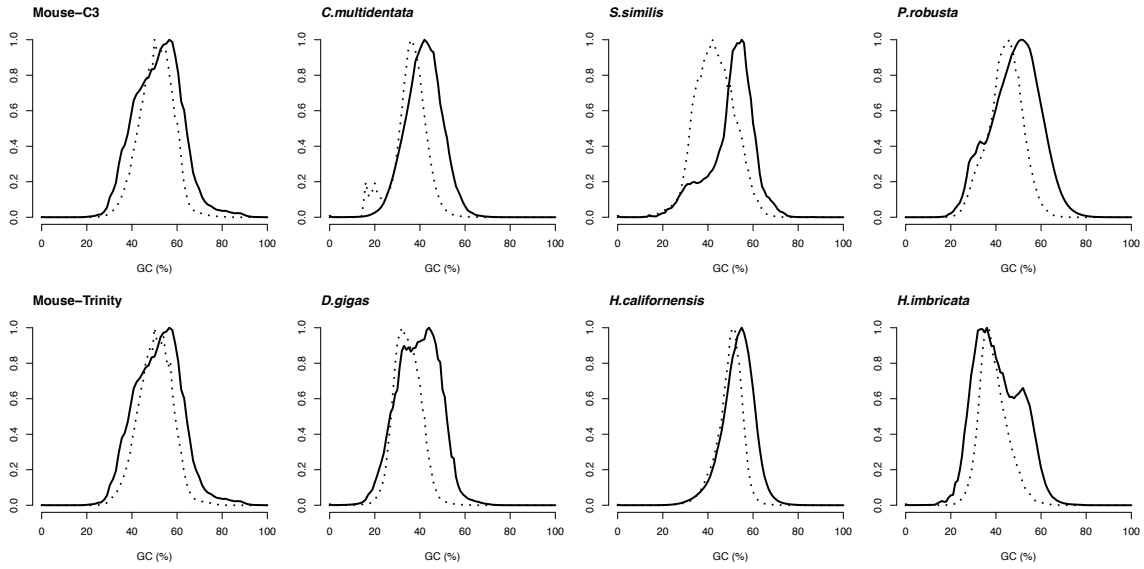

Figure 3.2: Histograms of GC distributions

Dashed lines show the normalized abundance of transcripts by GC content, while solid lines show normalized abundance of the raw reads.

Four of the animals showed modest gains in mean, median and N50 with more reads (average 20% from fewest to most reads), while *P.robusta* and *H.californensis* nearly doubled from the fewest to the most reads (Figure 3.3B-D). Most of the transcript-length increase occurred before 30 million reads, suggesting that adding more reads did not produce longer sequences beyond that threshold, or that they became longer at the same rate that new, short transcripts were generated. As with the mouse samples,

transcripts were added continually with more reads (Figure 3.3A). Compared to the mouse, on average these six animals all had more transcripts per locus (Figure 3.3H). It is unclear why this would be the case, though the C10 assembly had the fewest number of transcripts overall for all numbers of reads. The most pronounced gains in loci happened within the first 10 million reads, particularly for *P.robusta* and *H.californensis* (Figure 3.3E-F). Gains in loci tended to level out between 40 and 60 million reads, suggesting most genes (or parts of genes) were assembled by 60 million reads.

A very high number of transcripts for *C.multidentata* (Figure 3.3, circles) led to the lowest mean, median, and N50. The number of removed, low-quality reads is comparable in this sample to others, so low quality is unlikely to be the cause. As two sets of reads were combined into a whole animal, this may have created artifacts. However, another *C.multidentata* siphosome sample produced assemblies with large numbers of relatively short sequences (data unpublished). One possible explanation is that siphonophores have continuously developing differentiated zooids. [19] These zooids have specialized functions which are in some ways analogous to organs, and a whole organism can contain multiple developmental stages and express a large part of the genome, possibly confounding the assembly process. Assemblies of a number other siphonophores (data unpublished) similarly had many short transcripts. We speculate that alternate assembly strategies or very careful dissections might be required for animals in this lineage.

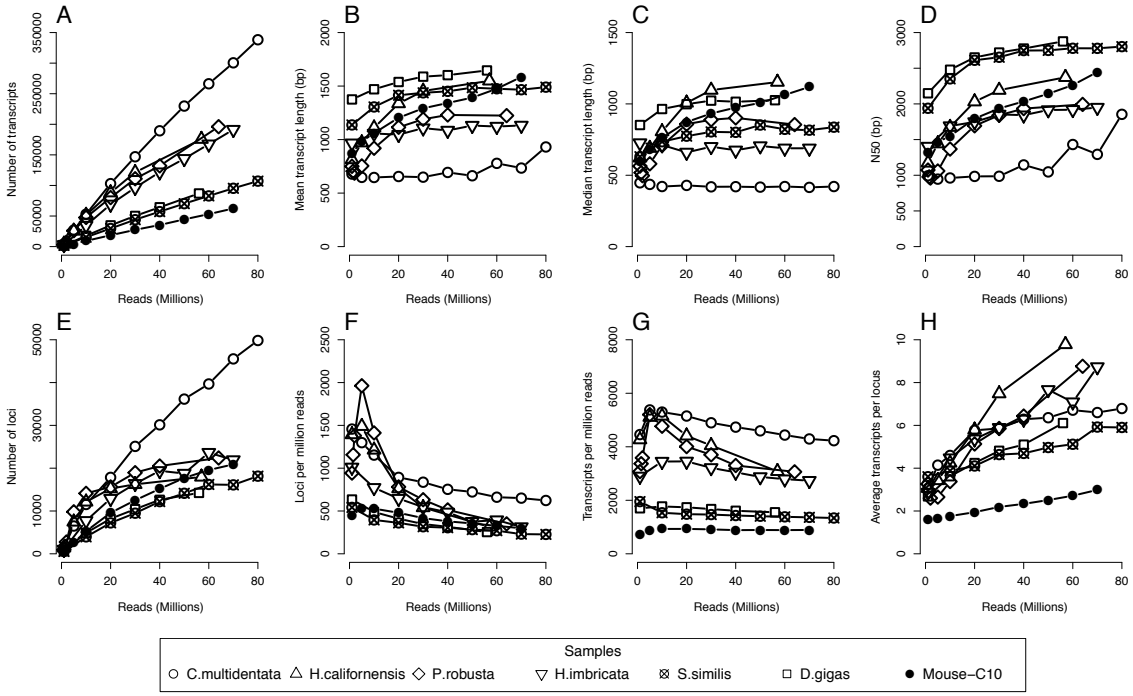

Figure 3.3: Assembly metrics for marine organisms

Assorted size metrics as in Figure 1; (A) number of transcripts; (B) mean length; (C) median length; (D) N50 of the assembly; (E) number of loci; (F) loci per million reads; (G) transcripts per million reads; (H) transcripts per locus.

3.3.2 Discovery of conserved genes

3.3.2.1 Conserved mouse genes

One approach used to assess genome completeness is to search only for conserved eukaryotic orthologous genes (KOGs). The current NCBI KOG database has 860 gene clusters across 7 eukaryotes with over 16000 proteins [99]. The KOG reference

genes did not include mouse sequences, and this provided an opportunity to test predictions about *de novo* transcriptome quality while still having a reference in the end to confirm the reliability of the sequences. For each KOG, the transcripts were aligned against the reference KOGs with tblastn, and the best coding sequence was kept. The putative proteins were classified by length relative to the range of sizes of the reference KOGs. The size range allowed some flexibility, as 12 mouse proteins were larger than the longest reference protein for that KOG, and 5 were shorter than the shortest reference protein. Finally the proteins were aligned with blastp against reviewed mouse proteins in Uniprot to determine accuracy. One protein was unreviewed (Q3UWL8, Mouse Prefoldin 4). For this test, Trinity and Oases are comparable at assembling full-length proteins, though Trinity appears to be slightly better at reconstructing canonical proteins (Figure 3.4A).

However, gene duplications present difficulties for such assessments unless one had *a priori* knowledge of how many copies should be present in the genome. For this study, we also used the subset of eukaryotic KOGs containing 248 genes from the CEGMA pipeline which were identified as single-copy orthologs in most genomes [71,72]. Almost one third of these KOGs are involved in processes like transcription and translation and were expected to be expressed in many tissues. Trinity and Oases with a lower coverage cutoff of 3 found similar numbers of KOGs at much lower numbers of reads (Figure 3.4B) than compared to the C10 assembly. Also more KOGs were found within expected length much faster with C3 than with the higher cutoff of 10, and the

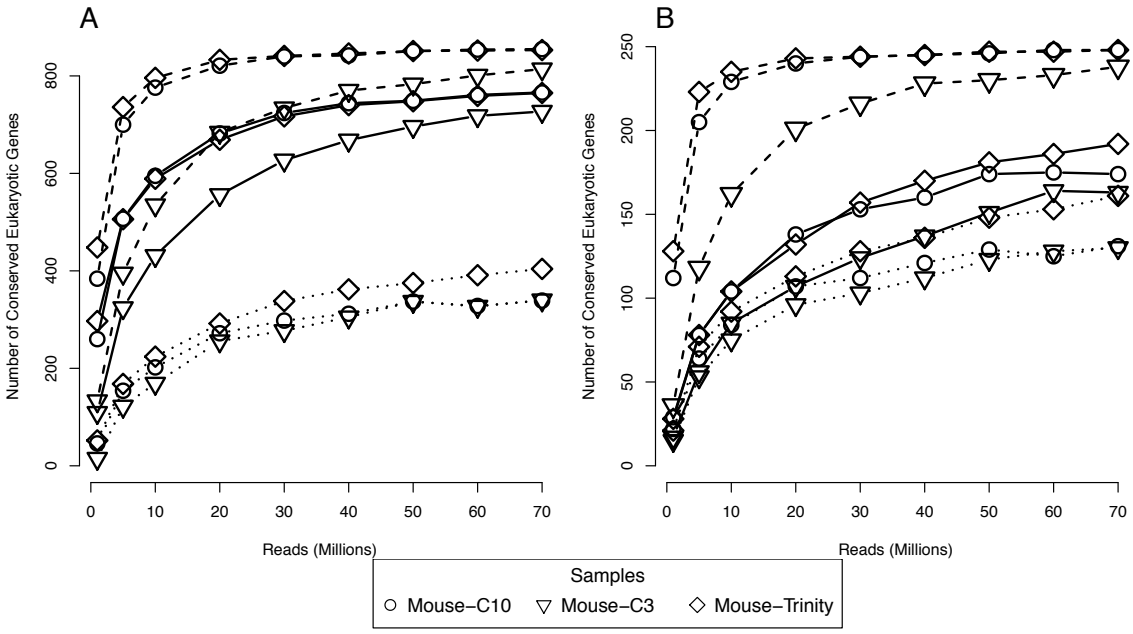

Figure 3.4: Conserved genes in the mouse transcriptome

Saturation curves of discovery of genes in the mouse heart from a set of (A) 860 conserved orthologs from NCBI and (B) a subset of 248 conserved orthologs; genes which have any blast hit are tracked in dashed lines; genes which the translated protein was within the expected size range of the conserved gene are solid lines; proteins which are 100% identical to a canonical protein in Uniprot/Swissprot mouse database are shown in dotted lines.

Trinity assembly outperformed both of these. These results suggest that it is better to have a lower cutoff and assemble more sequences. Likewise, the Trinity assembly had more transcripts than C10 and were shorter than those in C3, yet more KOGs were found with fewer reads and more coding transcripts were correctly assembled at greater numbers of reads. However, for the Oases assemblies this had remarkably little

effect on the number of correct canonical proteins that were found (Figure 3.4, dotted lines). Although there is some overestimation, no protein designated as too short was ever correct. Regarding the fate of the other full-length proteins, for C3 at 70 million reads, 186 KOGs were found within the expected range, though only 131 were correct. Eight of the 186 KOGs had only 1 mismatch in the amino-acid sequence compared to the reference protein which could be due to errors, splice variants, tissue-specific modifications or alleles. The remaining KOGs had at least two amino-acid changes but were within the size range. Thus for the mouse, the size range was a reliable predictor of true full-length proteins.

3.3.2.2 Conserved invertebrate genes

We then examined our invertebrate transcriptomes for completion using the same set of KOGs. There was a clear, qualitative difference between whole-body organisms (Figure 3.5A) and dissected tissues (Figure 3.5B). C10 mouse data are included for reference. For whole-body transcriptomes, over 90% of the KOGs were detectable at 20 million reads, yet the number of within-length KOGs went down with higher numbers of reads past 20 million. This could be caused if proteins declared to be within-range were longer than the true protein due to mis-assembly causing addition of pieces, or if the true protein became mis-assembled with addition of noisy reads. In nearly all of our assemblies, it was the latter: mis-assembly of the putative protein which generated stop codons. *C.multidentata* (Figure 3.4A, circles) was again exceptional, as the number of

within-length KOGs increased more slowly with addition of more reads than the other two whole-body animals (*H.californensis* and *P.robusta*) and only decreased after 50 million reads rather than 20 million.

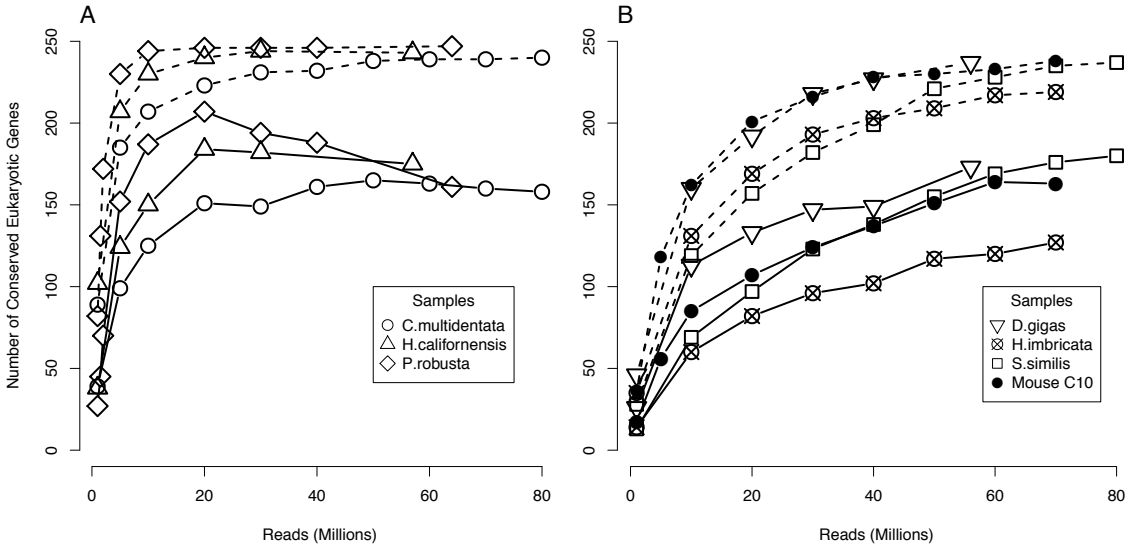

Figure 3.5: Conserved genes in marine organisms

As in Figure 3.4, genes with a reliable blast hit are shown in circles for all 6 marine organisms; genes which the translated protein was within the expected size range of the conserved gene are in solid lines.

For dissected-tissue transcriptomes (*Dosidicus gigas*, *Harmothoe imbricata*, and *Sergestes similis*), the rate of discovery of KOGs was much lower, with between 63% and 81% of KOGs detectable at 20 million reads (Figure 3.4B). This was not surprising since those genes may not be highly-expressed in all tissues and it is likely tissue-specific

genes account for the bulk of the assembly at low numbers of reads. Isolated tissues may express fewer universal KOGs that we selected in our test, and we expected that other abundant transcripts should mis-assemble at high numbers of reads in that tissue. However, the dissected-tissue transcriptomes had longer transcripts and fewer loci, suggesting this was not the case. Since whole-animal transcriptomes include all tissues, a greater proportion of the genome is expressed so coverage of any given transcript or splice-variant is proportionally much lower. The length saturation patterns appear to be different between whole-animal and tissue transcriptomes. However, using conserved genes as a metric, there appears to be limited benefit of sequencing beyond 60 million reads.

3.3.3 Mis-assembly at high numbers of reads

KOGs with single-exon coding sequences in the mouse were examined for mis-assembly. To increase the number of genes examined, another set of KOGs from only metazoans (*C.elegans*, *D.melanogaster* and *H.sapiens*, CDH) was used. The KOG database at NCBI contained 1147 clusters common to CDH. Again, only genes that were annotated as single copy in all three animals were used, leaving a final set of 202 KOGs specific to metazoans. These combined sets of 450 had 12 genes in mouse which were presumed single-copy and annotated in NCBI to have a single-exon coding sequence (GenBank:NP_062724.1, NP_666327.2, NP_082281.2, NP_058612.3, XP_899832.1, NP_001153802.1, NP_001104758.1, NP_077152.1, XP_486217.2, NP_598737.1, NP_032025.2, NP_075969.1).

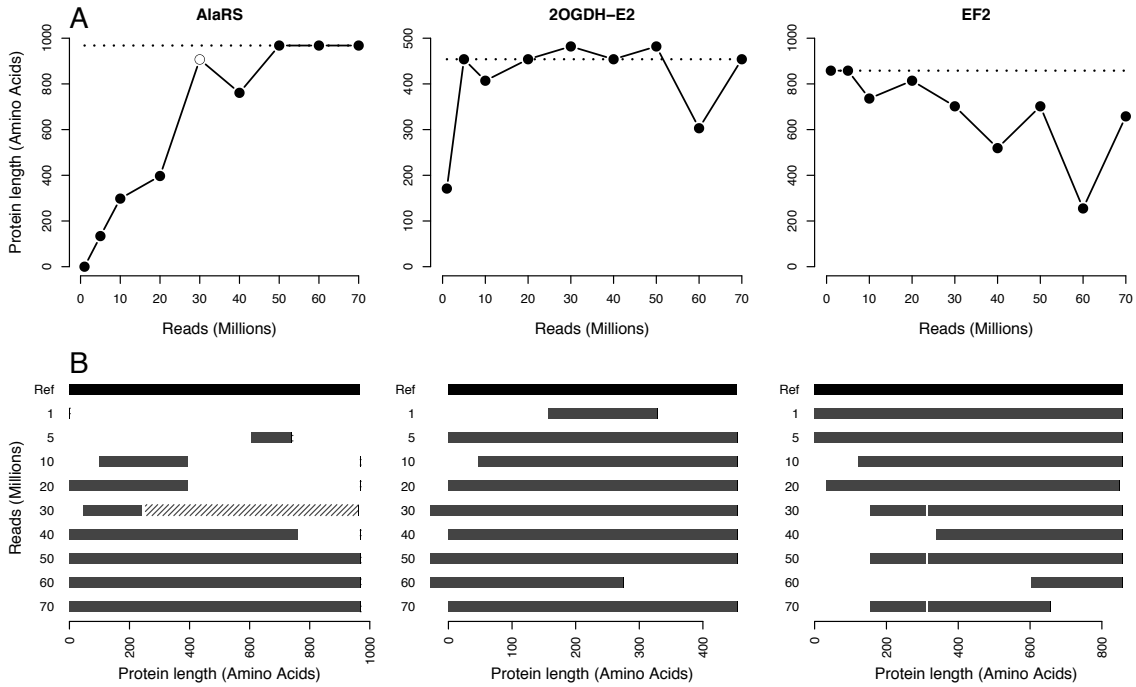

Figure 3.6: Selected cases of misassembly

Orthologs were tracked across multiple sequencing depths, and selected examples are here showing some of the pitfalls of assembly. (A) The lengths of three proteins are shown (AlaRS, Alanyl-tRNA synthetase; 2-OGDH-E2, 2-oxoglutarate dehydrogenase subunit E2; EF2, Elongation factor 2), and the canonical protein length is indicated by a dotted line. (B) Protein alignment view of the same three proteins compared to the Uniprot/Swissprot canonical protein, which is shown as the black bar. A chimeric portion of AlaRS at 30 million reads is indicated by the hashed bar, where it contains a sequence from the putative mitochondrial alanyl-tRNA synthetase 2 protein (NP_941010), and corresponds to the white point at 30 in (A). For AlaRS and EF2, some alignments produced a few short gaps compared to the reference proteins.

At 70 million reads, 3 genes in C3 had alternate erroneous coding sequences: NAT6, CHMP1B1/DID2, FTSJ (N-acetyl transferase 6, Charged multivesicular body protein 1b-1, Ribosomal RNA methyltransferase, respectively). The sequence of CHMP1B1 was never assembled correctly for any number of reads and the best version was missing 9 amino acids at the N-terminus including the start codon. Only NAT6 had extraneous coding sequence in C10, suggesting that such errors can be controlled by limiting read count as well as increasing k-mer coverage thresholds.

While some mis-assemblies can occur with more reads, overall this is not a problem, as shown by the curves in Figures 3.4 and 3.5. However, select cases of mis-assembly of the mouse genes are shown in Figure 3.6. AlaRS (Alanyl-tRNA synthetase) presents an example of the optimal scenario, whereby the protein is not found at all with few reads, but then pieces come together with the addition of more reads until the final protein is correctly assembled. The majority of proteins follow this trend. 2-OGDH shows an unusual oscillation between the reference protein and alternate forms. EF2 is assembled correctly with few reads, then errors accumulate as more reads are added. From this, it cannot be assumed that the largest set of reads will produce the best contigs. Schulz *et al.* indicated that between 10 and 20% of Oases transcripts had some degree of misassembly [85]. This value was found to correlate with the smallest k-mer used in assembly and the authors suggest using larger k-mers if problems arise due to chimeric transcripts. Thus if using more reads, it may be advisable to use larger k-mers or a higher static coverage cutoff.

3.4 Conclusions

In this study, a number of transcriptomes of whole animals and tissues from non-model organisms and one mouse organ were assembled and the completeness was assessed using a set of conserved genes. Additionally, a comparison was made between two high-performing assemblers with respect to the mouse data. Oases required much greater memory usage while Trinity had much longer run times (approximately 2-fold longer). Both Trinity and Oases perform comparably at assembling conserved genes across a large set, indicating that the saturation depth is not greatly affected by assembler choice.

Overall, these results suggest that for whole-body transcriptomes and individual organs or cells, 30 and 20 million reads are sufficient for mRNA level coverage, respectively. For the read length used in this study, that would produce 2-3 gigabases of sequence. It should be noted that the mouse data consisted of shorter reads than used for the invertebrates, but this did not appear to have substantial effect as this difference was only between 75bp reads and 100bp reads. Assembly errors are evident in whole-body transcriptomes after 30 million reads, and the average length appeared to level off at the same depth. Presumably this depth would apply for studies of differential expression as well, as the highly expressed transcripts should be present and

distinguishable at that sequencing depth. Based on these data, we found it was optimal to acquire between 50 and 60 million reads, and then sub-sample up around 20 or 30 million. This approach reliably assembled nearly all proteins of interest. There were still observable differences between assemblies, although some of these differences may ultimately be due to variations in RNA quality or properties of the animal.

3.5 Methods

3.5.1 Samples and sequencing

D.gigas and *H.californensis* were collected in the Gulf of California by jig and trawl net, respectively. *C.multidentata* and *S.similis* were collected in the Monterey Bay using remotely-operated-underwater vehicles. *H.imbricata* samples were given courtesy of T. Rivers. All samples were flash frozen in liquid nitrogen immediately following collection. Total RNA was extracted using RNeasy kit (Qiagen) as per instructions. *C.multidentata* RNA was extracted with Trizol and purified with the RNeasy kit. Preparation of RNA-seq libraries was done using Illumina TruSeq kit for paired end reads. Total RNA was sent for sequencing at University of Utah. Multiple individuals of *P.robusta* were sampled off the coast of Namibia and sequenced at the Institute for Clinical Molecular Biology, (IKMB, Kiel University). Sequencing was done using the Illumina HiSeq2000 platform on a paired-end protocol with 100 cycles. Mouse heart data were downloaded from NCBI accession GSE36025, sample SRR453174.

3.5.2 Transcriptome assembly

All computations were done on a computer with two quad-core processors and 96GB RAM. For each sample, the orders of all raw reads were randomized with the randomize.cpp program and processed with a modified version of the filter_illumina.cpp program in the Agalma transcriptome package (https://github.com/caseywdunn/agalma). This removed low-quality reads (with mean Phred score < 28), as well as reads containing adapters and reads that were mostly repeated bases, such as polyT tracts. Reads from pairs with one good read and one bad read retained the good read for the largest assembly. Otherwise, only good pairs were used in other assemblies. The transcriptome for each set was assembled *de novo* using Velvet v1.2.06 /Oases v0.2.06. Identical assembly parameters were used unless otherwise noted. Multiple k-mer assemblies were generated (21,25,29,33) and merged with Oases-M (k-mer of 27). A static coverage cutoff of 10 was used and insert size of the paired ends was estimated with the "-exp_cov auto" parameter, typically around 180bp, as expected. The minimum contig length was set to 100, which is the read length. The Trinity assembler was also used for comparison of mouse assemblies using the same filtered subsets of reads. Other than insert length being specified as the upper limit rather than the mean, default assembly parameters were used including a minimum transcript length of 200bp. Transcript lengths and GC content were measured with an in-house python script, sizecutter.py, available at the MBARI public repository (bitbucket.org/beroe/mbari-public/src).

3.5.3 Conserved gene analyses

All blast searches were done using the NCBI blast 2.2.25+ package [10]. We generated a script to blast and analyze the matches, kogblaster.py (on the public repository, as above). Briefly, the reference KOGs (860 orthologous groups from NCBI, or 248 orthologous groups, from http://korflab.ucdavis.edu/Datasets/cegma/) were aligned to each assembly with tblastn with an e-value cutoff of 10^{-6}. For each alignment, the subject hit was translated and coding sequences were only kept if they contained both start and stop codons. From this subset, the best alignment was declared to be the correct sequence. Next, the length of the correct sequence was used to estimate whether that sequence was full-length relative to the conserved orthologs. For each KOG in the CEGMA dataset, there were 6 proteins from 6 species and there was some variability in protein length (average 11.8% from longest to shortest). The variability from the the reference set was used to establish boundaries for size classifications which were made to watch the progression of assembly of individual genes: (1) within the size range of the KOG; (2) within the range but where the alignment was less than 90% of the length of the protein; (3) longer than those in the size range; (4) shorter than the size range; (5) shorter than the size range and shorter than the alignment, often indicative of a stop codon bridged by the alignment. The full-length size range was defined by ratios of the shortest protein to the second shortest, and analogously for the longest protein and second longest. For example, if the shortest protein within a KOG was 80AAs, and

the second shortest was 100AAs, the lower bound would be $(80 * (80/100))$, and thus 64AAs. This was calculated for each KOG, and was to account for proteins which could potentially become the 'new' shortest or longest. Ultimately, only those within the size range (1) were declared as full-length sequences.

Chapter 4

Occurrence of Isopenicillin-N-Synthase homologs in bioluminescent ctenophores

4.1 Abstract

The biosynthesis of the luciferin coelenterazine has remained a mystery for decades. While not all organisms that use coelenterazine appear to make it themselves, it is thought that ctenophores are likely producers. Here we describe a group of candidate genes for coelenterazine biosynthesis from the genomes and transcriptomes of 24 ctenophore species. These genes encode a group of highly conserved proteins that have the features of non-heme iron oxidases which are absent in the non-luminous species. Pairwise identities reveal an unusually high degree of identity even between the most unrelated species. Additionally, two related groups of proteins were found across all ctenophores, including those which are non-luminous, arguing against the involvement

of these two groups in luminescence. Important residues for iron-binding are conserved across all proteins in the three groups, indicating this function is likely still present. Given the known functions of other members of this protein superfamily are involved in heterocycle formation, we consider these genes to be top candidates for laboratory characterization or gene knockouts in the investigation of coelenterazine biosynthesis.

4.2 Background

Coelenterazine is the most widely occurring luciferin in marine bioluminescence [87], its use being reported in at least nine phyla [34]. The chemical structure was determined in parallel by two groups, one working on the sea pansy *Renilla* and the other working on the hydrozoan *Aequorea* [44, 90]. The structure contains an imidazopyrazinone core with three side groups that correspond to amino acid side chains, similar to the *Cypridina* luciferin [49]. Despite structural similarity, the two luciferins do not appear to be interchangeable [37, 38].

Although coelenterazine was first extracted from *Aequorea*, it was later shown that *A. victoria* gets the molecule from its diet [31]. It is unclear who is the prime synthesizer of coelenterazine and thus difficult to identify a biosynthetic pathway. However, several animals have been proposed as candidates based on reports of biolumienscence at early developmental stages. For example, a few very old reports had discussed "phosphorescence" from early-stage embryos of the ctenophores *Mnemiopsis leidyi* and a *Beroe* species [1, 73]. Various other reports had noted bioluminescence in embryos or

early developmental stages [27, 37], suggesting the possibility that ctenophores indeed produce their own coelenterazine.

It had been proposed that the luciferin biosynthesis could involve three amino acids forming a tripeptide and then cyclizing [61]. Indeed, feeding experiments using stable isotopes have shown that coelenterazine was synthesized from phenylalanine and tyrosine [68], however the mechanism of this is unknown. Likewise, the structurally similar *Cypridina* luciferin is synthesized from arginine, isoleucine, and tryptophan [67]. These experiments only demonstrated the dependence on amino acids, which potentially could occur several ways: a suite of enzymes link free amino acids to create di- and tripeptide intermediates, then cyclize that into the final structure; the residues "FYY" may be part of a larger peptide that is expressed normally and then cleaved and cyclized; a non-ribosomal peptide synthetase links the residues and then cyclizes them in a fashion similar to penicillin.

Here we identified candidate genes from the transcriptomes of luminous ctenophores that were not present in the non-luminous species. We compare these proteins to those from genomes of related ctenophores and show that this group of proteins are highly conserved even among distantly related animals, which is expected for critical biological processes.

4.3 Results

4.3.1 Sequencing and assembly of transcriptomes

We sequenced the transcriptomes of 21 luminous ctenophores and one non-luminous ctenophore (Table 4.1). Data from the genomes of two ctenophores, the luminous *Mnemiopsis leidyi* and the non-luminous *Pleurobrachia bachei* were used for comparison. Transcriptomes were assembled for each organism using both Velvet/Oases [85, 117] and Trinity [30], the results were pooled and redundant sequences were removed (see Methods). In general, more sequences appeared complete in the Trinity assemblies.

4.3.2 Transcriptomes include a broad set of expressed genes

Because the presence or absence of genes is difficult to address in transcriptomes as they reflect only genes expressed at the time of extraction or freezing, we examined a large set of genes to support that the transcriptomes are complete. We have previously used a set of genes housekeeping genes to assess transcriptome completeness [25]. Compared to the numbers of full-length annoated genes found in the reference genomes, many of the transcriptomes appear to contain full-length homologs of over 80% of target genes (Figure 4.1). Thus, from the set of housekeeping genes, we extrapolated that the transcriptomes contained most essential genes and the presence or absence of genes may be due to factors of biology rather than sequence analysis.

Table 4.1: List of ctenophore specimens

Species	Luminous? Y/N	Origin	Caught with	Extraction method	Library prep
Bathocyroe fosteria	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Bathyctena chuni	Yes	Monterey Bay	ROV	QR	TS-dT
Beroe aybssicola	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Beroe forskali	Yes	Monterey Bay	ROV	QR	TS-S-dT
Bolinopsis infundibulum	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Charistephane fugiens	Yes	Monterey Bay	ROV	QR	TS-S-dT
Dryodora glandiformis	Yes	Monterey Bay	Blue-water	QAP	TS-S-dT
Euplokamis dunlapae	Yes	Monterey Bay	ROV	QR	TS-S-dT
Haeckelia rubra	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Hormiphora californensis	No	Gulf of California	Trawl	QR	TS-dT
Lampea lactea	Yes	Monterey Bay	Blue-water	Trizol	TS-dT
Lampocteis cruentiventer	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Ocyropsis maculata	Yes	Monterey Bay	Blue-water	QR	TS-S-dT
Thalassocalyce inconstans	Yes	Monterey Bay	ROV	QR	TS-S-dT
Undescribed ctenophore B	Yes	Monterey Bay	ROV	QR	TS-S-dT
Undescribed ctenophore C	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Undescribed ctenophore N1	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Undescribed ctenophore N2	Yes	Monterey Bay	ROV	QAP	TS-S-dT
Undescribed ctenophore T	Yes	Monterey Bay	ROV	QR	TS-dT
Undescribed ctenophore V	Yes	Monterey Bay	ROV	QR	TS-dT
Undescribed ctenophore W	Yes	Monterey Bay	ROV	QR	TS-S-dT
Velamen parallelum	Yes	Monterey Bay	Blue-water	QAP	TS-S-dT

Specimens and origins for ctenophores used in this study. See Methods for details on specimen collection. Abbreviations for extraction and library preps are: QAP, Qiagen AllPrep; QR, Qiagen RNeasy; TS-S-dT, TruSeq Stranded prep with oligo-dT selection; TS-dT, TruSeq with oligo-dT selection.

4.3.3 The FYY motif is found in the ctenophore genome

The ctenophore *Mnemiopsis leidyi* has been a model organism for bioluminescence for over a century. The genome was recently sequenced and is the first genome of a bioluminescent organism [79,83]. We considered that one possible mechanism for coelenterazine biosynthesis may be from encoded "FYY" residues that are enzymatically cleaved. From the predicted 16,543 filtered gene models in the genome, we identified 374 gene products that contain the motif "FYY". Two of these genes, ML199826a and ML35201a, had the FYY motif at the C-terminus of the protein. The two genes are

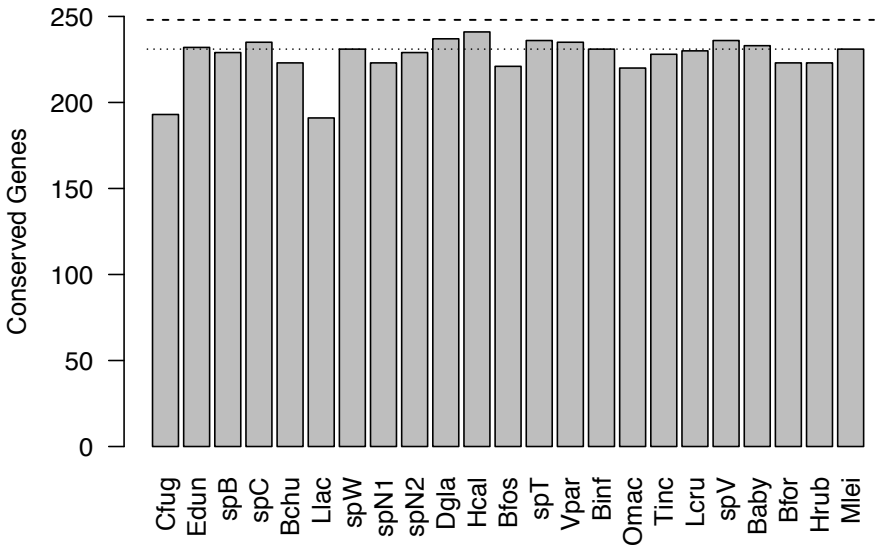

Figure 4.1: Survey of conserved genes across ctenophore transcriptomes

Dashed line indicates the maximum number of genes in this set, 248. The dotted line indicates the number of genes found in the *Mnemiopsis leidyi* genome. Most of the transcriptomes recovered a comparable number of genes as the genome. Species abbreviations are as follows: Bfos, *Bathocyroe fosteria*; Bchu, *Bathyctena chuni*; Baby, *Beroe aybssicola*; Bfor, *Beroe forskali*; Binf, *Bolinopsis infundibulum*; Cfug, *Charistephane fugiens*; Dgla, *Dryodora glandiformis*; Edun, *Euplokamis dunlapae*; Hrub, *Haeckelia rubra*; Hcal, *Hormiphora californensis*; Llac, *Lampea lactea*; Lcru, *Lampocteis cruentiventer*; Mlei, *Mnemiopsis leidyi*; Omac, *Ocyropsis maculata*; Tinc, *Thalassocalyce inconstans*; spB, *Undescribed ctenophore B*; spC, *Undescribed ctenophore C*; spN1, *Undescribed ctenophore N1*; spN2, *Undescribed ctenophore N2*; spT, *Undescribed ctenophore T*; spV, *Undescribed ctenophore V*; Vpar, *Velamen parallelum*

highly similar (Table 4.2). The shorter of the two proteins, ML35201a, was 99% identical to the other, varying only at a single residue but lacking a large piece of the N-terminus.

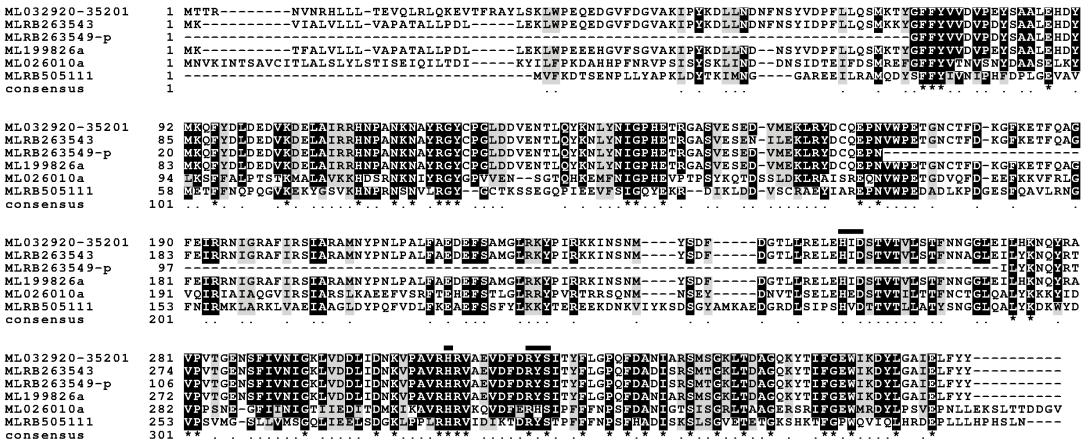

Figure 4.2: Multiple sequence alignment of *Mnemiopsis* proteins

ML032920-35201 is the putative full-length protein that connects ML032920a and ML35201a. MLRB263549-p indicates it is a partial sequence, as exons are missing in the scaffolds. The consensus sequence is indicated below, where identical residues are shown by '*' and similar residues are shown by '.'. Black boxes indicate the highly conserved residues putatively involved in iron and 2-oxoglutarate binding.

We then examined the unfiltered models and found two additional FYY-containing gene products in tandem on scaffold ML2635. The first one (MLRB263543) appeared to be complete and the second one (MLRB263549) was incomplete as several exons were clearly missing. Based on the alignment to the other proteins (Figure 4.2), some of the missing exons would fall in regions with low sequencing coverage, represented only by "N"s in the genomic scaffold. The two proteins appeared to be nearly identical to each other, varying at three residues.

Table 4.2: Percent Identity Matrix of *Mnemiopsis* genes and proteins

Gene	ML032920_35201	ML199826a	MLRB263543	MLRB263549	ML026010a	MLRB505111
ML032920_35201	=	97	93	94	54	51
ML199826a	100	=	91	94	52	50
MLRB263543	96	95	=	97	53	49
MLRB263549	97	97	98	=	56	50
ML026010a	48	46	45	47	=	49
MLRB505111	36	33	33	35	37	=

Pairwise identity for the *Mnemiopsis* genes. Protein sequence identity is shown on the lower portion and nucleotide sequences on the upper portion.

4.3.4 Four complete genes are annotated in *Mnemiopsis*

From the BLAST results we found two complete genes and two incomplete genes with the FYY ending. Because the predicted protein of ML35201a does not start with methionine and it is the first gene in its scaffold, we considered that the annotation may be incomplete and searched for other pieces of the gene. The unfiltered protein models (MLRB35201) and Cufflinks assembly (ML3520_cuf_1) show an additional exon at the N-terminus. Since these genes still would be missing almost 100 amino acids compared to ML199826a, we then searched for the N-terminal fragment in other scaffolds, and recovered two unfiltered protein models (MLRB032948 and MLRB032949) and the corresponding filtered model fragment (ML032920a) at the $3'$ end of scaffold ML0329. This suggests that scaffolds ML0329 and ML3520 are in proximity and are bridged by this gene. Using PCR, we were able to amplify a fragment of approximately 2kb using unique primers on each scaffold, confirming that these scaffolds are indeed adjacent (Figure 4.3).

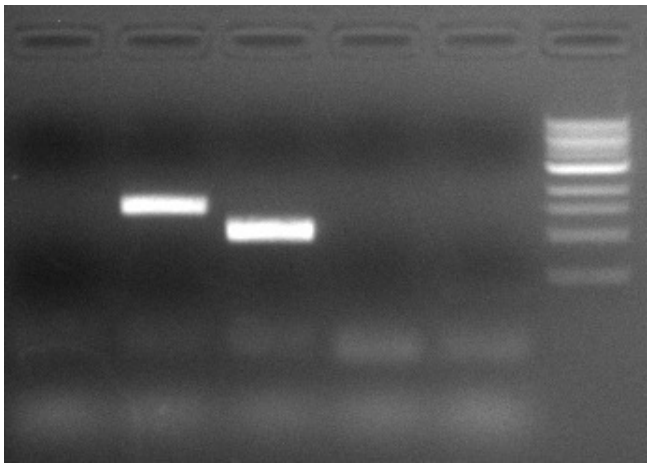

Figure 4.3: Agarose gel of PCR amplified genomic fragments from *Mnemiopsis leidyi* Amplification of gene ML35201a (right band) and the scaffold bridging ML032920-35201 (left band) with a 1kb ladder on the right.

Examining possible cellular locations, SignalP [74] indicated that ML199826a is likely to be cleaved at the "ATA-LL" site of the N-terminus and possibly secreted (D score: 0.899), likewise for MLRB263543 (D score: 0.919). While the rest of the gene is nearly identical, the putative full gene (ML032920a-ML35201a) differs from ML199826a at the N-terminus. An identical piece to the N-terminus of ML199826a (residues "MKVIAL") was found in ML0329, however if canonical splice sites are used, this would result in either a low similarity exon at the N-terminus or a stop codon, suggesting that the genomic sequence is wrong, the gene is inactive due to a nonsense mutation, or that the N-terminal exons are unused for this gene. Given the very high identity scores for both the protein and gene, it is possible that the RNA support (Trinity and Cufflinks tracks) for the gene were actually due to mis-alignments of reads from

ML199826a.

Another gene, ML026010a, was found to be similar to the FYY proteins (Figure 4.2 and Table 4.2) but lacked the FYY ending. Similarly, in the unfiltered models another homolog without the FYY was found (MLRB505111), which was different from both the FYY proteins and the other non-FYY protein (Table 4.2). This protein was not identified in the filtered models because it was split into two tandem pieces, ML50512a and ML50513a.

In all, there were four full-length annotated proteins and two incomplete proteins. As they were not entirely identical, re-sequencing may verify the presence and expression of the incomplete genes.

4.3.5 The FYY proteins are homologs of IPNS

To gain some insight as to the possible function of the FYY proteins, we compared the sequence to known proteins in various public databases. We BLASTed the FYY proteins against the nr (non-redundant) database on NCBI. Interestingly, nearly all of the top hits for all of the proteins were to a 2OG-Fe(II) oxygenase from the ciliate *Oxytricha trifallax* (Table 4.3). This was surprising since ciliates are unicellular eukaryotes and are not closely related to ctenophores. In a more restricted search using the Uniprot/Swissprot database, the top BLAST hits for many of the FYY proteins were to the same set of isopenicillin-N-synthase (IPNS) homologs, mostly from bacteria (Table 4.4). These proteins are members of a group of Fe-dependent oxygenases that

include IPNS and deacetoxycephalosporin C synthase (DAOCS), the enzymes responsible for the heterocycle-forming steps of penicillin biosynthesis and the ring expansion in cephalosporin biosynthesis, respectively [84].

Table 4.3: Top BLAST hits for FYY proteins in nr

Hit	Species	Accession	ML032920-ML35201	ML199826a	MLRB-263543	MLRB-263549	ML026010a	MLRB-505111
2OG-Fe(II) oxygenase	*Oxytricha trifallax*	EJY83212	2e-24	1e-24	2e-23	8e-5	6e-25	3e-16
2OG-Fe(II) oxygenase	*Oxytricha trifallax*	EJY68314	2e-17	2e-17	1e-17	-	2e-21	-
2OG-Fe(II) oxygenase	*Oxytricha trifallax*	EJY86133	1e-16	1e-16	3e-15	-	4e-27	8e-26
Isopenicillin N synthetase	*Crassostrea gigas*	EKC20116	5e-16	6e-16	1e-16	1e-5	3e-23	3e-23
Isopenicillin N synthetase	*Crassostrea gigas*	EKC29048	1e-15	1e-15	4e-15	-	2e-23	1e-21
Unnamed protein product	*Oikopleura dioica*	CBY23383	8e-15	1e-14	5e-16	-	2e-25	1e-19
Unnamed protein product	*Oikopleura dioica*	CBY34089	4e-14	3e-14	3e-15	-	3e-25	2e-19
2OG-Fe(II) oxygenase	*Oceanibaculum indicum* P24	ZP_11130131	2e-13	1e-13	-	7e-21	-	3e-14
Isopenicillin N synthase family	*Gordonia rubripertincta* NBRC 101908	ZP_11242214	1e-12	1e-12	-	-	-	-
2OG-Fe(II) oxygenase	*Mesorhizobium opportunistum* WSM2075	YP_004613268	2e-12	2e-12	-	-	1e-20	-
2OG-Fe(II) oxygenase family	*Campylobacter jejuni* 81116	YP_001482719	-	-	2e-12	-	-	-
2OG-Fe(II) oxygenase family	*Campylobacter jejuni* 414	ZP_06372273	-	-	2e-12	-	-	-
Putative iron/ascorbate-dependent oxidoreductase	*Campylobacter jejuni* ATCC 33560	ZP_14173854	-	-	5e-12	-	-	-
Putative isopenicillin N synthetase	*Talaromyces marneffei* ATCC 18224	XP_002152319	-	-	-	9e-4	-	-
Isopenicillin N synthase	*Mycobacterium phlei* RIVM601174	ZP_09977466	-	-	-	-	1e-20	-
2OG-Fe(II) oxygenase	*Mesorhizobium alhagi* CCNWXJ12-2	ZP_09292393	-	-	-	-	-	1e-14
Oxidoreductase	*Acidocella sp.* MX-AZ02	ZP_11251216	-	-	-	-	-	2e-14
Unnamed protein product	*Oikopleura dioica*	CBY11707	-	-	-	-	-	2e-13

Best ten BLASTP hits against the NCBI nr database for each of the proteins from *Mnemiopsis*. Numbers indicate e-values, for which a cutoff of 1e-3 was used. MLRB263549 was truncated and therefore did not align to many proteins.

Several conserved binding-pocket positions were detected when compared to the structures of IPNS and DAOCS [**?**,77]. In ML199826a, we identified the iron-binding positions, H245, D247 and H301, suggesting that this function is still present (Figure 4.2). We also identified the conserved RXS motif at R310-S312, involved in coordinating the 2-oxoglutarate in DAOCS or the carboxyl group of valine in the tripeptide (ACV) in IPNS. Y221 was also a conserved residue that coordinates the ACV-valine in IPNS, however the same tyrosine in DAOCS points the opposite direction towards a backbone helix.

4.3.6 FYY proteins are expressed only in luminous species

We found a homolog of the FYY protein in nearly every ctenophore in our transcriptome (Figure 4.4). In *Charistephane fugiens* we only found a partial sequence, though the assembly was among the worst of the set (Figure 4.1). Among the ctenophores examined here, both *Hormiphora californensis* and *Pleurobrachia bachei* were reported to be non-luminous [33]. Because these ctenophores belong to a family of other non-luminous species, we considered that this may be due to the genes being absent or unexpressed in that lineage. Several BLAST searches (blastn, blastp, and tblastn) failed to identify a similar sequence to the FYY proteins in *Hormiphora* transcriptome, however did find proteins similar to the non-FYY IPNS-homologs (Figure 4.4).

We considered that this absence could be due to a very low expression of the FYY protein which was removed during assembly. To address this, we then examined

Table 4.4: Top BLAST hits for FYY proteins in Swissprot

Hit	Species	Accession	ML032920-ML35201	ML199826a	MLRB-263543	MLRB-263549	ML026010a	MLRB-505111
Isopenicillin N synthase	*Streptomyces clavuligerus*	P10621	6e-12	8e-12	8e-12	-	2e-14	6e-12
Isopenicillin N synthase	*Lysobacter lactamgenus*	Q48739	2e-10	3e-10	1e-10	-	1e-17	4e-08
Isopenicillin N synthase	*Flavobacterium sp.* (strain SC 12,154)	P16020	1e-10	4e-10	1e-10	-	9e-18	2e-08
Isopenicillin N synthase	*Streptomyces griseus*	Q54243	4e-09	5e-09	1e-09	-	-	4e-07
Isopenicillin N synthase	*Streptomyces jumonjinensis*	P18286	5e-09	7e-09	4e-09	-	7e-15	1e-07
Isopenicillin N synthase	*Streptomyces microflavus*	P12438	1e-08	1e-08	2e-08	-	2e-11	-
Isopenicillin N synthase	*Streptomyces cattleya*	Q53932	1e-08	3e-08	2e-08	-	-	-
Isopenicillin N synthase	*Penicillium chrysogenum*	P08703	1e-05	1e-05	2e-06	-	1e-16	-
Isopenicillin N synthase	*Cephalosporium acremonium*	P05189	-	-	-	-	1e-17	-
Isopenicillin N synthase	*Emericella nidulans*	P05326	-	-	6e-05	-	7e-17	-
Isopenicillin N synthase	*Nocardia lactamdurans*	P27744	1e-05	2e-05	1e-05	-	8e-13	1e-11
1-aminocyclopropane-1-carboxylate oxidase	*Dictyostelium mucoroides*	A6BM06	-	-	-	-	1e-10	-
1-aminocyclopropane-1-carboxylate oxidase homolog 8	*Arabidopsis thaliana*	Q9M2C4	-	-	-	-	-	3e-07
Leucoanthocyanidin dioxygenase	*Petunia hybrida*	-	-	-	-	-	7.33e-07	
1-aminocyclopropane-1-carboxylate oxidase homolog 10	*Arabidopsis thaliana*	Q9LSW6	-	-	-	-	-	9e-06
1-aminocyclopropane-1-carboxylate oxidase homolog 1	*Arabidopsis thaliana*	Q84MB3	-	-	-	-	-	9e-06
Gibberellin 2-beta-dioxygenase	*Arabidopsis thaliana*	Q9XFR9	7e-05	6e-05	-	-	-	-

Best BLASTP hits against the Uniprot/Swissprot database for the FYY proteins from *Mnemiopsis*. Numbers indicate e-values, for which a cutoff of 1e-3 was used.

whether any fragments of the FYY proteins could be identified in the pre-assembled contigs (called "contigs.fa" by Velvet and "inchworm.K25.L25.DS.fa" by the first stage of Trinity.) We found 75 contigs this way and most were redundant when translated. Two putatively full-length proteins were identified from the contigs both of which group to non-FYY homologs in other ctenophores (Figures 4.5 and 4.6).

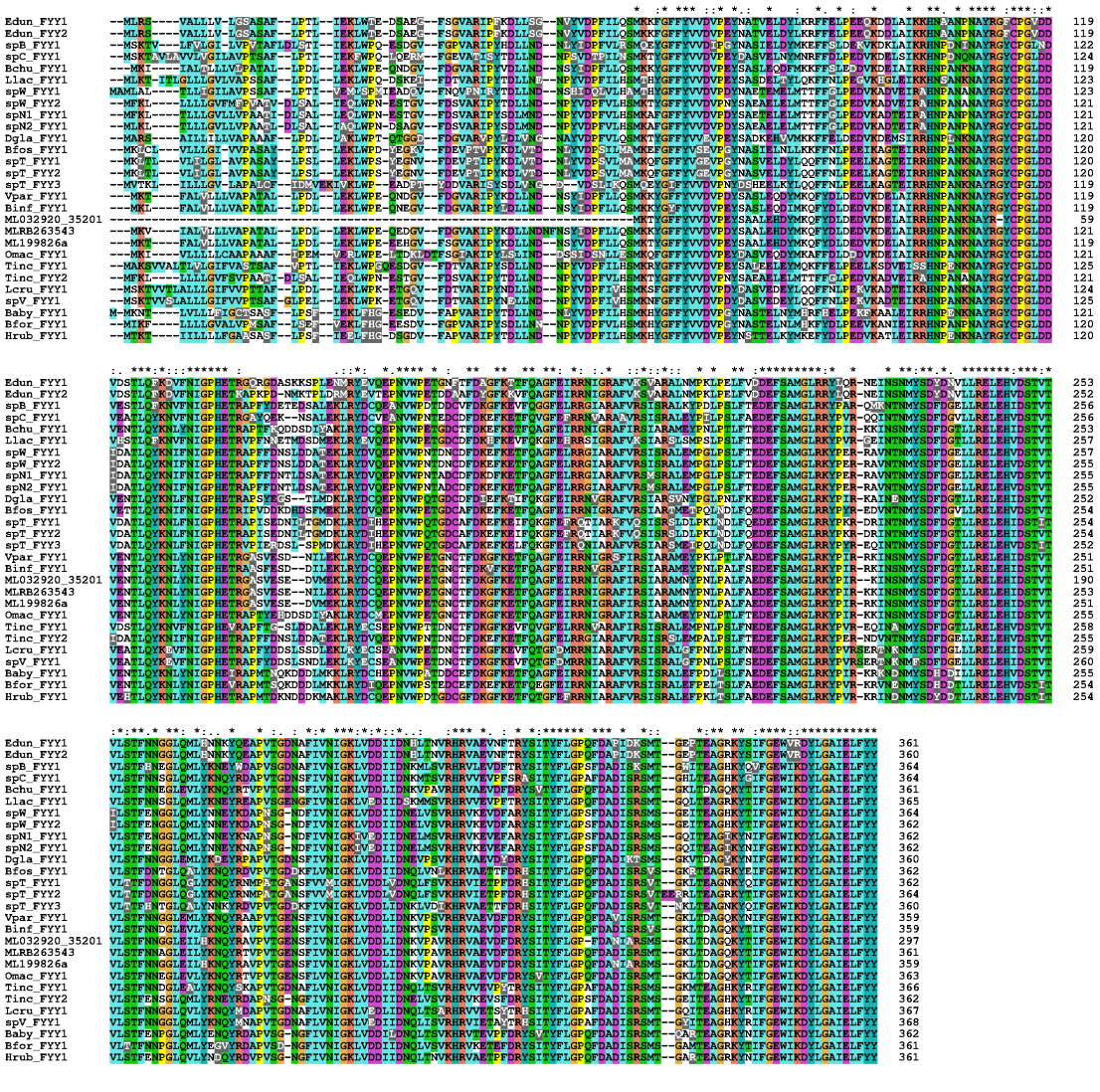

Figure 4.4: Multiple sequence alignment of all FYY proteins

Alignment of all FYY proteins across ctenophores. Partial sequences were excluded to show the high degree of identity, though were used for subsequent analysis. Species abbreviations are as in Figure 4.1

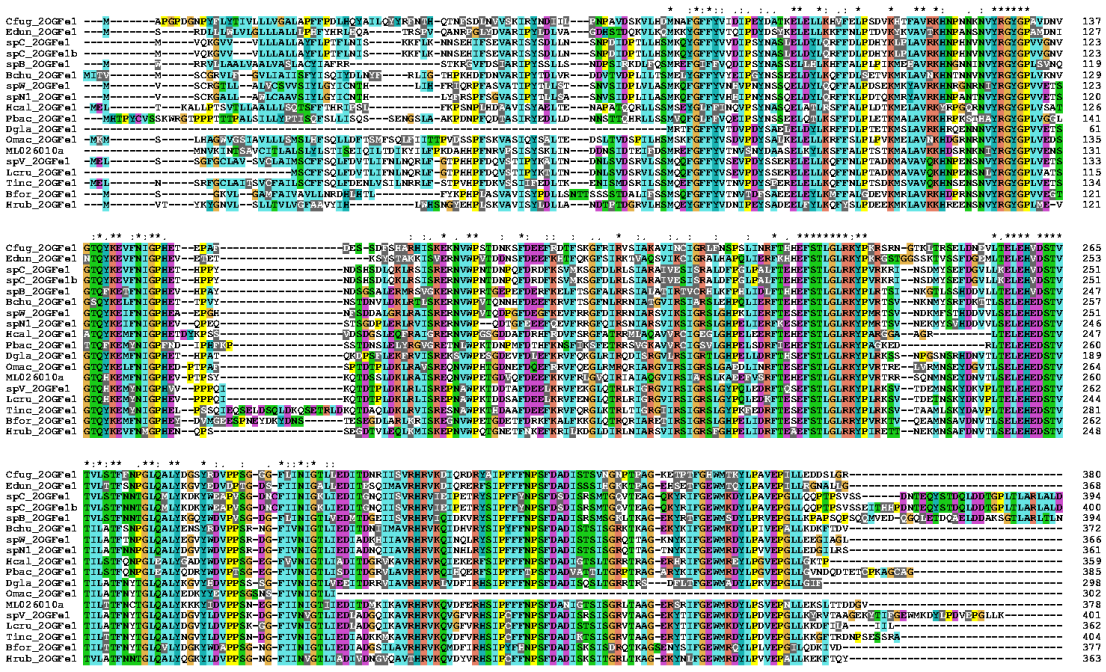

Figure 4.5: Multiple sequence alignment of all group 1 non-FYY proteins

Alignment of all group 1 non-FYY proteins across ctenophores. Partial sequences were excluded, though were used for subsequent analysis. Species abbreviations are as in Figure 4.1, with addition of Pbac as *Pleurobrachia bachei*.

We then further examined the predicted genes from *Pleurobrachia*. As with *Hormiphora*, two different genes which are most similar to the non-FYY IPNS-homologs (sp2669069 to ML026010a and sp3466438 to MLRB505111) were found in the unfiltered models (Figure 4.7). BLAST searches did not yield any sequence similar to the FYY proteins, nor were any of the conserved motifs found in any of the unfiltered models or translated adult mRNA datasets (RELEHXD, iron-binding site; GAIELFYY, conserved C-terminus). The absence of these proteins our searches in the genome of *Pleurobrachia*

and the transcriptome of *Hormiphora* indicated that these genes may have been lost in the *Pleurobrachiidae* clade. Without the genomic scaffolds to verify, we cannot resolve whether they were lost entirely or pseudogenized and unexpressed.

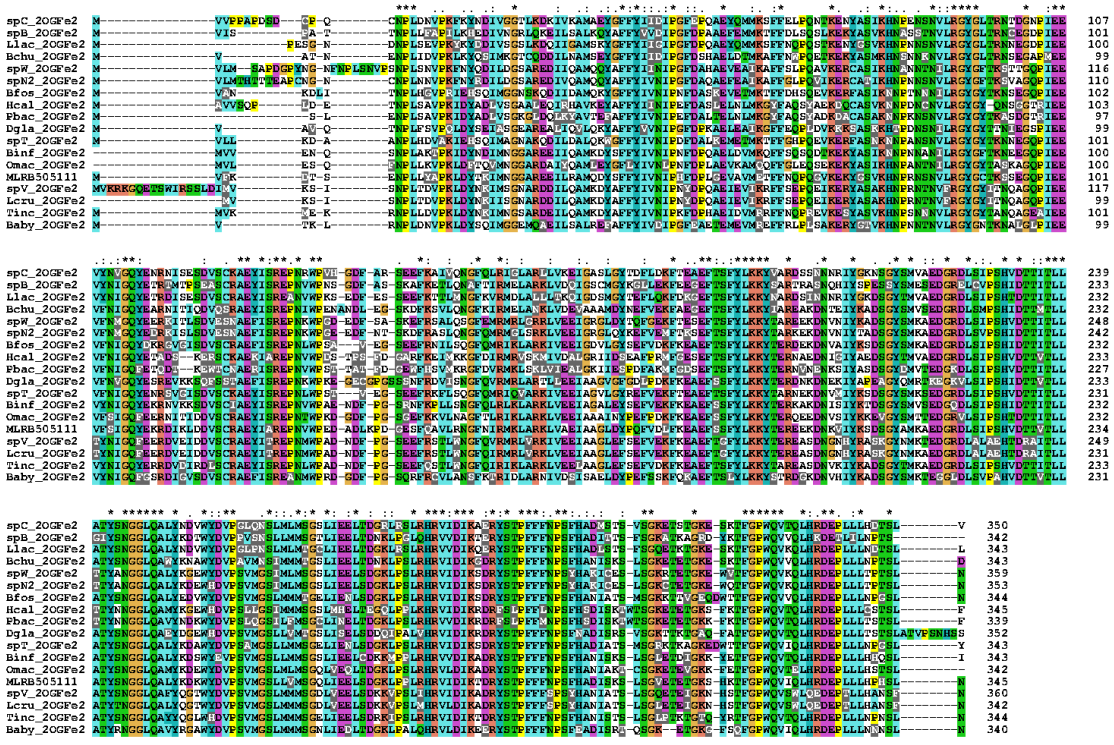

Figure 4.6: Multiple sequence alignment of all group 2 non-FYY proteins

Alignment of all group 2 non-FYY proteins across ctenophores. Partial sequences were excluded to show the identical portions, though were used for subsequent analysis. Species abbreviations are as in Figure 4.5.

4.3.7 Other luminescence genes are absent in non-luminous species

While the lack of luminescence may be due to the absence of the FYY proteins, other proteins involved in the process may be responsible instead. One report suggests that under several conditions, none of members of the family *Pleurobrachiidae* including *Hormiphora* produced any light [33]. It was shown that even when extracts for photoproteins were incubated with coelenterazine, no light was detectable, suggesting that photoproteins are absent in these species [33]. Indeed, thorough searching in the transcriptome assemblies of *Hormiphora* only identified one putative photoprotein (Supplemental Data) which was closer in sequence to the non-luminous protein from *Nematostella vectensis* [83]. A homolog was found in the *Mnemiopsis* genome, which is composed of four exons instead of one for all other photoproteins [83], suggesting it arose at a different time and may function in another way.

We then checked for photoproteins in *Pleurobrachia* and only found a partial gene of the homolog in *Hormiphora* (Figure 4.8) and no true photoproteins. Other hits to various photoprotein queries from other animals included two hits from Obelin (sb2644252, top hit back to hypothetical calmodulin-like protein; sb2643469, calmodulin), and one hit to a *Mnemiopsis* photoprotein (sb2667296, top hit back to NOX5, a calcium-dependent NADPH-oxidase), all due to the presence of EF-hand motifs.

We constructed a phylogenetic tree from these photoprotein-like genes in ctenophores and proper photoproteins from cnidarians and ctenophores, which show a clear difference between these photoprotein-like genes and true ctenophore photoproteins (Figure

4.8). True photoproteins are closer in sequence to cnidarian photoproteins than to these photoprotein-like genes, suggesting that duplication of the common ancestor of the two gene sets was before the divergence of metazoans. As the putative photoprotein-like genes in these three species lack the canonical EF-hand residues for calcium binding in photoproteins, it is questionable whether these proteins bind calcium at all. It is therefore likely that these putative genes are not photoproteins and perform some other function unrelated to bioluminescence. Ultimately, because we were unable to identify any photoproteins in the transcriptome of *Hormiphora* or the genome of *Pleurobrachia*, we conclude that those species are not bioluminescent in part because they lack photoproteins.

4.3.8 The FYY proteins are highly conserved

Because long segments of the FYY proteins appeared to be identical across many ctenophores, we then measured the degree of identity and base substitution across the proteins. FYY proteins had much higher pairwise percent identities (Table 4.5) than either of the groups of the non-FYY proteins (Tables 4.6 and 4.7). The lowest identity among the most distantly related members in the FYY group was 60% (average:71.61%) compared to 44% (average:56.00%) and 50% (average:62.17%) for non-FYY groups 1 and 2, respectively.

We then examined whether these genes were conserved across the ctenophore clade using codeml [115]. We found that FYY proteins were characterized by low ratios of non-synonymous to synonymous substitutions and generally much lower numbers of

Table 4.5: Percent Identity Matrix of all ctenophore FYY proteins

Edun_FYY1	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Edun_FYY2	95	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spB_FYY1	66	66	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spC_FYY1	67	66	75	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bchu_FYY1	65	65	71	70	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Llac_FYY1	71	71	74	74	71	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spW_FYY1	63	62	73	72	68	71	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spW_FYY2	65	64	75	72	69	73	89	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spN1_FYY1	65	65	74	70	69	74	85	91	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spN2_FYY1	66	65	74	70	69	74	85	91	98	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Dgla_FYY1	65	65	68	69	78	71	66	69	69	69	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bfos_FYY1	65	65	67	69	71	70	69	69	70	69	69	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spT_FYY1	60	62	65	68	67	68	65	66	67	67	69	79	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spT_FYY2	61	62	64	68	67	69	66	66	67	67	68	79	99	=	-	-	-	-	-	-	-	-	-	-	-	-	-
spT_FYY3	62	62	64	67	67	68	65	67	66	66	67	75	75	73	=	-	-	-	-	-	-	-	-	-	-	-	-
Vpar_FYY1	69	67	72	73	90	72	68	70	70	70	78	70	68	68	66	=	-	-	-	-	-	-	-	-	-	-	-
Binf_FYY1	66	64	70	72	91	71	68	70	69	69	79	70	67	67	66	94	=	-	-	-	-	-	-	-	-	-	-
ML032920_35201	71	69	73	77	89	73	72	73	71	71	82	73	70	70	71	93	93	=	-	-	-	-	-	-	-	-	-
MLRB263543	68	66	71	72	90	72	67	69	68	68	78	70	68	68	67	95	92	96	=	-	-	-	-	-	-	-	-
ML199826a	69	67	70	72	87	71	67	70	68	68	78	70	67	67	66	93	92	100	95	=	-	-	-	-	-	-	-
Omac_FYY1	64	63	68	70	88	67	66	66	65	64	74	68	66	66	66	84	84	89	85	83	=	-	-	-	-	-	-
Tinc_FYY1	68	68	74	74	71	75	73	76	74	73	70	69	66	66	68	72	73	75	72	71	69	=	-	-	-	-	-
Tinc_FYY2	65	64	75	72	70	75	86	96	90	90	69	71	67	68	68	70	71	73	70	70	67	76	=	-	-	-	-
Lcru_FYY1	66	66	74	73	70	74	70	72	73	72	68	71	70	70	68	71	70	71	70	70	68	80	73	=	-	-	-
spV_FYY1	66	66	74	73	68	73	71	72	73	72	67	70	70	70	66	71	69	71	70	69	68	79	72	95	=	-	-
Baby_FYY1	64	65	72	71	70	71	70	72	70	70	68	69	68	68	66	71	69	74	70	70	68	72	73	70	69	=	-
Bfor_FYY1	62	63	71	70	70	71	72	74	73	73	69	71	69	69	68	69	70	73	69	69	67	72	75	69	70	83	=
Hrub_FYY1	66	66	70	71	72	72	69	72	71	71	69	70	69	68	69	72	71	76	72	72	70	73	73	71	70	85	81

Pairwise percentage identity for the FYY proteins.

non-synonymous substitutions compared to the non-FYY proteins that were relatively more neutral (Table 4.8, Supplemental Data). Combined with the high identities across different ctenophore groups, this suggests that the FYY proteins are under strong purifying selection and any given mutation might result in the loss of activity for the protein, perhaps due to backbone changes which may affect a binding pocket or to interfaces with other proteins.

Table 4.6: Percent Identity Matrix of all Group-1 2OGFe proteins

Cfug_2OGFe1	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Edun_2OGFe1	60	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spC_2OGFe1	53	51	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spC_2OGFe1b	53	51	100	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spB_2OGFe1	54	52	58	57	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bchu_2OGFe1	51	54	57	57	56	=	-	-	-	-	-	-	-	-	-	-	-	-	-
spW_2OGFe1	52	53	56	56	56	64	=	-	-	-	-	-	-	-	-	-	-	-	-
spN1_2OGFe1	53	55	58	58	56	65	84	=	-	-	-	-	-	-	-	-	-	-	-
Hcal_2OGFe1	49	50	50	50	54	51	55	53	=	-	-	-	-	-	-	-	-	-	-
Pbac_2OGFe1	44	46	47	47	48	51	50	51	63	=	-	-	-	-	-	-	-	-	-
Dgla_2OGFe1	57	55	58	58	59	60	63	65	61	57	=	-	-	-	-	-	-	-	-
Omac_2OGFe1	49	47	51	51	48	57	55	57	51	47	67	=	-	-	-	-	-	-	-
ML026010a	48	48	52	52	52	57	58	58	53	47	65	59	=	-	-	-	-	-	-
spV_2OGFe1	48	51	49	48	48	59	60	61	51	50	66	61	58	=	-	-	-	-	-
Lcru_2OGFe1	51	52	52	51	53	61	61	62	53	53	66	63	61	95	=	-	-	-	-
Tinc_2OGFe1	47	49	49	48	50	58	60	60	52	49	67	58	61	79	80	=	-	-	-
Bfor_2OGFe1	48	50	54	54	52	56	60	61	52	49	65	58	59	59	60	58	=	-	-
Hrub_2OGFe1	50	52	54	54	51	60	58	59	52	49	64	56	56	57	58	58	64	=	-

Pairwise percentage identity for the ctenophore Group-1 2-oxoglutarate Iron (2OGFe1) proteins.

Table 4.7: Percent Identity Matrix of all Group-2 2OGFe proteins

spC_2OGFe2	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spB_2OGFe2	63	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Llac_2OGFe2	79	65	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Bchu_2OGFe2	67	64	68	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spW_2OGFe2	65	58	63	68	=	-	-	-	-	-	-	-	-	-	-	-	-	-	-
spN2_2OGFe2	65	58	63	67	86	=	-	-	-	-	-	-	-	-	-	-	-	-	-
Bfos_2OGFe2	64	59	63	65	68	68	=	-	-	-	-	-	-	-	-	-	-	-	-
Hcal_2OGFe2	56	51	56	57	58	56	55	=	-	-	-	-	-	-	-	-	-	-	-
Pbac_2OGFe2	57	50	58	55	57	57	56	75	=	-	-	-	-	-	-	-	-	-	-
Dgla_2OGFe2	58	54	55	57	61	61	61	55	54	=	-	-	-	-	-	-	-	-	-
spT_2OGFe2	64	59	63	66	68	68	84	55	57	59	=	-	-	-	-	-	-	-	-
Binf_2OGFe2	65	60	63	67	65	66	69	57	57	63	67	=	-	-	-	-	-	-	-
Omac_2OGFe2	64	57	60	64	64	63	66	55	57	60	65	71	=	-	-	-	-	-	-
MLRB505111	63	59	63	68	67	66	68	59	59	61	68	71	69	=	-	-	-	-	-
spV_2OGFe2	57	53	57	61	62	62	63	50	50	56	63	66	64	64	=	-	-	-	-
Lcru_2OGFe2	57	54	56	60	63	62	63	50	50	55	62	67	64	64	97	=	-	-	-
Tinc_2OGFe2	64	60	64	68	66	65	70	57	56	63	69	72	68	73	74	74	=	-	-
Baby_2OGFe2	61	58	60	61	61	61	62	53	54	56	63	59	61	63	58	58	64	=	-

Pairwise percentage identity for the ctenophore Group-2 2-oxoglutarate Iron (2OGFe2) proteins.

4.4 Discussion

Here we have sequenced and searched the transcriptomes of 22 ctenophore species for putative genes in the coelenterazine biosynthetic pathway. While it was pre-

Table 4.8: Base substitution ratios for *Mnemiopsis* genes

	ML199826a			MLRB263543			ML026010a			MLRB505111		
Species	dN/dS	dN	dS	dN/dS	dN	dS	dN/dS	dN	dS	dN/dS	dN	dS
Cfug	-	-	-	-	-	-	0	1.4037	0	-	-	-
Edun	0.0235	0.038	1.6201	0.0235	0.038	1.6201	0.4249	0.4006	0.9429	-	-	-
spC	0	0	0.4271	0	0	0.9644	0.2061	0.3011	1.4612	0.7789	0.4126	0.5297
spB	0.0393	0.0382	0.9738	0.0393	0.0382	0.9738	0	0.3139	0	0.3703	0.5359	1.447
Llac	0.1224	0.0788	0.6437	0.0825	0.0788	0.9552	0	0.365	0	2.0579	1.3566	0.6592
Bchu	0.0612	0.0385	0.6301	0.0268	0.0385	1.4356	0.3233	0.2546	0.7874	0	0.4472	0
spW	0.0408	0.0771	1.8886	0.0408	0.0771	1.8886	0	0.2075	0	0.446	0.4038	0.9053
spN1	0.1516	0.1394	0.9198	0.1516	0.1394	0.9198	0	0.2317	0	-	-	-
spN2	0.0933	0.1394	1.4947	0.0933	0.1394	1.4947	-	-	-	0.2764	0.4	1.447
Bfos	0.1406	0.1871	1.3311	0.0626	0.1871	2.9917	0.1943	0.1587	0.817	0	0.1798	0
Hcal	-	-	-	-	-	-	0	0.2961	0	0.7218	0.5637	0.781
Pbac	-	-	-	-	-	-	-	-	-	0	0.5105	0
Dgla	0	0	0.4271	0	0	0.9644	0	0.2876	0	0	0.1506	0
spT	0.1191	0.2111	1.7732	0.1191	0.2111	1.7732	-	-	-	0	0.0962	0
Binf	0	0	0.2563	0	0	0.2563	-	-	-	0	0.1695	0
Omac	0.0612	0.0385	0.6301	0.0268	0.0385	1.4356	0.3586	0.3413	0.9517	0	0.1401	0
Vpar	0	0	1.6201	0	0	1.6201	0.3273	0.3244	0.9911	-	-	-
ML199826a	-	-	-	0	0	0.2577	0	0.5831	0	0.4804	1.092	2.2734
MLRB263543	0	0	0.2577	-	-	-	0.5407	0.5831	1.0785	0.8796	1.092	1.2415
spV	0.0202	0.0792	3.9178	0	0.0792	0	0.0814	0.2063	2.5357	0	0.2292	0
Lcru	0.0202	0.0792	3.9178	0	0.0792	0	0	0.1839	0	0	0.2296	0
Tinc	0.0552	0.0792	1.4356	0.0202	0.0792	3.9178	0	0.1158	0	0	0.1373	0
Baby	0	0.0788	0	0	0.0788	0	0.1124	0.3499	3.1139	0.1821	0.182	0.9992
Bfor	0	0.1404	0	0	0.1404	0	0.0887	0.2071	2.3351	-	-	-
Hrub	0.0464	0.0779	1.6773	0.0761	0.0779	1.0228	0	0.3183	0	-	-	-

Base substitution rates of *Mnemiopsis* genes compared to the various non-heme iron oxidases of the other species. 0 indicates the model was inadequate for this analysis due to a lack of detected substitutions. Abbreviations are as in Figure 4.

viously demonstrated that coelenterazine can be synthesized from isotopically-labeled amino acids [68], several mechanisms could involve amino acids, including normal ribosomally-synthesized peptides. This led us to search for peptides including the motif "FYY", and also to search for non-heme iron oxidases, a class of enzymes known for many heterocycle-forming reactions such as those which create the heterocyclic structure of the tripeptide penicillin. We have identified one family of genes across luminous ctenophores which both contain the residues "FYY" which occur in coelenterazine as well as having

detectable similarity to non-heme iron oxidases. This includes several closely related genes in the genome of *Mnemiopsis leidyi* as well as two more distant non-heme oxidase families. These three protein families all appear to be closer to each other than to any other non-heme oxidases, which might be expected for an isolated clade such as the ctenophores.

The evident conservation of the FYY proteins between species suggests that whatever the function is, it is very important to the physiology of the animals. Bioluminescence is known to have functional importance in ctenophores [32], and photoprotein genes appeared to be under tight purifying selection [83]. It could then be expected that the production of luciferin would be tightly controlled as well, as disruptions to either luciferin biosynthesis or photoproteins would result in a loss of bioluminescence.

Of the initial hypotheses of possible biosynthetic pathways, we were quite surprised to find two in the same protein, that is, a FYY-containing protein that is also a non-heme iron oxidase. The apparent explanation is that, under some circumstance, these enzymes would be capable of auto-catalytic cleavage and cyclization of the C-terminal FYY residues to form coelenterazine. While there is no precedent for this type of reaction, it is evident from the types of chemistries displayed by other non-heme iron oxidases that the full range of activities of these enzymes is poorly characterized.

Verification of the functions could be realized two ways: cloning and knockout experiments. While cloning a gene is straightforward, expressing a functional protein is often challenging, given that the conditions for activity are unknown. For example, because several slightly different isoforms were found in a few of the transcriptomes and

the *Mnemiopsis* genome, it could be that multiple proteins are required for activity, perhaps as a hetero-dimer. These could, however, also just be redundant copies or very recent duplications in a species specific fashion. Knockouts and other genetic manipulations would be ideal to confirm the overall involvement in a process, though cannot easily dissect functions without something like LCMS to confirm any intermediates. It was recently demonstrated that *Mnemiopsis* specimens could be maintained in the lab for generations, suggesting the possibility of genetic manipulations that may ultimately resolve the functions.

New genetically-encoded optical tools are always desired for potential cell biology applications. Coelenterazine, for example, is the substrate of the calcium-activated photoprotein Aequorin, yet its complex heterocyclic structure makes it expensive to produce synthetically and limits the use in reporter technologies. Because the biosynthetic pathways for all eukaryotic luciferins are still unknown or incomplete, both attempts to genetically engineer a eukaryote to be self-luminous have used codon-optimized versions of the bacterial Lux genes, one in tobacco plants [51], the other in cultured human cells [16]. Discovery of the biosynthetic pathway of coelenterazine would enable a broad range of novel reporter systems and may ultimately provide insights into the evolution of bioluminescence in marine systems.

4.5 Materials and Methods

4.5.1 Specimens and sequencing

Specimens were collected either by trawl net, during blue-water dives, or captured at depth using remotely-operated-underwater vehicles (ROVs) (Tables 1 and 2). All samples were frozen in liquid nitrogen immediately following collection. All animals used in this study were treated ethically and responsibly. As no vertebrates or octopus were involved, no formal statement is required by the Helsinki Declaration. All specimens were sequenced at the University of Utah using the Illumina HiSeq2000 platform paired-end with 100 cycles.

4.5.2 Transcriptome assembly

All computations were done on a computer with two quad-core processors and 96GB RAM. For each sample, raw RNAseq reads were processed as previously published [25]. Briefly, read order was randomized. Low-quality reads, adapters, and repeats were removed. For efficiency, subsets of reads were used to assemble transcriptomes. Assembly was done with both Velvet/Oases [85, 117] and Trinity [30], though better sequences were often observed with Trinity. Transcripts from both assemblers were combined and redundant sequences were removed using the "sequniq" program in the GenomeTools package [**?**].

4.5.3 Genomic reference data

Gene models, scaffolds, and proteins for the *Mnemiopsis leidyi* genome [79] v2.2 were downloaded from NCBI at the *Mnemiopsis* Genome Portal (including browser at http://research.nhgri.nih.gov/mnemiopsis/). Gene models and transcripts for *Pleurobrachia bachei* genome v1.1 were downloaded from the the Moroz Lab (ftp linked from http://moroz.hpc.ufl.edu/). As the genomic scaffolds were unpublished, nucleotide analyses were excluded.

4.5.4 Gene identification

All BLAST searches were done using the NCBI BLAST 2.2.28+ package [10]. Various *Mnemiopsis* genes were examined manually using the genome browser and in-house Python scripts which can be downloaded at the MBARI public repository (https://bitbucket.org/beroe/mbari-public/src).

4.5.5 Alignments and phylogenetic tree generation

Alignments for proteins sequences were created using MAFFT v7.029b, with L-INS-i parameters for accurate alignments [48]. Trees were generate using RAxML-HPC-MPI v7.2.8 [93], using the PROTCATWAG model for proteins and 100 bootstrap replicates with the "rapid bootstrap" (-f a) algorithm.

4.5.6 Purifying selection analyses

Pairwise percentage identity calculations were generated among a suite of output files using ClustalX. The program implements a simple calculation and ignores gapped positions. To assess for evidence of purifying selection, ratios of non-synonymous to synonymous substitutions (dN/dS) were calculated using codeml in the PAML v4.7 package [115]. The previously generated tree was used to provide branch topology. Other parameters were as follows: seqtype=1 (codons); CodonFreq=2 (the F3X4 model); model=2.

4.5.7 PCR amplification

PCR was performed as follows: 98° C for 1 min; 30 cycles of 98° for 10s, 56° for 15s, 72° for 60s; final extension phase of 72° for 7min. Reactions were 50μ L using Phusion High-Fidelity PCR Master Mix with HF Buffer (New England Biolabs). Primers used were: ML0329-end-F2 5′, CCA TGA AGA CTT ACG GAT TTT TCT ACG; ML3250-start-F 5′, GAG ATC AGG AGG AAC ATC GG; ML3250-R 3′, GGA GAA ACA GAA GAA AAA ACA TAC TGT TTA G. Genomic sequence failed to amplify when an alternate 5′ primer for ML0329-end-F1 (TTT CGT TAA TAG CTA TGA AGG TTA TCG C) suggesting there may be base errors. The 1% agarose gel containing 5μ L ethidium bromide was visualized and photographed under UV light. 5μ L of Quick-Load 1kb DNA Ladder (New England Biolabs) were used for band size comparison.

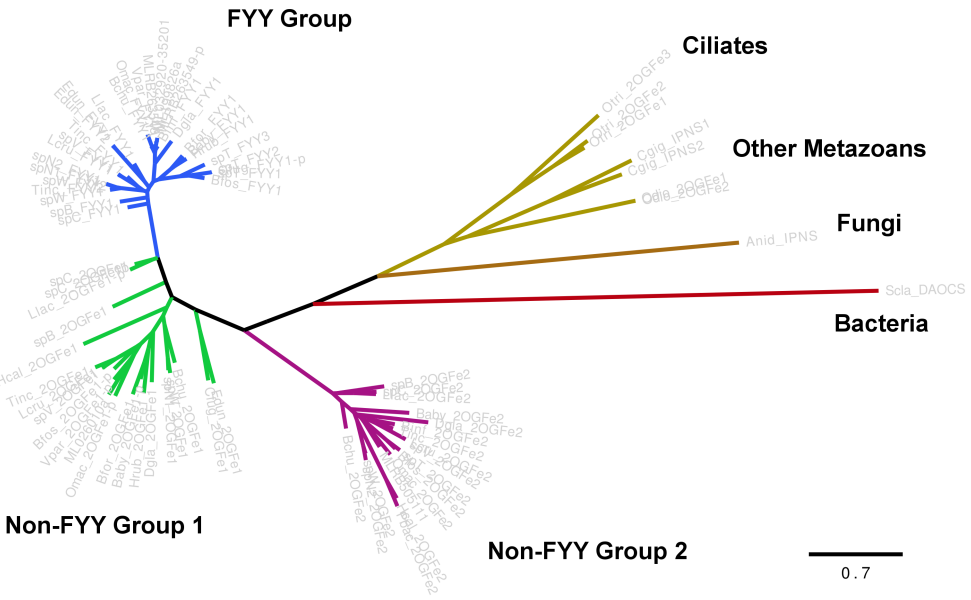

Figure 4.7: Maximum-likelihood tree of all putative ctenophore non-heme oxygenase proteins

Maximum-likelihood tree of all ctenophore non-heme oxygenase proteins including both FYY-containing (blue branches) and non-FYY groups (green and purple branches). Outgroups from top BLAST hits (gold branches) and model enzymes (brown and red branches) show long branches compared to the FYY proteins. Scale bar indicates substitutions per site. Partial or incomplete sequences are indicated by -*p* as in Figure 4.4. Species abbreviations are as follows: Anid, *Aspergillus nidulans*; Bfos, *Bathocyroe fosteria*; Bchu, *Bathyctena chuni*; Baby, *Beroe aybssicola*; Bfor, *Beroe forskali*; Binf, *Bolinopsis infundibulum*; Cfug, *Charistephane fugiens*; Cgig, *Crassostrea gigas*; Dgla, *Dryodora glandiformis*; Edun, *Euplokamis dunlapae*; Hrub, *Haeckelia rubra*; Hcal, *Hormiphora californensis*; Llac, *Lampea lactea*; Lcru, *Lampocteis cruentiventer*; ML, *Mnemiopsis leidyi*; Odio, *Oikopleura dioica*; Omac, *Ocyropsis maculata*; Otri, *Oxytricha trifallax*; Pbac, *Pleurobrachia bachei*; Scla, *Streptomyces clavuligerus*; Tinc, *Thalassocalyce inconstans*; spB, *Undescribed ctenophore B*; spC, *Undescribed ctenophore C*; spN1, *Undescribed ctenophore N1*; spN2, *Undescribed ctenophore N2*; spT, *Undescribed ctenophore T*; spV, *Undescribed ctenophore V*; Vpar, *Velamen parallelum*

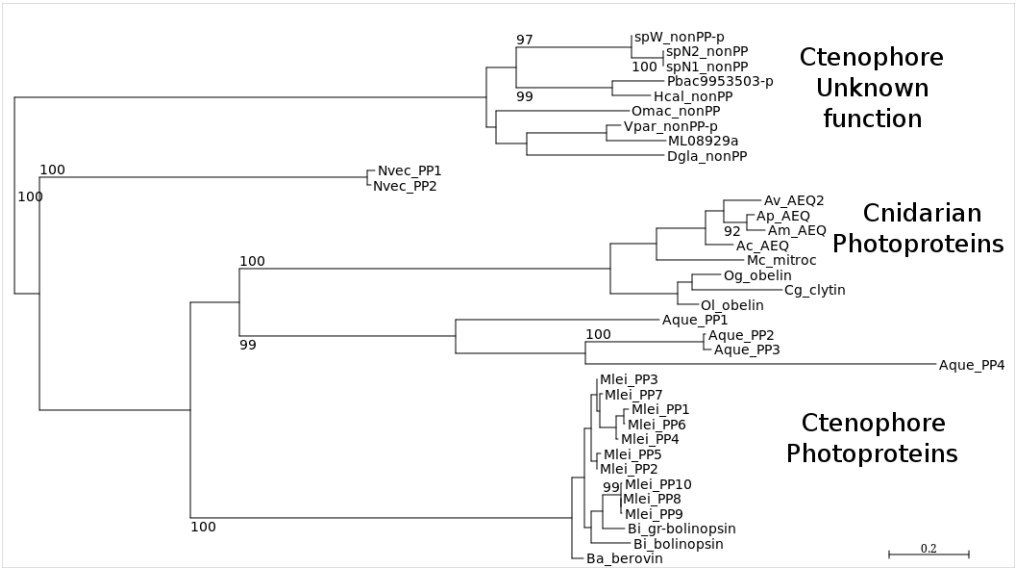

Figure 4.8: Maximum-likelihood tree of putative ctenophore photoprotein-like proteins Maximum-likelihood tree of recovered ctenophore photoprotein-like genes and a set of verified cnidarian and ctenophore photoproteins from Schnitzler *et al.* (2012) [83]. Bootstrap values above 90 are shown. Abbreviations are as in Figure 4.5 with a few changes and additions: Ac, *Aequorea coerulescens*; Aque, *Amphimedon queenslandica*; Am, *Amphimedon macrodactyla*; Ap, *Aequorea parva*; Av, *Aequorea victoria*; Ba, *Beroe aybssicola*; Bi, *Bolinopsis infundibulum*; Cg, *Clytia gregaria*; Mc, *Mitrocoma cellularia*; Nvec, *Nematostella vectensis*; Og, *Obelia geniculata*; Ol, *Obelia longissima*

Chapter 5

Conclusion

The ocean is the largest habitat on the planet, and many animals there create light to communicate in ways that science is only beginning to understand. However, this was not broadly appreciated until cloning and expression of the *Aequorea* green-fluorescent protein, which had transformed cell biology and ultimately was the subject of a Nobel Prize. Soon afterward there was a surge of papers aimed at modifying fluorescent proteins for more specialized applications and to develop new optical tools for time-lapse, calcium imaging, two-photon microscopy, *etc.* possibly with the hope that they would become as successful. Yet, in fact, it was the study of bioluminescence in *Aequorea* that brought about the incidental finding of the fluorescent protein that was interfering with purification of the photoprotein Aequorin [91]. Because of the calcium-sensitivity of Aequorin, it has promise as a calcium-sensor with much greater sensitivity than the calcium-binding dyes and fluorescent proteins, however the repeated addition of coelenterazine has limited its adoption.

Although *Aequorea* does not appear to produce coelenterazine [33], there is a wealth of work suggesting that ctenophores make it. The genes found in this work are promising candidates, due to both the presence of the "FYY" motif and the similarity to isopenicillin-N-synthases. While this may be very coincidental, the absence of these genes in non-luminous species is also telling. Of course, other genes are certainly lost in that lineage. Indeed, photoproteins were absent in both of the non-luminous species as well. Yet until the genomic scaffolds of *Pleurobrachia bachei* are made available, no true comparison is possible between the luminous *Mnemiopsis leidyi* and the non-luminous *Pleurobrachia bachei*, and even so the FYY proteins are a likely case.

Discovery of the coelenterazine biosynthetic pathway theoretically would enable making any organism self-luminous. This could work by coupling with either a luciferase (such as from *Renilla reniformis* or *Gaussia princeps*) or photoproteins from hydrozoans or ctenophores. Because there are many other clades that make use of coelenterazine [34], completely novel proteins may still be discovered that could produce light from coelenterazine under alternative conditions. This, like the previous reporter systems, has obvious biological applications, but could possible even be extended to more artistic projects, such as novelty glowing plants along walkways or golf courses. Engineering of eukaryotes was attempted already using the codon-optimized versions of all of the bacterial Lux genes [16, 51], though the efficiency of light emission was low. There may be metabolic issues with that particular system that have to be resolved,

or possibly other luminous systems work better in eukaryotic cells. Naturally, there is still a need to explore new possibilities in case an alternate luminescence system is more easily applied to plants or animals.

Along those lines, an abundant fluor was found in the luminous exudate of *Tomopteris* worms and from whole specimens that was identified as aloe-emodin. The connection between fluorescence and bioluminescence has been discussed at length, and while it is still speculative, based on the properties of other anthraquinones it is logical to deduce that aloe-emodin is the oxyluciferin. However, it is difficult to imagine what the luciferin might be. Early structure determination studies on the anthraquinone aloin (sometimes called barbaloin in the literature) had noted that it can be converted to aloe-emodin under relatively mild conditions [3,56]. Aloin is structurally similar to aloe-emodin, differing by the presence of a hexose sugar at the 10-position on the central ring rather than the carbonyl. These studies were attempting to determine the sugar group, however cleaving the sugar from the ring structure made the 10-position carbon reactive and labile to oxidation, resulting in aloe-emodin. This was hypothesized to occur via an intermediate structure called aloe-emodin anthrone, that is, the single carbonyl version. The anthrone could not be purified, probably because it was very unstable and rapidly oxidized into the stable aloe-emodin. In solution, intracellular conditions may not be suitable for the anthrone, but when bound stably by a protein it may be possible to secure a relatively unstable molecular until the needed light-producing conditions.

It should be no surprise that the technical applications in the lab are what drive this field of research. Nature has evolved under the conditions of the environment at the time, for the purposes of nature, and the systems that are used in the lab typically have to be engineered for the purposes of the lab. However it is important to remember that the world is large and biological systems have had billions of years to form stable, yet complicated interactions. Although new optical tools may enable manipulation and understanding of these systems, it was a chance finding from an unassuming jellyfish in the Puget Sound that brought about a revolution in technical advancements. We might fair better to pay attention to other such creatures in the world.

Bibliography

[1] G.J. Allman. Note on the phosphorescence of Beroe. *Proc roy soc Edinb*, 4:518–519, 1862.

[2] Roberto a Barrero, Brett Chapman, Yanfang Yang, Paula Moolhuijzen, Gabriel Keeble-Gagnère, Nan Zhang, Qi Tang, Matthew I Bellgard, and Deyou Qiu. De novo assembly of Euphorbia fischeriana root transcriptome identifies prostratin pathway related genes. *BMC genomics*, 12(1):600, January 2011.

[3] AJ Birch and FW Donovan. Barbaloin. I. Some observations on its structure. *Australian Journal of Chemistry*, 8(4):523–528, 1955.

[4] Benjamin J Blencowe, Sidrah Ahmad, and Leo J Lee. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes & development*, 23(12):1379–86, June 2009.

[5] M. R. Bowlby and J. F. Case. Flash kinetics and spatial patterns of bioluminescence in the copepod Gaussia princeps. *Marine Biology*, 110(3):329–336, October 1991.

[6] Gerhard Bringmann and Andreas Irmer. Acetogenic anthraquinones: biosynthetic convergence and chemical evidence of enzymatic cooperation in nature. *Phytochemistry Reviews*, 7(3):499–511, February 2008.

[7] Gerhard Bringmann, Torsten F Noll, Tobias a M Gulder, Matthias Grüne, Michael Dreyer, Christopher Wilde, Florian Pankewitz, Monika Hilker, Gail D Payne, Amanda L Jones, Michael Goodfellow, and Hans-Peter Fiedler. Different polyketide folding modes converge to an identical molecular architecture. *Nature chemical biology*, 2(8):429–33, August 2006.

[8] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11:94, January 2010.

[9] EJ Buskey and DE Stearns. The effects of starvation on bioluminescence potential and egg release of the copepod Metridia longa. *Journal of plankton research*, 13(4):885–893, 1991.

[10] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC bioinformatics*, 10:421, January 2009.

[11] a. K. Campbell and P. J. Herring. Imidazolopyrazine bioluminescence in copepods and other marine organisms. *Marine Biology*, 104(2):219–225, June 1990.

[12] Steven A Carlson and David M Hercules. Photoinduced luminescence of

9,10-anthraquinone. Primary photolysis of 9,10-dihydroxyanthracene. *Analytical Chemistry*, 45(11):1794–1799, September 1973.

[13] Steven A. Carlson and David M. Hercules. Studies on some intermediates and products of the photoreduction of 9,10-anthraquinone. *Photochemistry and Photobiology*, 17(2):123–131, February 1973.

[14] Andrew Clarke, Lesley J Holmes, and Deborah J Gore. Proximate and elemental composition of gelatinous zooplankton from the Southern Ocean. *Journal of Experimental Marine Biology and Ecology*, 155(1):55–68, February 1992.

[15] Nicole Cloonan, Alistair R R Forrest, Gabriel Kolle, Brooke B A Gardiner, Geoffrey J Faulkner, Mellissa K Brown, Darrin F Taylor, Anita L Steptoe, Shivangi Wani, Graeme Bethel, Alan J Robertson, Andrew C Perkins, Stephen J Bruce, Clarence C Lee, Swati S Ranade, Heather E Peckham, Jonathan M Manning, Kevin J McKernan, Sean M Grimmond, K Mellissa, C Andrew, and J Kevin. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods*, 5(7):613–9, July 2008.

[16] Dan M Close, Stacey S Patterson, Steven Ripp, Seung J Baek, John Sanseverino, and Gary S Sayler. Autonomous bioluminescent expression of the bacterial luciferase gene cassette (lux) in a mammalian cell line. *PloS one*, 5(8):e12441, January 2010.

[17] Jacob E Crawford, Wamdaogo M Guelbeogo, Antoine Sanou, Alphonse Traoré,

Kenneth D Vernick, N'Fale Sagnon, and Brian P Lazzaro. De novo transcriptome sequencing in Anopheles funestus using Illumina RNA-seq technology. *PloS one*, 5(12):e14202, January 2010.

[18] R Phillips Dales. Bioluminescence in Pelagic Polychaetes. *Journal of the Fisheries Research Board of Canada*, 28(10):1487–1489, October 1971.

[19] Casey Dunn. Siphonophores. *Current biology : CB*, 19(6):R233–4, March 2009.

[20] T Eisner and J Meinwald. Defensive secretions of arthropods. *Science (New York, N.Y.)*, 153(3742):1341–50, September 1966.

[21] M Eley, J Lee, J M Lhoste, C Y Lee, M J Cormier, and P Hemmerich. Bacterial bioluminescence. Comparisons of bioluminescence emission spectra, the fluorescence of luciferase reaction mixtures, and the fluorescence of flavin cations. *Biochemistry*, 9(14):2902–8, July 1970.

[22] Barbara Feldmeyer, Christopher W Wheat, Nicolas Krezdorn, Björn Rotter, and Markus Pfenninger. Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. *BMC genomics*, 12(1):317, January 2011.

[23] Harlan Foster and John H. Gardner. The Preparation and Hydrolysis of Some Polyhydroxyanthraquinone Glucosides. *Journal of the American Chemical Society*, 318(3):1934–1936, 1936.

[24] Denis Fox. Flavones. In *Animal Biochromes and Structural Colors*, pages 211–214. University of California Press, 1953.

[25] Warren R Francis, Lynne M Christianson, Rainer Kiko, Meghan L Powers, Nathan C Shaner, Steven H D Haddock, Steven H D Haddock, and B M C Genomics. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics*, 14(1):167, January 2013.

[26] Warren R. Francis, Meghan L. Powers, and Steven H. D. Haddock. Characterization of an anthraquinone fluor from the bioluminescent, pelagic polychaete Tomopteris. *Luminescence*, (February):n/a–n/a, April 2014.

[27] Gary Freeman and Geo.T. Reynolds. The development of bioluminescence in the ctenophore Mnemiopsis leidyi. *Developmental Biology*, 31(1):61–100, March 1973.

[28] Rohini Garg, Ravi K Patel, Akhilesh K Tyagi, and Mukesh Jain. De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA research : an international journal for rapid publication of reports on genes and genomes*, 18(1):53–63, February 2011.

[29] Q H Gibson and J W Hastings. The oxidation of reduced flavin mononucleotide by molecular oxygen. *The Biochemical journal*, 83:368–77, May 1962.

[30] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn a Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong

Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, 29(7):644–52, July 2011.

[31] S H Haddock, T J Rivers, and B H Robison. Can coelenterates make coelenterazine? Dietary requirement for luciferin in cnidarian bioluminescence. *Proceedings of the National Academy of Sciences of the United States of America*, 98(20):11148–51, September 2001.

[32] S. H. D. Haddock and J. F. Case. Bioluminescence spectra of shallow and deep-sea gelatinous zooplankton: ctenophores, medusae and siphonophores. *Marine Biology*, 133(3):571–582, April 1999.

[33] Steven H. D. Haddock and James F. Case. Not All Ctenophores Are Bioluminescent: Pleurobrachia. *Biological Bulletin*, 189(3):356, December 1995.

[34] Steven H.D. Haddock, Mark A. Moline, and James F. Case. Bioluminescence in the Sea. *Annual Review of Marine Science*, 2(1):443–493, January 2010.

[35] Matthew C Hale, Cory R McCormick, James R Jackson, and J Andrew Dewoody. Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (Acipenser fulvescens): the relative merits of normalization and rarefaction in gene discovery. *BMC genomics*, 10:203, January 2009.

[36] E. Newton Harvey. Bioluminescence and fluorescence in the living world. *American Journal of Physiology*, 77(3):555–561, 1926.

[37] Edmund Newton Harvey. *Bioluminescence*. Academic Press, 1st edition, 1952.

[38] EN Harvey. Additional data on the specificity of luciferin and luciferase, together with a general survey of this reaction. *American Journal of Physiology–Legacy ...*, 77(3):548–554, 1926.

[39] J W Hastings. Biological diversity, chemical mechanisms, and the evolutionary origins of bioluminescent systems. *Journal of molecular evolution*, 19(5):309–21, January 1983.

[40] J.Woodland Hastings. Chemistries and colors of bioluminescent reactions: a review. *Gene*, 173(1):5–11, January 1996.

[41] Usama W Hawas, Ahmed Atef El-Beih, and Ali M El-Halawany. Bioactive anthraquinones from endophytic fungus Aspergillus versicolor isolated from red sea algae. *Archives of pharmacal research*, 35(10):1749–56, October 2012.

[42] J. Evelyn Hay and L J Haynes. 605. The aloins. Part I. The structure of barbaloin. *Journal of the Chemical Society (Resumed)*, pages 3141–3147, 1956.

[43] K. Hori and M. J. Cormier. Structure and synthesis of a luciferin active in the bioluminescent systems of sea pansy (Renilla) and certain other bioluminescent coelenterates. *Chemiluminescence and bioluminescence*, pages 361–368, 1972.

[44] Kazuo Hori, Harry Charbonneau, R.C. Hart, and M.J. Cormier. Structure of native Renilla reniformis luciferin. *Proceedings of the National Academy of Sciences of the United States of America*, 74(10):4285, 1977.

[45] D F Howard, D W Phillips, T H Jones, and M S Blum. Anthraquinones and anthrones: Occurrence and defensive function in a Chrysomelid Beetle. *Naturwissenschaften*, 69(2):91–92, February 1982.

[46] Akira Kanakubo and Minoru Isobe. Isolation of brominated quinones showing chemiluminescence activity from luminous acorn worm, Ptychodera flava. *Bioorganic & medicinal chemistry*, 13(8):2741–7, April 2005.

[47] Akira Kanakubo, Kazushi Koga, Minoru Isobe, and Kenji Yoza. Tetrabromohydroquinone and riboflavin are possibly responsible for green luminescence in the luminous acorn worm, Ptychodera flava. *Luminescence : the journal of biological and chemical luminescence*, 20(6):397–400, 2005.

[48] Kazutaka Katoh and Daron M Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–80, April 2013.

[49] Toshito Kishi, Toshio Goto, Yoshimasa Hirata, Osamu Shimomura, and Frank H. Johnson. Cypridina bioluminescence I Sructure of luciferin. *Tetrahedron Letters*, 7(29):3427–3436, January 1966.

[50] Y. Kishi, H. Tanino, and T. Goto. The structure confirmation of the light-emitting moiety of bioluminescent jellyfish. *Tetrahedron Letters*, 13(27):2747–2748, 1972.

[51] Alexander Krichevsky, Benjamin Meyers, Alexander Vainstein, Pal Maliga, and Vitaly Citovsky. Autoluminescent Plants. *PLoS ONE*, 5(11):e15461, November 2010.

[52] Sujai Kumar and Mark L Blaxter. Comparing de novo assemblers for 454 transcriptome data. *BMC genomics*, 11(1):571, January 2010.

[53] M. I. Latz, T. M. Frank, and J. F. Case. Spectral composition of bioluminescence of epipelagic organisms from the Sargasso Sea. *Marine Biology*, 98(3):441–446, June 1988.

[54] Do Yup Lee, Benjamin P Bowen, and Trent R Northen. Mass spectrometry-based metabolomics, analysis of metabolite-protein interactions, and imaging. *BioTechniques*, 49(2):557–65, August 2010.

[55] Hairi Li, Michael T Lovci, Young-Soo Kwon, Michael G Rosenfeld, Xiang-Dong Fu, and Gene W Yeo. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51):20179–84, December 2008.

[56] R.E. Lister and R.R.A. Pride. The characterisation of crystalline and amorphous aloin. *Journal of Pharmacy and Pharmacology*, 11(S):278T, 1959.

[57] M Lucas and F Solano. Coelenterazine is a superoxide anion-sensitive chemiluminescent probe: its usefulness in the assay of respiratory burst in neutrophils. *Analytical biochemistry*, 206(2):273–7, November 1992.

[58] K E Malterud, T L Farbrot, A E Huse, and R B Sund. Antioxidant and radical scavenging effects of anthraquinones and anthrones. *Pharmacology*, 47 Suppl 1:77–85, October 1993.

[59] Jeffrey a Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82, October 2011.

[60] Tiina M Mattila, Jesper S Bechsgaard, Troels T Hansen, Mikkel H Schierup, and Trine Bilde. Orthologous genes identified by transcriptome sequencing in the spider genus Stegodyphus. *BMC genomics*, 13(1):70, January 2012.

[61] Frank McCapra and Martin Roth. Cyclisation of a dehydropeptide derivative: a model for cypridina luciferin biosynthesis. *Journal of the Chemical Society, Chemical Communications*, 13(15):894, January 1972.

[62] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8, July 2008.

[63] Marvin Mundry, Erich Bornberg-Bauer, Michael Sammeth, and Philine G D Feulner. Evaluating characteristics of de novo assembly software on 454 transcriptome data: a simulation approach. *PloS one*, 7(2):e31410, January 2012.

[64] Ugrappa Nagalakshmi, Zhong Wang, Karl Waern, Chong Shou, Debasish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881):1344–9, June 2008.

[65] M.T. Nicolas, J.M. Bassot, and O. Shimomura. Polynoidin: a membrane photoprotein isolated from the bioluminescent system of scale-worms. *Photochemistry and Photobiology*, 35(2):201–207, 1982.

[66] Yuichi Oba, Shin-ichi Kato, Makoto Ojika, and Satoshi Inouye. Biosynthesis of luciferin in the sea firefly, Cypridina hilgendorfii: l-tryptophan is a component in Cypridina luciferin. *Tetrahedron Letters*, 43(13):2389–2392, March 2002.

[67] Yuichi Oba, Shin-ichi Kato, Makoto Ojika, and Satoshi Inouye. Biosynthesis of Cypridina Luciferin in Cypridina noctiluca. *HETEROCYCLES*, 72(1):673, 2007.

[68] Yuichi Oba, Shin-Ichi Kato, Makoto Ojika, and Satoshi Inouye. Biosynthesis of coelenterazine in the deep-sea copepod, Metridia pacifica. *Biochemical and biophysical research communications*, 390(3):684–8, December 2009.

[69] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature genetics*, 40(12):1413–5, December 2008.

[70] Kevin Pang and Mark Q Martindale. Mnemiopsis leidyi Spawning and Embryo Collection. *CSH protocols*, 2008:pdb.prot5085, January 2008.

[71] Genis Parra, Keith Bradnam, and Ian Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, 23(9):1061–7, May 2007.

[72] Genis Parra, Keith Bradnam, Zemin Ning, Thomas Keane, and Ian Korf. Assessing the gene space in draft genomes. *Nucleic acids research*, 37(1):289–97, January 2009.

[73] Amos W. Peters. Phosphorescence in ctenophores. *Journal of Experimental Zoology*, 2(1):103–116, April 1905.

[74] Thomas Nordahl Petersen, Sø ren Brunak, Gunnar von Heijne, and Henrik Nielsen. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–6, January 2011.

[75] Jörn Piel. A polyketide synthase-peptide synthetase gene cluster from an uncultured bacterial symbiont of Paederus beetles. *Proceedings of the National Academy of Sciences of the United States of America*, 99(22):14002–7, October 2002.

[76] J F Rees, B de Wergifosse, O Noiset, M Dubuisson, B Janssens, and E M Thompson. The origins of marine bioluminescence: turning oxygen defence mechanisms into deep-sea communication tools. *The Journal of experimental biology*, 201(Pt 8):1211–21, April 1998.

[77] Peter L Roach, Ian J Clifton, C M Hensgens, N Shibata, C J Schofield, Janos Hajdu, and Jack E Baldwin. Structure of isopenicillin N synthase complexed with

substrate and the mechanism of penicillin formation. *Nature*, 387(6635):827–30, June 1997.

[78] Jonathan Romiguier, Vincent Ranwez, Emmanuel J P Douzery, and Nicolas Galtier. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome research*, 20(8):1001–9, August 2010.

[79] Joseph F. Ryan, Kevin Pang, Christine E. Schnitzler, Anh-dao a. D. Nguyen, R. Travis Moreland, David K. Simmons, Bernard J. Koch, Warren R. Francis, Paul Havlak, Stephen a. Smith, Nicholas H. Putnam, Steven H. D. Haddock, Casey W. Dunn, Tyra G. Wolfsberg, James C. Mullikin, Mark Q. Martindale, Andreas D. Baxevanis, Nisc Comparative, and Sequencing Program. The Genome of the Ctenophore Mnemiopsis leidyi and Its Implications for Cell Type Evolution. *Science*, 342(6164):1242592–1242592, December 2013.

[80] Meena Kishore Sakharkar, Bagavathi S Perumal, Kishore R Sakharkar, and Pandjassarame Kangueane. An analysis on gene architecture in human and mouse genomes. *In silico biology*, 5(4):347–65, January 2005.

[81] a Sali, L Potterton, F Yuan, H van Vlijmen, and M Karplus. Evaluation of comparative protein modeling by MODELLER. *Proteins*, 23(3):318–26, November 1995.

[82] Steven L Salzberg and James a Yorke. Beware of mis-assembled genomes. *Bioinformatics (Oxford, England)*, 21(24):4320–1, December 2005.

[83] Christine E Schnitzler, Kevin Pang, Meghan L Powers, Adam M Reitzel, Joseph F Ryan, David Simmons, Takashi Tada, Morgan Park, Jyoti Gupta, Shelise Y Brooks, Robert W Blakesley, Shozo Yokoyama, Steven Hd Haddock, Mark Q Martindale, and Andreas D Baxevanis. Genomic organization, evolution, and expression of photoprotein and opsin genes in Mnemiopsis leidyi: a new view of ctenophore photocytes. *BMC biology*, 10(1):107, January 2012.

[84] C J Schofield and Z Zhang. Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. *Current opinion in structural biology*, 9(6):722–31, December 1999.

[85] Marcel H Schulz, Daniel R Zerbino, Martin Vingron, and Ewan Birney. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, pages 1–12, February 2012.

[86] Cheng-Ying Shi, Hua Yang, Chao-Ling Wei, Oliver Yu, Zheng-Zhu Zhang, Chang-Jun Jiang, Jun Sun, Ye-Yun Li, Qi Chen, Tao Xia, and Xiao-Chun Wan. Deep sequencing of the Camellia sinensis transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds. *BMC genomics*, 12(1):131, January 2011.

[87] Osamu Shimomura. *Bioluminescence: Chemical Principles And Methods*. World Scientific Publishing Company, Incorporated, 2006.

[88] Osamu Shimomura and Frank H. Johnson. Chaetopterus photoprotein: crystallization and cofactor requirements for bioluminescence. *Science*, 159(3820):1239–1240, 1968.

[89] Osamu Shimomura and Frank H. Johnson. Calcium binding, quantum yield, and emitting molecule in Aequorin bioluminescence. *Nature*, 227(5265):1356–1357, September 1970.

[90] Osamu Shimomura and Frank H Johnson. Chemical nature of bioluminescence systems in coelenterates. *Proceedings of the National Academy of Sciences of the United States of America*, 72(4):1546–9, April 1975.

[91] Osamu Shimomura, Frank H. Johnson, and Yo Saiga. Extraction, Purification and Properties of Aequorin, a Bioluminescent Protein from the Luminous Hydromedusan,Aequorea. *Journal of Cellular and Comparative Physiology*, 59(3):223–239, June 1962.

[92] Osamu Shimomura and Katsunori Teranishi. Light-emitters involved in the luminescence of coelenterazine. *Luminescence : the journal of biological and chemical luminescence*, 15(1):51–8, 2000.

[93] Alexandros Stamatakis. RAxML-VI-HPC: maximum likelihood-based phyloge-

netic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, 22(21):2688–90, November 2006.

[94] Joachim Stauff and Peter Bartolmes. Chemiluminescence on Oxidative Formation of Triplet States of Anthrasemiquinone- and Anthraquinone-2-sulfonate. *Angewandte Chemie International Edition in English*, 9(4):307–308, April 1970.

[95] N Suzuki and Toshio Goto. Firefly bioluminescence II. Identification of 2-(6'-hydroxybenzothiazol-2'-yl)-4-hydroxythiazole as a product in the bioluminescence of firefly lanterns and as a product in the chemiluminescence of firefly luciferin in DMSO. *Tetrahedron Letters*, 22:2021–2024, 1971.

[96] Nobutaka Suzuki, Masamitsu Sato, Kunisuke Nishikawa, and Toshio Goto. Synthesis and spectral properties of 2-(6'-hydroxybenzothiazol-2'-yl)-4-hydroxythiazole, a possible emitting species in the firefly bioluminescence. *Tetrahedron Letters*, 10(53):4683–4684, January 1969.

[97] Gerard Talavera and Jose Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology*, 56(4):564–77, August 2007.

[98] Sonia Tarazona, Fernando García-Alcalde, Joaquín Dopazo, Alberto Ferrer, and Ana Conesa. Differential expression in RNA-seq: a matter of depth. *Genome research*, 21(12):2213–23, December 2011.

[99] Roman L Tatusov, Natalie D Fedorova, John D Jackson, Aviva R Jacobs, Boris

Kiryutin, Eugene V Koonin, Dmitri M Krylov, Raja Mazumder, Sergei L Mekhedov, Anastasia N Nikolskaya, B Sridhar Rao, Sergei Smirnov, Alexander V Sverdlov, Sona Vasudevan, Yuri I Wolf, Jodie J Yin, and Darren a Natale. The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, 4:41, September 2003.

[100] B. Terio. Su un pigmento fluorescente presente nella pinna del remo dorsale di un annelide polichete (Tomopteris septentrionalis Steenstrup). *Boll. Soc. Ital. Biol. Sper.*, 36:725–727, 1960.

[101] B. Terio. Possibili interrelazioni tra bioluminescenza e fluorescenza di materiali fotosensibili presenti nelle pinne e sui parapodi dei Tomopterid. *Atti Soc. Peloritana Sci. Fis. Mat. Natur*, 10:1–11, 1964.

[102] C. M. Thomson, P. J. Herring, and a. K. Campbell. Coelenterazine distribution and luciferase characteristics in oceanic decapod crustaceans. *Marine Biology*, 124(2):197–207, December 1995.

[103] Catherine M. Thomson, Peter J. Herring, and Anthony K. Campbell. Evidence For De Novo Biosynthesis of Coelenterazine in the Bioluminescent Midwater Shrimp, Systellaspis Debilis C. *Journal of the Marine Biological Association of the United Kingdom*, 75(01):165, May 1995.

[104] Ronald Hunter Thomson. *Naturally Occurring Quinones*. Academic Press, 1st edition, 1971.

[105] Ronald Hunter Thomson. *Naturally Occurring Quinones III: Recent Advances*. Chapman and Hall, 1st edition, 1987.

[106] Bing Tian and Yuejin Hua. Concentration-dependence of prooxidant and antioxidant effects of aloin and aloe-emodin on DNA. *Food Chemistry*, 91(3):413–418, July 2005.

[107] Richard D. Towner, Harold A. Neufeld, and Philip B. Shevlin. Some characteristics of riboflavin chemiluminescence. *Archives of Biochemistry and Biophysics*, 137(1):102–108, March 1970.

[108] Wayne G Wamer, Peter Vath, and Daniel E Falvey. In vitro studies on the photobiological properties of aloe emodin and aloin A. *Free Radical Biology and Medicine*, 34(2):233–242, January 2003.

[109] Xiao-Wei Wang, Jun-Bo Luan, Jun-Min Li, Yan-Yuan Bao, Chuan-Xi Zhang, and Shu-Sheng Liu. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC genomics*, 11:400, January 2010.

[110] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[111] Robert H Waterston, Kerstin Lindblad-Toh, Ewan Birney, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–62, December 2002.

[112] By John Weatherston. The Chemistry of Arthropod Defensive Substances. *Quarterly Reviews*, 1966.

[113] Klaus Wolkenstein, Wolfgang Schoefberger, Norbert Müller, and Tatsuo Oji. Proisocrinins A-F, brominated anthraquinone pigments from the stalked crinoid Proisocrinus ruberrimus. *Journal of natural products*, 72(11):2036–9, November 2009.

[114] Deying Yang, Yan Fu, Xuhang Wu, Yue Xie, Huaming Nie, Lin Chen, Xiang Nong, Xiaobin Gu, Shuxian Wang, Xuerong Peng, Ning Yan, Runhui Zhang, Wanpeng Zheng, and Guangyou Yang. Annotation of the transcriptome from Taenia pisiformis and its comparative analysis with three Taeniidae species. *PloS one*, 7(4):e32283, January 2012.

[115] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–91, August 2007.

[116] Gow-chin Yen. Antioxidant activity of anthraquinones and anthrone. *Food Chemistry*, 70(4):437–441, September 2000.

[117] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18(5):821–9, May 2008.

[118] Zhihong Zhang, Jingshan Ren, and DK Stammers. Structural origins of the selectivity of the trifunctional oxygenase clavaminic acid synthase. *Nature structural & molecular biology*, 7(2):127–133, 2000.

[119] Qiong-Yi Zhao, Yi Wang, Yi-Meng Kong, Da Luo, Xuan Li, and Pei Hao. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC Bioinformatics*, 12(Suppl 14):S2, 2011.