TRANSCRIPTOME *DE NOVO* ASSEMBLY, CLUSTERING AND ANNOTATION OF NOVEL TRANSCRIPTS

by

FATEMEH (SHAADI) POOYAEI MEHR

A dissertation submitted to the Graduate Faculty in Biology in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

2013

This manuscript has been read and accepted for the
Graduate Faculty in Biology in satisfaction of the
dissertation requirement for the degree of Doctor of Philosophy.

[typed name] David Gruber

09/09/2013 [required signature] David Gruber

Date Chair of Examining Committee

[typed name] Rob DeSalle

09/09/2013 [required signature] Rob DeSalle

Date Executive Officer

[typed name] Eugenia Naro-Maciel

[typed name] Estefania Rodriguez

[typed name] Vincent A. Pieribone

Supervisory committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

TRANSCRIPTOME *DE NOVO* ASSEMBLY, CLUSTERING AND ANNOTATION OF NOVEL TRANSCRIPTS

by

Fatemeh (Shaadi) Pooyaei Mehr

Advisors: Professor David Gruber / Robert DeSalle

Recent advances in Next Generation Sequencing (NGS) have allowed for unparalleled access to genetic information for organisms in both the functional and phylogenetic realms of biology. Analysis of the RNA transcripts of cells of organisms using Next Generation Sequencing (called RNA-seq) has opened doors for unique insights into the genomic complexity of organisms and has provided researchers with invaluable tools for analysis of function of gene products and phylogenetic relatedness. Application of this method has moved beyond model organisms. It has provided a lot of potentials, in ecological research and comparative transcriptomics, in non-model organisms. This thesis presents an overview on existing applications of RNA-seq in non-model organisms. Furthermore, it presents a new clustering design on handling the data, which led to identification of twelve new fluorescent protein isoforms in corals. In addition, *de novo* assembly and annotation of the data from polychaete *Hermodice carunculata* made possible the identification of one new phylogenetic marker and eight bioluminescent

protein isoforms. Also, twelve new bilirubin-induced fluorescent proteins were identified from false moray eel *Kaupichthys hyproroides*. This approach can be applied on any other data.

Acknowledgements

I would like to take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project. I would like to show my greatest appreciation to professors Rob DeSalle and David Gruber. Without their support and guidance this project would have not been materialized. I am grateful for their continual stimulate discussion on all the aspects concerning this thesis.

I would like to give my deep appreciation to my parents, family and friends.

Finally, I would like to thank the staff of the Biology Department, Faculty of Biology at City University of New York, Baruch College and Sackler Institute for Comparative Genomcis, Faculty of Invertebrate Zoology and Molecular Systematics, at American Museum of Natural History.

Author

Fatemeh (Shaadi) Pooyaei Mehr

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES AND BOXES

Chapter 1: Introduction

Part 1:

C1.1: Fluorescence and Bioluminescence

C1.1.1: Fluorescence

The mechanism of light production through absorbing photons, which temporarily excites electrons to higher energy states, distinguishes fluorescence from other natural optical phenomena such as bioluminescence and phosphorescence. As the excited electrons relax to their basal state, they release energy at a longer wavelength. Since this excitation and relaxation happens almost within picoseconds, the emission of light is visible while the specimen is being illuminated. Chlorophyll, phycobiliproteins, and green fluorescent proteins (GFPs) are examples of fluorescent proteins.

GFP was first detected in the hydromedusa *Aequorea victoria* [1, 2], and was at the time considered a component of aequorin, a bioluminescent protein that was then thought to be made up of multiple components. Aequorin produces blue bioluminescence in the presence of Ca^{+2}, and homologues have been reported in the sea pansy *Renilla reniformis* [3] and in the hydroids *Obelia [4]*and *Clytia gregaria* [5] . The blue bioluminescent light (470nm) that is radiated by aequorin excites the closely located GFP, and green light is emitted.

There are about 200 GFPs in the protein Data Bank distributed over about 125 species. Up until the discovery of a luminescent species of eel [6], GFPs were defined as a group of structurally homologous proteins from marine bioluminescent animals and non-bioluminescent reef organisms, mostly corals, sea anemone [7] and others such as species of comb jellies [8] and lancelets [9, 10]. Despite the high degree of structural homology within this protein family, the amino acid sequence identity is very low. However, the newly identified fluorescent protein (FP) from the luminescent eel is the exception to this rule. A puzzling observation about this protein is the fact that it requires bilirubin as its chromophore for being luminous. In addition, there are structural differences between this Billirubin-induced Fluorescent Protein (BiFP) and canonical GFP such as wider beta barrel, and maturation independent of oxygen in BiFP compared to GFP [6].

A major established application of FPs is in Forster resonance energy transfer (FRET) [3] in which a fluorescent or bioluminescent donor is coupled with a fluorescent acceptor. Light that is being emitted from the donor results in excitation of the acceptor. In the process of this excitation, the fluorophore spontaneously forms by an oxidative cyclization of three amino acids, and in the absence of a fluorescence cofactor ligand. While the acceptor molecule is always fluorescent, the donor can be either fluorescent itself or bioluminescent. FRET technology can be incorporated in microscopy, electrophoresis, chromatographic assays and flow cytometry [11].

In order for the FRET to be efficient some fundamental criteria should be met: 1) substantial spectral overlap between the donor and acceptor, 2) strong fluorescence of the

donor, 3) matched orientation of transition dipoles, 4) and ~50Å separation between donor and acceptor [5]. In order to test the donor-acceptor coupling possibility and their applicability for FRET, a test has to be carried out in solution. For instance, it has been suggested that an efficient complex forms in the case of *Renilla* luciferase and its conjugate GFP [12] but not with other bioluminescent proteins [13]. In order to expand our repertoire of FRET pairs, it is important to identify and test a diverse range of bioluminescent molecules, both fluorescent and luminescent.

C1.1.2: Bioluminescence

Many species, from bacteria to fishes, utilize bioluminescence for a diverse range of functions such as predation (dragonfish [14]), defense (arrow worms [15]) and intraspecific communication (fireflies). In fact, many marine animals acquire their primary visual stimulation via their biologically generated lights rather than sunlight [8]. Bioluminescence is produced as a result of a chemical reaction. The oxidation of a light-emitting molecule, luciferin, generates light. An enzyme, called luciferase, controls the rate of the bioluminescence reaction. These photoproteins are triggered to produce light upon binding to ions such as Ca2+ or Mg2+ [16].

While luciferins are conserved, luciferases and photoproteins are variable, and have been reported from a wide range of species. Luciferins of marine organism are reported to exist in five major types including 1) Bacterial, 2) Dinoflagellate, 3) Cypridina, 4) Coelenterazine, 5) other unknown types. For instance, Coelenterazine is thought to be the

light emitter in nine phyla, including protozoans, jellyfish, crustaceans, arrow worms, mollusks and vertebrates. The accepted hypothesis implies that luciferins are acquired though the diet [17]. However, their ultimate origins remain unknown.

It has been suggested that each luminescent hydrozoan genera (*Aequorea, Obelia, Mitrocoma,* and *Clytia*) has one or more genes encoding bioluminescent photoproteins with detectable sequence homology within hydrozoans, but with little homology to luciferases from other cnidarians [8]. Chapter 3 of this thesis work is dedicated to test this hypothesis.

The initial step required for the identification of any novel protein (such as fluorescent and luminescent proteins) is the identification of its coding sequence. In the second part of this chapter, I present the transcriptomic deep sequencing approach that I used to find novel proteins from species whose genomes are not available. I will furthermore describe how I used this approach to identify novel bioluminescent proteins.

Part 2:

C1.2.1: RNA-seq shines new light on the Tree of Life: How to use recent methodological advances to break the transcriptome barrier

C1.2.2: Abstract

Recent advances in Next Generation Sequencing (NGS) have allowed for unparalleled access to genetic information for organisms in both the functional and phylogenetic realms of biology. Analysis of the RNA transcripts of cells of organisms using NGS (called RNA-seq) has opened doors for unique insights into the genomic complexity of organisms and has provided researchers with invaluable tools for analysis of function of gene products and phylogenetic relatedness. We describe some of the recent approaches to collecting, manipulating and analyzing RNA-seq data and provide a description of pipelines for analysis of RNA-seq data. We discuss four major phases of data analysis in RNA-seq studies – quality control, assembly, annotation and quantification – and provide critical assessment of tools developed for these steps in data management for RNA-seq studies. Finally, we provide a sample pipeline for how most RNA-seq studies can and should proceed.

C1.2.3: Introduction

Over the past five years, massively parallel sequencing of cDNA libraries (RNA-Seq) has proven itself as an efficient, robust and cost-effective means of rapidly obtaining entire transcriptomes of organisms. Applications of RNA-Seq such as expression profiling, monitoring splicing events, and novel transcript identification have been widely adopted in model organisms. Yet, there is an increasing demand to apply RNA-Seq to improve and deepen ecological and phylogenetic studies for organisms with little or no available sequence information. Within non-model organisms, RNA-Seq approaches are beginning to be used for applications such as developing novel gene markers for phylogenetic studies [18], allele frequency detection for population genetics [19], novel toxin or protein identification [20], metabolic pathway reconstruction for functional studies [21], and even differential gene expression profiling [22]. Here, we review RNA-Seq applications, advances and challenges and explore their potential to transform our genomic understanding of non-model organisms. We also provide a brief application strategy for utilizing short-read RNA-seq data for investigations into non-model organisms.

Information on the expression of transcript isoforms in organisms is a valuable data regarding the functional status of a cell. Many methods for exploring this landscape have been implemented, providing important mechanistic insights into the regulation and expression of RNA molecules. Quantitative analysis of transcript abundance was developed using *in situ* hybridization as early as 1980 [23], followed by qRT-PCR [24]

and high-throughput trancriptome, microarrays, in 1995 [25]. Microarrays later became the primary method for monitoring gene expression fluctuation, surveying genome wide DNA and protein binding interactions and long-range DNA interactions [26, 27].

Yet, as the new sequencing methods emerge, the existing microarray applications are gradually shifting to sequence-based methods [28]. Generating complete transcriptomes using Next-Generation Sequencing (NGS) methods, known as RNA-Seq, not only provides high-resolution and dynamic range of gene expression measurements, but also provides the opportunity to find novel transcribed sequences, and alternative splice variants in organisms with available genomes [29, 30].

Surging interest in exploring the landscape of the comparative genomics in non-model organisms has led to the application of NGS methods to generating cDNA libraries. These libraries have an enormous sequencing depth and better reproducibility, producing at least 100 to 10,000 times higher throughput than classical Sanger sequencing [31]. This level of sensitivity in examination of thousands of transcripts from species with no available genome renders it easy to perform a wide range of biological studies including phylogenomics [32], regulatory gene discovery [21, 22, 33, 34], molecular marker development [35, 36], single nucleotide polymorphism (SNP) identification for trait adaptation [37, 38], haplotype detection [19, 39], and differential gene expression profiling [22, 40, 41].

Although most of the toolkits available for transcriptomic landscape analysis are based on research involving model organisms, there is an unprecedented potential for research involving other less well-studied organisms. This could be particularly useful for rare or endangered species in which access to samples is limited or those of economic and ecological importance. One of the major pitfalls in RNA-seq data analysis is the bioinformatics challenge dealing with the massive amounts of data generated (300-600Gb per lane). These bioinfomatics challenges include maintaining computational resources as well as creating novel software to perform downstream analysis. Since most of the available data analysis tools are based on their application in model organisms, their application in non-model organism requires *ad hoc* implementation studies of the genomics of organisms they were not designed for. However, as the number of available software packages expands, so will the scope of RNA-seq applications to better understand and investigate the Tree of Life.

C1.2.4: Available sequencing platforms

Many massively parallel DNA sequencing platforms have been developed in the last decade [42]. Currently there are a handful of platforms that researchers use for NGS. Read length and depth coverage in platforms such as 454, Illumina, Helicos, SOLiD ABI, PacBio, and IonTorrent vary (Table C1.1). For example, Illumina and SOLiD ABI are categorized under short-read sequencer methods and generate relatively short paired-end (PE) reads (2 × 100bp for Illumina and 2 × 75bp for SOLiD ABI). They also sequence with considerably high depth (200Gb for Illumina and 300Gb for SOliD ABI). On the

other hand, 454 has reached read lengths of 500 base pairs, but its depth of coverage is much less. While technical details vary among these platforms, they all generate single DNA molecules at small dilutions. They allow direct production of DNA fragment with no need for plasmid or vector for cloning [28].

C1.2.5: RNA-seq advances in model organisms

RNA-seq was developed in 2008 [29] for transcriptomic analysis, and has been widely applied in model organisms to provide information on genes and their alternate isoform mapping [43-45], gene expression [28, 41, 46], non-coding RNA identification [47]post-transcriptional single nucleotide variation [48], and gene fusions [49]. It has been effective at determining exon boundaries in *Saccharomyces cerevisiae* [30], *Caenorhabditis elegans [50], Mus musculus* [29], and *Homo sapiens* [51]. RNA-seq also provides better sensitivity in predicting exon boundaries compared to tilling arrays [52]. The quest for improvement in transcription start site mapping, strand specific measurements, and small RNA characterization is constantly underway in model species [53].

C1.2.6: RNA-seq potentials and challenges in non-model organisms

The advent of RNA-seq makes it possible to quickly and cheaply generate transcriptome sequences in any organism for which RNA can be obtained. It provides access to the coding sequences of hundreds of genes and opens up a diverse range of previously

unexplored questions in phylogenomics, population genetics, and functional genomics research in non-model organisms. We provide selected examples of applications to each endeavor below.

Phylogenomics: Within the scope of insect phylogenomics, using long-read NGS transcriptomic data, a recent study explored the unresolved phylogenetic issues within the Polyneopteran orders and provided support for monophyletic Polyneoptera [18]. Furthermore, recent implementation of transcriptomic data, combined with morphological data from ostracod and pancrustacean taxa, has led to better phylogenetic resolution of long-standing questions such as interordinal relationship related to this clade of organisms [54].

Population Genetics: Genetic variation between populations can be utilized as molecular markers in population genetics studies. In this regard, RNA-seq can be use to find genetic variation in microsatellites, simple sequence repeats (SSR), and single-nucleotide polymorphisms (SNP). For instance, recent work on the rainbow trout (*Oncorhynchus mykiss*) reported SNPs in genes associated with variation in growth rate [37]. In another study, comparative SNP analysis between two alfalfa (*Medicago sativa*) genotypes led to the identification of candidate genes that may play a role in stem development and different cell composition [55]. Furthermore, identification of simple sequence repeats (SSRs), as a molecular marker, in the chickpea (*Cicer arietinum*) was accomplished using the application of transcriptome sequencing and analysis [35].

Functional Genomics: Other applications focus on regulatory gene discovery and metabolic pathway reconstruction. For instance, tea-specific major metabolic pathways were reconstructed with deep RNA sequencing [21]. In addition, applications of this method have led to identification of genes involved in sex differentiation, cold adaptation and immunity in the tuatara (*Sphenodon punctatus*), the only surviving species of the reptilian order Rhynchocephalia [56].

One advantage of RNA-seq is the ability to count the number of reads mapped on certain areas of the transcriptome. This provides the opportunity of transcript abundance measurement and quantification. Therefore it provides the possibility of capturing differentially expressed genes under different conditions. Using this approach, for example, a recent study captured differential up-regulation of anthocyanin biosynthesis during the fruit ripening processes in the Chinese bayberry (*Myrica rubra*) [22].

RNA-seq provides intensely rich datasets that can be probed to investigate multitude research questions. Therefore, the application of these methods has expanded the potential for a much deeper genomic and evolutionary understanding. Depending on the questions and available computational resources, certain strategies can be applied.

C1.2.7: RNA-seq application strategy in non-model organisms

RNA-seq data analysis in species with no available genome sequence consists of four major steps; 1) Quality control and filtering, 2) *De novo* assembly, 3) Annotation and 4)

Transcript abundance quantification (**Fig. C1.1A; C1.1B; C1.1C**). In this review we focus on short-read (i.e. Illumina) *de novo* assembly. Methods for other platforms vary, but in general the approaches described below are a good starting point for understanding the complexity of analysis of NGS RNA-Seq.

C1.2.7.1: 1) Quality control and filtering

Before assembly, quality assessment is a critical step to guarantee high accuracy of the data for the subsequent steps. Sometimes, it is necessary to filter the data by removing linkers, and low quality sequences which contain either numerous Ns (ambiguously called nucleotides) or are shorter than 17-20bp. Several programs and toolkits have been developed to perform various aspects of these steps in the workflow. FastQC is a useful program to generate a quality overview [57]. NGSQC provides a collection of command line programs to process and filter low quality reads [58]. Programs such as AdapterRemoval [59] and Cutadapt [60] locate and remove adapter residues from the read. Other programs such as Clean-Reads [61] and Seqtk [62] filter out the reads under the minimum length and score quality criteria, and Clean-Reads can also remove vectors and adapters (Table C1.2). Lastly, the filtered reads should be deposited to Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) at the National Center for Biotechnology Information (NCBI) or some other archival resource.

C1.2.7.2: 2) *De novo* assembly

Assembling transcriptomes from short-read RNA-seq data is a computationally intensive task. In genome-dependent transcriptome reconstruction, researchers take advantage of the mapping-first approach to analyze the transcriptome data. The main goal in this kind of analysis is gene expression measurements or novel transcript or gene variant discovery. In such analysis, exon guided assemblies are carried out using programs such as G.Mo.R.Se [63], Scripture [64], and Cufflinks [65]. These programs are useful for capturing differentially expressed genes in samples with well-annotated genomes.

On the other hand, the surge of interest in the application of RNA-seq methods in non-model species has led to the development of several genome-independent (*de novo*) transcriptome reconstruction pipelines. Initially, the short-read assembly programs were mainly based on *de Bruijn* graph genomic assemblers (Box C1.1) such as Velvet, ABySS, Soapdenovo2 [66-68]. *De Bruijn* graph assembly algorithms do not use a reference genome and attempt to classify reads into distinct overlapping components (k-mer) that represent all gene isoforms. The major problem with these genome-based programs lies in their assumption that the read coverage is even. However, the depth of coverage fluctuates significantly between transcripts isoforms. This phenomenon made it hard to isolate erroneous sequences with low coverage when the mentioned programs were used for transcriptome data.

Recently these programs adopted post-processing pipelines of *de novo* genome assemblers such as Oases [69], Trans-ABySS [70] and SOAPdenovo-Trans [71] to account for uneven range of coverage depth. The common denominator of these pipelines is that they run the assembler at different k-mer lengths and merge these assemblies into one [72]. Therefore, they are called mulit k-mer assemblers. The purpose of this method is to capture both low expressed transcripts (lower values of k) and highly expressed transcripts (high values of k) by combining a range of k-mers [73].

The pipeline presented in Trans-ABySS [70], SOAPdenovo-Trans [71] also detects alternative splicing variants by searching for connected groups of contigs. Alternatively, Trinity [74] is designed to recover assembly paths supported by actual reads and removes ambiguous edges. It uses only one single k-mer to reconstruct highly expressed transcripts to full-length assemblies. Trinity has been used to reconstruct full-length splice variants in yeast, mouse and whitefly [74, 75]. Oases [69], uses a greedy topological strategy, similar to Trinity, and combines it with multi k-mer strategy used in Velvet. Oases claims to generate longer transcript assemblies compared to other methods [69]. Another program, Rnnotator [76], orients the assembled sequences in the correct mRNA sense strand orientation, which can be useful when the library preparation is not strand specific. In contrast to other programs, KISSPLICE [77] does not generate the entire transcriptome, instead it only outputs the variable regions of transcripts in different conditions. It identifies SNPs, indels and alternative splicing events variants under each experimental condition, which are useful in trait adaption or population structure studies.

Choice of assembler program depends on depth of coverage, complexity of the genome, available computational resources (run-time), and the purpose of generating the data. Quality statistics of assembled sequences can be used to assess the accuracy of the assembled transcripts. The quality statistics varies depending on 1) number of reads being mapped back to transcripts (RMBT), 2) N50 (50th percentile of length distribution), 3) sequences mean length 4) longest generated transcripts. It is essential to measure these parameters to choose the best assembly program, and make sure the contigs with low coverage are removed from the assembly before progressing to the annotation stage. Since these *de novo* assembly methods are new, tools to evaluate their performance are yet being developed. Most of the existing statistic benchmarking toolkits for *de novo* transcriptome assembly evaluation and quality control require *ad hoc* implementation. Perl scripts for measuring some of these statistics such as N50, mean length measurements and RMBT calculations, are available at (https://github.com/spooyaei/NGS-eval).

Programs such as Trinity requires extensive computational resources (100 hours of 80 G RAM to assemble 117 Mb genome), but it has been shown to perform the best for the percentage of reads mapped back to transcripts measurement, when applied to *Drosophila melanogaster* transcriptome [72]. However, the performance and specification of these programs require detailed evaluation in more datasets with different complexities. Based on recent work on the central nervous system of *Lymnaea stagnalis*, OASES, Trinity and Rnnotator performed at the same level in terms of N50 length, maximum contig length and contig number.

After the assembly step, an optional clustering step can be used to remove sequences with 99% identity to reduce the size of the data set. Some of the recently developed clustering programs such as CD-hit [78], USEARCH [77] can accomplish this clustering step. Lastly, the assembled sequences should be deposited in Transcriptome Shotgun Assembly (http://www.ncbi.nlm.nih.gov/subs/tsa/) at the National Center for Biotechnology Information (NCBI) or in other archival resources.

De novo assembly programs serve as a complementary approach when genome sequences are unavailable or partially annotated (Table C1.3). However successfully applying these programs presents unique challenges to bioinformaticians. Further analysis and comparative studies in different species, with different transcriptome complexity are required to provide guidelines and help direct the choice of assembly programs for RNA-Seq.

C1.2.7.3: 3) Annotation

Homology searches against available annotated databases are the backbone of the annotation step [79]. The parameters of the homology search step are dependent on the availability of the database, computational recourses and the type of questions that are addressed. Typically the annotation represents the percentage of the homology match between the assembled sequences and the reference sequences.

If the genome of a subject species is not available, the sets of assembled contigs are aligned to a reference transcriptome or proteome for the annotation. The reference transcriptome most commonly is the closest evolutionary relative of the subject species. Typically the reciprocal homology search is performed using BLAST [80]. This means that the assembled transcriptome will be searched against the reference dataset, and all the pairwise connections should be the best hit in both directions. Generating Open Reading Frames (ORF) [81] from the transcriptome is optional and it depends on the program of choice for the homology search.

One possible approach is to cluster the matched sequences around each homologous reference sequence (**Fig. C1.1C**). For pre-clustering step, a list of matches with the best bit-score is prepared and used as the input for these clustering programs. There are two approaches to accomplish clustering. One approach is to use similarity based criteria to establish a cluster. TribeMCL [82] is a useful program for clustering using this approach. The second approach is to use a tree based approach to assess homology. OrthologID [83] is an example of this approach. Homologous/orthologous sequences are then used for further analysis such as functional annotation and transcript isoform identification.

For functional annotation and metabolic pathway reconstruction, Gene Onthology (GO) [84] terms and categories are assigned to the sequences with matches in the NCBI protein (NR) and nucleotide (NT) databases, SWISS-PROT [85], and Refseq [86]. Other alternatives for functional annotation are domain-ID assignment to sequences by using

InterPro scan [87] algorithm and Blast2GO [88] which is a functional annotation (Table C1.4).

C1.2.7.4: 4) Transcript abundance quantification

Another major goal of RNA-Seq studies is to quantify transcript/allele specific abundance, which facilitates capturing the gene expression heterogeneity. This can be achieved by mapping reads to a well-annotated reference genome, or to the annotated assembled transcripts.

Expression profiling in RNA-seq is based on read count, which requires normalization to account for variation in RNA fragmentation and variation between different runs. In order to normalize read count, the total number of reads per kilobase of transcripts per million mapped reads (RPKM) is used as a metric [29]. However, in case of paired-end reads, the normalization is based on total fragment counts per kilobase of transcript per million reads (FPKM).

In order to compute FPKM and RPKM, short-read alignment programs such as BWA [89], BOWTIE [90], BOWTIE2 [91], MAQ [92], and SSAHA2 [93], among many others, can be used. The alignments which are in SAM (Sequence Alignment/MAP) format can be manipulated (sorting, merging, changing the test format to binary format (BAM), and generating alignment per position format) with programs such as SAMtools [94]. Application of SAMtools can provide SNP or variation calling too. In addition, the

generated alignments can be visualized with programs such as IGV [95] and Bamview [96] (Table C1.5).

In addition to command line tools, some bioinformatics groups have developed genome utility tools and implemented some of these programs for NGS data handling and analysis [97]. For example, we have implemented a BOWTIE runner into a Perl script for FPKM calculation (https://github.com/spooyaei/NGS-eval). Regardless of program used, accurate alignment to repetitive elements, or alignment to multiple locations is still hard to resolve. According to a recent benchmarking of some of the short read alignment programs, there is no "best-tool program" and each tool performs best under certain conditions [98]. Evaluation of short-read mapping is an actively researched problem with many aspects to be addressed. Further comparison of the performance of the current transcriptome assemblers is still required to help users decide which mapping program best fits their intended application.

C1.2.8: Conclusion

While massively parallel sequencing of cDNA libraries (RNA-Seq) is a fairly new technology, it has already led to revolutionary changes in genomics research and other related fields. Although it is proven to be an efficient, robust and cost-effective means of rapidly obtaining entire transcriptomes, how the data handling is pipelined is still being developed. The programs and technologies reviewed in this article for the application of RNA-seq can be applied to any species, from model organisms to the newly discovered.

The application strategies for utilizing short-read RNA-seq data presented in this article may be helpful to those considering using this approach to test wide ranges of biological hypotheses in non-model species. However, challenging steps still lie ahead of the development of methods to utilize these data for large-scale genome annotation.

C1.2.9: Glossary

BAM (Binary Alignment/MAP) format: Binary format for storing (compressing) SAM files.

Bit-score: Log scale version of an alignment score (A numerical value that describes the overall quality of an alignment).

BLAST: A program that finds similar regions between sequences.

Contigs: Set of overlapping DNA that represents a consensus region of DNA.

Coverage: The percent of a reference genome/transcriptome that is covered by read sequences.

De Bruijn graph: A directed graph representing the overlap between DNA sequence reads.

Depth Coverage: Specific number of aligned reads that are counted per base position

Homology searches: A given sequence of DNA or protein is searched against all of the different sequences in the database.

K-mer: Specific number of nucleotide sequences that can be used to identify fragments within DNA

Mapping-first approach: Mapping the raw reads directly into the genome for assembly.

N50: A statistical value representing the weighted median of assembled transcripts (50% of the assembled transcripts are longer than this value).

NGS - Next Generation Sequencing: DNA sequencing that operates in a massively parallel way and generates orders of magnitude more data than Sanger sequencing.

Normalization: The process of compensating for technical differences between genes/samples, to find real biological differences between them.

qRT-PCR (Quantitative reverse transcriptase PCR): A PCR technique used to determine the amount of cDNA in the sample.

Quality score of the reads: The integer mapping of p (i.e. the probability that the base is correct).

Reads: Small DNA fragments, between 20 to 1000 bases, depending on the platform, produced by DNA sequencing technology.

RMBT: Number of reads that could be mapped back to de novo assembled transcripts

RNA-Seq: Next Generation Sequencing can generate millions of reads. If the sequences come from cDNA, the generated reads represent the RNA content of the cell and are known as RNA-seq.

SAM (Sequence Alignment/MAP) format: Generic format for storing large nucleotide sequence alignments.

Transcriptome: The collection of all mRNA products of a cell. Transcriptomes will differ from cell type to cell type.

Tree of Life: A phylogenetic hypothesis about the relationships of organisms on our planet.

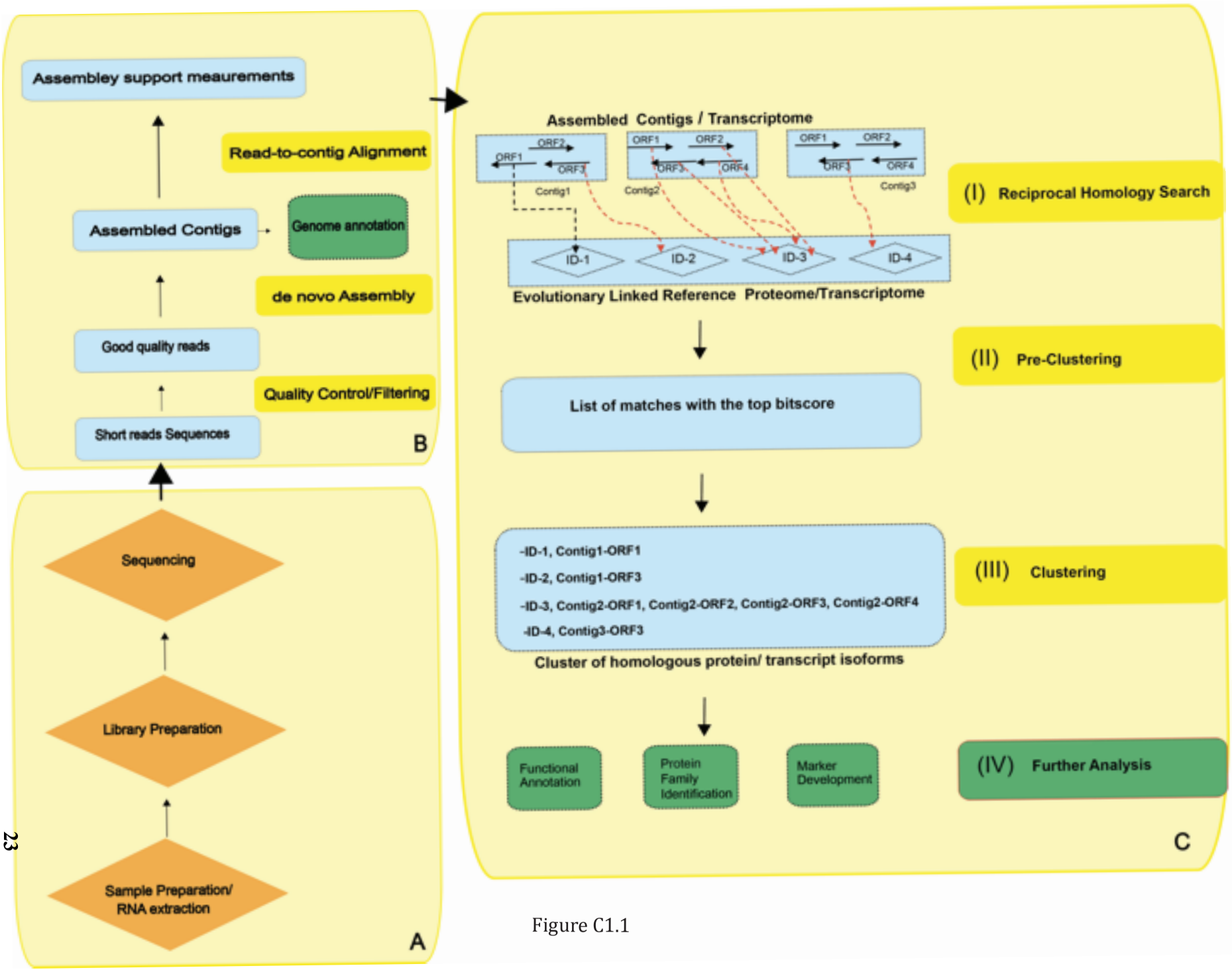
Assembley support meaurements
Read-to-contig Alignment
Assembled Contigs
Genome annotation
de novo Assembly
Good quality reads
Quality Control/Filtering
Short reads Sequences
B
Sequencing
Library Preparation
Sample Preparation/
RNA extraction
A
Assembled Contigs / Transcriptome
ORF1
ORF2
ORF3
ORF4
Contig1
Contig2
Contig3
ID-1
ID-2
ID-3
ID-4
Evolutionary Linked Reference Proteome/Transcriptome
(I) Reciprocal Homology Search
(II) Pre-Clustering
List of matches with the top bitscore
(III) Clustering
-ID-1, Contig1-ORF1
-ID-2, Contig1-ORF3
-ID-3, Contig2-ORF1, Contig2-ORF2, Contig2-ORF3, Contig2-ORF4
-ID-4, Contig3-ORF3
Cluster of homologous protein/ transcript isoforms
Functional Annotation
Protein Family Identification
Marker Development
(IV) Further Analysis
C

Figure C1.1

Figure C1.1: Work flow the genome independent RNA-seq data handling in non-model species. This pipeline consists of three steps: A. Sample preparation and sequence data generation, content in orange boxes denotes the material and procedures to generate the data; B. Read quality filtering and assembly; C. Annotation. Blue boxes denotes the input and output files produced during assembly and annotation; yellow boxes denote the software and scripts used in this workflow; Green boxes refer to the potential application and further analysis. (See text for more information)

Box C1.1. De Bruijn graphs assembly

Some transcriptome de novo assembly programs are built based on de Bruijn graphs. In a de bruijn assembly graph, the **edges** of a graph are unique small sequences of length k within reads. In this assembly graph, the **nodes** of the graph common subsequecnes of length k-1 between edges. Therefore, and edge connects two nodes if the suffix of one node shares an exact match of length k-2 with the prefix of the other connected node (*Note 1*). The de Bruijn assembly algorithms compress the graph into contigs. These contigs are non-branching subgraphs of the de Bruijn graph(*Note 2*). Contigs are extended until the extenstion branches. Repeats cause branches in the graph and makes the graph ambiguous. The role of a good assembler is to identify the correct path, representing the full picture of a transcriptome.

Note 1: Pop, Mihai. "Genome assembly reborn: recent computational challenges." Briefings in bioinformatics 10.4 (2009): 354-366.
Note 2: Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg. "Assembly of large genomes using second-generation sequencing." Genome research 20.9 (2010): 1165-1173.

Table C1.1: Comparative features of Next Generation Sequencing Platforms

Platform	Generated Sequences Characteristic	
	Read length	Depth Coverage
Illumina	2 x 100 bp	200 Gb
454/Roche	2 x 400 bp	0.6 Gb
Helicos	2 x 55 bp	25 Gb
SOLiD ABI	2 x 75 bp	300 Gb
PacBio	1000 bp	3 Gb

Table C1.2: Quality Control Programs

Program/Repository	Source	Reference
FastQC	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/	[57]
NGSQC	http://brainarray.mbni.med.umich.edu/brainarray/ngsqc/	[58]
AdapterRemoval	http://code.google.com/p/adapterremoval/	[59]
Cutadapt	https://github.com/marcelm/cutadapt	[60]
Clean-Reads	http://bioinf.comav.upv.es/clean_reads/	[61]
Seqtk	https://github.com/lh3/seqtk	[62]
Sequence Read Archive	http://www.ncbi.nlm.nih.gov/sra	
Transcriptome Shotgun Assembly	http://www.ncbi.nlm.nih.gov/subs/tsa/	

Table C1.3. Assembly Programs

Program	Source	Reference
G.Mo.R.Se	http://www.genoscope.cns.fr/externe/gmorse/	[63]
Scripture	http://www.broadinstitute.org/software/scripture/	[64]
Cufflinks	http://cufflinks.cbcb.umd.edu	[65]
Velvet	http://www.ebi.ac.uk/~zerbino/velvet/	[66]
ABySS	http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases/1.3.0	[67]
Soapdenovo2	http://sourceforge.net/projects/soapdenovo2/	[68]
Oases	http://www.ebi.ac.uk/~zerbino/oases/	[69]
transABySS	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	[70]
SOAPdenovo-Trans	http://sourceforge.net/projects/soapdenovotrans/	[71]
Rnnotator	http://www.jgi.doe.gov/software/	[76]
KISSPLICE	http://kissplice.prabi.fr/download/	[77]
Trinity	http://sourceforge.net/projects/trinityrnaseq/	[74]

Table C1.4. Annotation programs

Program	Source	Reference
NCBI BLAST	http://www.ncbi.nlm.nih.gov/guide/howto/run-blast-local `	[80]
GetORF	http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html	[81]
TribeMCL	http://freecode.com/projects/mcl-algorithm	[82]
OrthologID	http://nypg.bio.nyu.edu/orthologid/	[83]
GO	http://www.geneontology.org	[84]
SWISS-PROT	http://www.uniprot.org/downloads	[85]
Refseq	http://www.ncbi.nlm.nih.gov/Ftp/	[86]
InterPro scan	http://www.ebi.ac.uk/interpro/download.html;jsessionid=97AB129A255625E9D160F458661D877A	[87]
Blast2GO	http://www.blast2go.com/b2glaunch	[88]

Table C1.5. RNA Quantification Programs

Program	Source	Reference
BWA	http://bio-bwa.sourceforge.net	[89]
BOWTIE	http://bowtie-bio.sourceforge.net/index.shtml	[90]
BOWTIE2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml	[91]
MAQ	http://maq.sourceforge.net/maq-man.shtml	[92]
SSAHA2	http://www.sanger.ac.uk/resources/software/ssaha2/	[93]
SAMtools	http://samtools.sourceforge.net	[94]
IGV	http://www.broadinstitute.org/igv/	[95]
Bamview	http://bamview.sourceforge.net	[96]

Chapter 2: Transcriptome deep-sequencing and clustering of expressed isoforms from *Favia* corals

PUBLISHED (http://www.biomedcentral.com/1471-2164/14/546/abstract)

Shaadi F Pooyaei Mehr[1,2,*]
Email: fpooyaei-mehr@amnh.org

Rob DeSalle[2]
Email: desalle@amnh.org

Hung-Teh Kao[3]
Email: htkao2@gmail.com

Apurva Narechania[2]
Email: anarechania@amnh.org

Zhou Han[4]
Email: zhan@jbpierce.org

Dan Tchernov[5]
Email: dtchernov@univ.haifa.ac.il

Vincent Pieribone[4]
Email: vpieribo@jbpierce.org

David F Gruber[1,2,6]
Email: david.gruber@baruch.cuny.edu

[1] The Graduate Center, Molecular, Cellular and Developmental Biology, City University of New York, New York, NY 10065, USA

[2] American Museum of Natural History, Sackler Institute of Comparative Genomics, New York, NY 10024, USA

[3] Department of Psychiatry and Human Behavior, Division of Biology and Medicine, Warren Alpert Medical School, Brown University, Providence RI 02912, USA

[4] John B. Pierce Laboratory, Cellular and Molecular Physiology, Yale University, New Haven, CT 06519, USA

[5] Marine Biology Department, The Leon H. Charney School of Marine Sciences, University of Haifa, Mount Carmel, Haifa 31905, Israel

[6] Department of Natural Sciences, City University of New York, Baruch College, Box A-0506, 17 Lexington Avenue, New York, New York 10010, USA

[*] Corresponding author. American Museum of Natural History, Sackler Institute of Comparative Genomics, New York, NY 10024, USA

C2.1: Abstract

Background

Genomic and transcriptomic sequence data are essential tools for tackling ecological problems. Using an approach that combines next-generation sequencing, *de novo* transcriptome assembly, gene annotation and synthetic gene construction, we identify and cluster the protein families from *Favia* corals from the northern Red Sea.

Results

We obtained 80 million 75 bp paired-end cDNA reads from two *Favia* adult samples collected at 65 m (*Fav1*, *Fav2*) on the Illumina GA platform, and generated two *de novo* assemblies using ABySS and CAP3. After removing redundancy and filtering out low quality reads, our transcriptome datasets contained 58,268 (*Fav1*) and 62,469 (*Fav2*) contigs longer than 100 bp, with N50 values of 1,665 bp and 1,439 bp, respectively. Using the proteome of the sea anemone *Nematostella vectensis* as a reference, we were able to annotate almost 20% of each dataset using reciprocal homology searches. Homologous clustering of these annotated transcripts allowed us to divide them into 7,186 (*Fav1*) and 6,862 (*Fav2*) homologous transcript clusters (E-value $\leq 2e^{-30}$). Functional annotation categories were assigned to homologous clusters using the functional annotation of *Nematostella vectensis*. General annotation of the assembled transcripts was improved 1-3% using the *Acropora digitifera* proteome. In addition, we screened these transcript isoform clusters for fluorescent proteins (FPs) homologs and identified seven potential FP homologs in *Fav1*, and four in *Fav2*. These transcripts were validated as bona fide FP transcripts via robust fluorescence heterologous expression. Annotation of the assembled contigs revealed that 1.34% and 1.61%

(in *Fav1* and *Fav2*, respectively) of the total assembled contigs likely originated from the corals' algal symbiont, *Symbiodinium spp*.

Conclusions

Here we present a study to identify the homologous transcript isoform clusters from the transcriptome of *Favia* corals using a far-related reference proteome. Furthermore, the symbiont-derived transcripts were isolated from the datasets and their contribution quantified. This is the first annotated transcriptome of the genus *Favia*, a major increase in genomics resources available in this important family of corals.

Keywords

K-mer, Contig, Open reading frame, Fluorescent protein, Blast, Clustering, High-throughput sequencing, Illumina paired-end, Coral

C2.2: Background

With the advent of Next-Generation Sequencing (NGS) technology, genomic data acquisition has become much easier, especially for non-model organisms [1]. The generation of transcriptomes from non-model organisms has also benefitted from NGS advances. Transcriptomic datasets can facilitate genome annotation, single-nucleotide polymorphism (SNP) analysis [2], marker development for population genetic and adaptive evolutionary studies [3], as well as functional classification [4] in non-model species. The application of transcriptome deep sequencing in metabolic pathway reconstruction and gene marker development has already shown great promise in *Camellia sinesis* [5], *Cicer arietinum* [6], *Sphenodon punctatus* [7], and *Anopheles funestus* [8].

This method is also valuable for relatively understudied species, such as *Favia* corals. Though corals are high in economic and ecological value, limited genomic resources are available, largely because samples are difficult to obtain. Because NGS requires only small amounts of animal tissue, it is possible get large amounts of information from very small samples (1–2 coral polyps). Recently, anthropogenic threats such as climate change, metal pollution and oceanic acidification [9] have led to rapid declines in worldwide coral populations, lending increased urgency to the need for genomic data. Detailed understanding at the genomic and transcriptomic level will allow for the development experimental studies to assess how the intensity and frequency of disturbances affects coral health and abundance.

Several studies have reported NGS long reads transcriptome sequencing of coral species such as *Acropora millepora* [10,11] and *Pocillopora damicornis* [12]. In addition, other recent studies have used the Short Sequence Reads (SSR) platform [13], or combined SSR and long reads approach to explore whole transcriptome modulation in response to low pH in adult *Pocillopora damicornis* [13], and in early life stages of *Acropora millepora* [14]. Yet, these coral clades are quite phylogenetically divergent from *Favia* [15].

Favia is one of the most widely and uniformly distributed of all coral genera and is phenotypically presented as massive, dome-shaped and flat. In many cases *Favia* species exhibit cryptic species complexes and their phylogeny has been parodied as being a "Bigmessidae" [16]. *Favia* is within the *Faviidae* that contains twenty-four genera, more than any other coral family [17]. *Faviidae* is one of the highly fragmented families in their population structure, and Indo-Pacific members appear to be distinct from Atlantic counterparts. Therefore, adding more molecular markers to resolve their phylogeny will add further resolution to coral systematics.

We sequenced and assembled 58 Mbp of Illumina cDNA reads from two coral *Favia* samples, termed "*Fav1*" and "*Fav2,*" that were collected at 65 m in the northern Red Sea (Figure C2. 1). Reads were assembled into contigs and annotated to: 1) identify protein family clusters using the proteome of *Nematostella vectensis* as a reference; 2) assign functions to the protein family clusters using *Nematostella vectensis* GO, InterPro and KOG functional annotation*;* 3) identify homologous proteins in *Acropora digitifera* using sequence-based similarity searches; 4) identify symbiont-derived contigs in our assembly; and 5) conduct phylogenetic assessment using three molecular markers (Cytb, COI, 28S) and eleven full-length fluorescent proteins. The resulting data provide a valuable resource for future studies in faviids and other corals.

C2.3: Results and discussion

C2.3.1: *De novo* assembly

Holobiont cDNA libraries were synthesized from the RNA of two individual adult *Favia* sp. collected from the Gulf of Eilat in the Red Sea. Illumina runs performed on each separate, normalized, cDNA pool generated approximately 80 million reads per sample with average quality scores > Q20 at each base. The first step of assembly was carried out with ABySS [18,19], a *de Brujin* graph assembler. In order to recover transcripts across a range of expression levels, we carried out assembly across a range of k-mer values. Transcripts with low depth (i.e. weakly expressed) are best recovered with low k-mer values, while high depth (i.e. highly expressed) transcripts are best recovered with high k-mer values [20]. Using a range of k-mer values also allows for the identification of expressed splice variants arising from a single gene. As the Illumina read length was set to 75 bp, we chose initial k-mer values ranging from 29 to 45 bp for each sample run.

We evaluated various assembly parameters (e.g. total number of contigs, contigs longer than 100 bp, N50 length, and average contig length) as a function of k-mer length. The three k-mer values (35, 39, 45 for *Fav1* and 31, 35, 39 for *Fav2*) with the highest N50 length [21] were selected as being most informative. In each sample, we eliminated contigs shorter than 150 bp [20] in the two k-mer assemblies with the shortest median contig length, but kept all the contigs in the assembly with the longest median contig length in order to retain any information useful for bridging in the subsequent assembly steps. Within each sample, the three k-mer assemblies were then combined, and the combined contigs were assembled with CAP3 (using default parameters), which computes overlaps to correct errors in constructing contigs and generates consensus sequences for contigs [22], thus eliminating redundant contigs. It has been suggested that assembly of ABySS followed CAP3 yield better contigs [19]. As a result, the N50 length distribution improved after using CAP3, and the best N50 values increased from 1027 to 1665 in *Fav1*, and 742 to 1439 in *Fav2* (Figure C2. 2). The final assembled datasets, which were used for all subsequent analyses, contained 58,848 sequences in *Fav1* and 62,469 sequences in *Fav2*. The N50 values of these two datasets were higher than previous short-read publications [5,7,23] (675 bp, 1438 bp, 506 bp, respectively), suggesting that the quality of our data was comparable to results in other non-model species (For all commands and parameters, see Additional file 1: File S1).

C2.3.2: Homologous clustering of expressed coral transcripts

After using the EMBOSS package [24] to generate all possible open reading frames (ORFs) from stop to stop for each assembled contig, the resulting predicted ORFs were searched for sequence similarity against the *Nematostella vectensis* proteome [25], using reciprocal BlastP (E-value $\leq 2e^{-30)}$ [26] (Script 1). For the 519,766 predicted ORFs longer than 150 bp, 12,141 unique ORFs in *Fav1* showed considerable sequence similarity to 7,186 existing protein

sequences in *Nematostella vectensis*. Similarly, 12,425 unique ORFs in *Fav2* showed similarity to 6,862 *Nematostella vectensis* protein sequences (Additional file 2: File S2, Additional file 3: File S3). The top Blast hits for each sample were saved in a pre-clustering list using a Perl script (Script 2; Output files reported in Additional file 4: File S4, Additional file 5: File S5). These lists were then used in TRIBE-MCL [27] to identify homologous protein family clusters in a comprehensive and uniform way (Additional file 6: File S6, Additional file 7: File S7). The main clustering parameter, inflation value (r), was selected as default (r = 2.5). *Fav1* and *Fav2* had similar numbers (7,186 and 6,862, respectively) of protein family clusters homologous to unique *Nematostella vectensis* proteins. These clusters were subjected to further functional annotation.

In order to evaluate the completeness of our annotation using *Nematostella vectensis* as the reference as opposed to using another available Cnidarian non-annotated proteome (*Acropora digitifera)*, we applied a newly-developed completeness metric [28] (*In prep.*) to determine the proportion of the reference proteome covered by our sets of assembled transcripts. Only those ORFs with length coverage ≥80% of the matched protein from the *Nematostella vectensis or Acropora digitifera* proteome were included. Completeness measurements in *Fav1* and *Fav2* compared to *Nematostella vectensis* were 29.54% and 28.20%, respectively; when the same procedure was carried out using the unannotated proteome of *Acropora digitifera* as a reference (23,677 ORFs downloaded from http://marinegenomics.oist.jp/genomes/downloads?project_id=3. This showed an improvement of only 1-3%, thus validating our usage of *Nematostella vectensis* as a reference proteome (Additional file 8: Table C2. C2. S1).

C2.3.3: Functional annotation and characterization of the isoform clusters in *Fav1*

To identify the putative function of 7,187 isoform clusters, Gene Ontology (GO) and protein domain (KOG, InterPro) searches were performed using the functional annotation of the *Nematostella vectensis*. (Data downloaded from the JGI genome project http://genome.jgi-psf.org/Nemve1/Nemve1 .download.html). The clusters were assigned gene names based on the gene name annotation of the best Blast match for the sequences (Additional file 9: File S8). This process successfully assigned gene names for 6,632 (92.27%) clusters using GO term, KOG description, and InterPro description. Among 12,141 annotated best hits, 11,411 (93.98%) gene names were assigned to sequences. These provide a rough estimate of the number of different genes expressed in *Fav1* libraries. Broadly, the putative homologs of genes involved in various cellular processes and pathways found to be functionally conserved.

Based on GO terms assignment to clusters, a total of 4,678 (65%) clusters were assigned at least one GO term, among which 11% were assigned at least one GO term in biological processes, 48% in molecular function and 6% in cellular component category (Additional file 10: Figure C2. S1). Among the various biological processes, protein metabolism, and electron transport were mostly highly represented (Table C2. 1). Protein metabolism is also highly represented in other transcriptome characterization studies [6,7,29].

Table C2. 1 Top 30 frequent annotated functions of homologous protein clusters in *Fav1*

Top frequent GO-annotated homologous protein clusters in *Fav1*		
Go Categories/Description	Count	Percentage
Total Clusters	7,187	
Total (GO-annotated)	4,677	65.1%
Molecular function	3,477	48.37%
1-Nucleic acid binding	241	5.15%
2-Protein kinase activity	218	4.66%
3-DNA binding	208	4.45%
4-Catalytic activity	173	3.70%
5-Calcium ion binding	158	3.38%
6-ATP binding	129	2.76%
7-Protein binding	119	2.54%
8-GTP binding	114	2.44%
9-Transporter activity	96	2.05%
10-Structural constituent of ribosome	82	1.75%
Biological process	776	16.59%
1-Metabolism	122	2.61%
2-Electron transport	88	1.88%
3-Intracellular signaling cascade	54	1.15%
4-Proteolysis and peptidolysis	48	1.03%
5-Protein folding	47	1.00%
6-Protein modification	31	0.66%
7-Cell adhesion	29	0.62%
8-Intracellular protein transport	26	0.56%
9-Carbohydrate metabolism	21	0.45%
10-Regulation of cell cycle	18	0.38%
Cellular Component	424	6%
1-Ubiquitin ligase complex	68	1.45%
2-Integral to membrane	58	1.24%

3-Membrane	58	1.24%
4-Nucleus	46	0.98%
5-Intracellular	41	0.88%
6-Cytoplasm	26	0.56%
7-Cytoskeleton	24	0.51%
8-Nucleosome	16	0.34%
9-Chromatin	10	0.21%
10-Extracellular region	8	0.17%

Top 30 high frequent annotated homologous protein clusters under cellular component, molecular function and biological processes. Full annotation included in Additional file 9: File S8.

According to assigned KOG descriptions to *Fav1* clusters, a total of 6,326 (88%) clusters were assigned at least one KOG description. However, this was 4,489 (62.45%) with InerPro description assignment. This implies that the KOG description was most useful in assigning domain description to our dataset compared to InterPro. The top most frequently detected domain, associated with KOG and InterPro assignment, include conserved domain associated with predicted E3 ubiquitin ligase, fibrillins and related proteins containing Ca2 + −binding EGF-like domains, FOG: Zn-finger, GPCR Rhodopsin, and Ras GTPase superfamily. One of the utilities of domain annotation is that it provides quick access to homologs of genes with known roles in intercellular signaling pathway. The representation of genes involved in intracellular signaling pathway was very similar to that of *Acropora millepora* [10]. However, a few families showed the events of expansion (for example, Patched, Hepatocyte nuclear factor 4 and Activin-like kinase) and contraction (for example, Notch-delta, Frizzled, Wnt etc.) indicating their functional significance (Table C2. 2).

Table C2. 2 Intracellular signaling pathway genes annotated in *Fav1*

Intracellular signaling pathway proteins annotated in *Fav1*		
Pathway	Protein name	Sequences (n)
Hedgehog	Patched	27
	Sonic	2
	Fused	1
	Receptor activity (IFRD-C)	1
	DUF699	2
	Smoothened	12
JAK/STAT	STAT protein	1
NFKB/Toll	Nuclear factor NF-kappa-B	1
	Intermediate in Toll-signaling	1
	Toll-like receptor	1
NHR	Hepatocyte nuclear factor 4	2
Notch	Notch	4
	TACE	3
RTK	RTK signaling protein	1
TGF-beta	Activin-like kinase	8
	SMAD	9
	TGF-beta-receptor	1
WNT	Frizzled	9
	Wnt	2

Further, we identified major transcription factors encoding transcripts. In comparison to *Acropora millerpora* [10], the represented genes were somewhat similar. However, a few families were newly reported in our dataset (For example, HMG box, T-box, ETSDomain, MADS) (Table C2. 3).

Table C2. 3 Major transcription factor families identified by conserved domain annotation

Transcription factors identified by KOG/InterPro/GO annotation in *Fav1*	
Sequence description	**Sequences(n)**
CBF	1
Transcriptional Coactivator P50	1
Transcriptional Coactivator P100	6
Transcriptional Coactivator CAPER	2
Homeobox domain	7
HSF-type DNA-binding	1
P53 DNA-binding domain	2
NF-X1-type zinc finger protein	3
Dimerization partner (TDP)	2
Fork head	15
Basic region leucine zipper & bZIP	6
Helix-loop-helix DNA binding domain	12
Myb-like DNA-binding domain	3
Zinc finger C2H2 type	3
Zinc finger MIZ type	1
HMG box	12
TBOX	5
ETS domain	12
MADS domain	4

C2.3.4: Annotation of *Symbiodinium*-derived contigs

Holobiont coral tissues also contain eukaryotic dinoflagellate endosymbionts of the genus *Symbiodinium* [30,31]. We therefore determined the contribution of symbiont-derived transcripts in our analysis. First, we extracted the regions of cDNA contigs that corresponded to each individual annotated ORF in two datasets (For commands, see Additional file 1: File S1). Furthermore their similarity search against two *Symbiodinium* transcriptomes (http://medinalab.org/zoox/) was performed using BlastN. In order to define an E-value as a cutoff threshold, a reciprocal BlastN search between the *Nematostella vectensis* genome and the two *Symbiodinium* transcriptomes showed an average E-value of e^{-80}. Thus all contigs with similarity higher than this threshold to *Symbiodinium* were defined as likely to be symbiont-derived. Based on these results, 9% of the annotated ORFs (1.34% of the total assembled contigs) of *Fav1* were labeled as symbiont sequences, and 8.7% (1.61% of total assembled contigs) of *Fav2*. FASTA files of these non-symbiont transcripts are reported (Additional file 11: File S9, Additional file 12: File S10). Finally, we performed BlastX (E-value equal to at least e^{-30}) on the non-symbiont derived cDNA fragments against the *Nematostella vectensis* proteome to confirm correct initial annotation by BlastP. All the cDNA sequences matched to the same *Nematostella vectensis* IDs that were predicted using BlastP.

C2.3.5: Phylogenetic assessment

Molecular markers, with enough resolution signals, are essential tools for population genetic studies. Typically, combination of mitochondrial and nuclear markers are used to examine the species relationships. In order to generate a *Favia* molecular marker dataset, we downloaded *Favia* related sequences from NCBI. Similarity searches were carried out against this *Favia*

dataset. Among various molecular markers, we chose COI, Cytb and 28S. Individual sequence regions were identified and extracted from the cDNA contig files in both samples. DNA alignments for each locus were generated using ClustalW2 with default parameters [32] (Additional file 13: File S11, Additional file 14: File S12, Additional file 15: File S13). Consequently, a matrix of these three loci was generated using FASconCAT [33]. A Maximum likelihood phylogenetic analysis (RaxML) was carried out [34]. Maximum likelihood phylogenetic analysis using three loci (COI, Cytb, 28S) suggests that these *Favia* samples belong to faviids (Additional file 16: Figure C2. S2). Morphological analysis places them as *Favia albidus* [17], a species that is not yet represented in NCBI. For example, out of 18 *Favia* species that have been described morphologically, only 15 of them have molecular data in NCBI. *Favia albidus*, *Favia helianthoides*, and *Favia marshae* lack sequences of any molecular markers in NCBI. Based on geological distribution [17] and morphology, we suggest these two species belong to *F. albidus*. In fact, *F. helianthoides* has no morphological similarities with our samples, and *F. marshae* habitat has never been reported in Red Sea [17]. However, further skeletal samplings are required for final validation [35,36]. Regardless, this study increases the protein information of the faviids from 496 proteins to over 12,000 proteins in NCBI.

C2.3.6: Characterization of one exemplary homologous protein cluster

From the protein clustering results, we chose to characterize a protein family with a natural fluorescent property. One of the benefits of utilizing scleractinian corals as our model organism is that they posses genes for fluorescent proteins (FPs), a rare characteristic in most other phyla besides Cnidaria [37-40]. In *Nematostella vectensis*, six protein IDs encode for FPs [41]. A search among the homologous sequence clusters with E values of at least $2e^{-30}$ in each transcriptome led to the identification of one protein cluster group per sample

representing potential fluorescent proteins (FPs). A total of 11 new potential FPs were identified, six belonging to the *Fav1* sample and four belonging to the *Fav2* sample. One additional sequence, s23Contig9635-2 was found by increasing the E-value to $2e^{-10}$ in *Fav1*. The alignment of these sequences with *Nematostella vectensis* fluorescent protein sequences (JGI ID:205348, ID:206334), *Branchiostoma GFPa1* [42] and GFP of *Aequorea victoria* (GI:17943301) showed a considerable homology (Figure C2. 3). The conserved chromophore region is located at the residues 303 to 305 based on the top sequence. Our data shows that one of the newly identified potential fluorescent protein sequences (*Fav1* s23Contig16657-5) is 185 amino acids longer at the N-terminus (416 amino acids in total) and two of them were shown to be 49 (*Fav2* s62Contig19888-6) and 41 amino acids (*Fav2* s62Contig41210-3) longer than the consensus length of reported sequences in NCBI (wild-type GFP from *Aequorea victoria* is 236 amino acids) (Additional file 17: Figure C2. S3). This extended region does not seem to interfere with the proper folding and expression of FP, however further studies are required to reveal the function of these upstream domains.

Furthermore, the maximum likelihood trees were generated from the alignment of 156 fluorescent sequences, including the 11 newly identified sequences (Additional file 18: Figure C2. S4, Additional file 19: File S14 contains all the accession numbers). There was a strong bootstrap support for basal clade relationships within tree. This includes the order Ceriantharia, and Pennatulacea, although low bootstrap support for FPs within order Scleractinia. Ctenophore FP clustered with hydrozoan FP, therefore the cnidarian clade was not monophyletic. Others have shown that incongruence with taxonomy is not unusual in fluorescent proteins [43]. For better visualization, a smaller maximum likelihood sub-tree was generated from 46 scleractinian FP sequences (Figure C2. 4). Although the bootstrap values improved compared to Additional file 18: Figure C2. S4, some branches still exhibited low bootstrap values. Nonetheless, using RaXML [27], we categorized the newly identified

sequences into four clades and using ProtTest [44] we identified “PROTGAMMAWAGF” as the best-fit model.

In order to evaluate our assembly method and the possible impact of ABySS-specific errors on the annotation accuracy of the long candidate FP sequence, we performed both Trans-ABySS [20] and Trinity [45] on reads from *Fav1*. Both assembly programs led to the generation of sequences identical to *Fav1* s23Contig16657-5 as predicted using ABySS and CAP3. (Additional file 20: File S15).

C2.3.7: Validation of the identified protein clusters as fluorescent proteins

The intrinsic fluorescence of FPs includes a unique chromophore that is formed post-translationally within the protein upon autocatalytic cyclization and oxidation of residues X-Tyr-Gly [46]. The fluorophore is located almost at the center of the cylinder and is inaccessible to outside enzymes [46,47]. The GFP fluorophore is capable of forming under a wide range of conditions and once formed is highly stable. The entire structure is very resistant to denaturation by heat and denaturants. The three sequences with longer N-terminal domains (s23Contig16657-5, s62Contig19888-6 and s62Contig41210-3) were cloned into mammalian expression vectors. We used Kozak analysis [48] to pick the best potential start codon, and reading frames were generated using gene synthesis. The start codons are underlined in red in Additional file 21: Figure C2. S5. The synthesized sequences were optimized for expression in mammalian cell lines. The synthesized sequences showed fluorescence when expressed in HEK-293 mammalian cells, thus validating them as genuine FPs (Figure C2. 5).

C2.3.8: *In Silico* quantification of faviids transcripts

In order to rule out the possibility of promiscuous domain assembly, we assessed the quality of the *de novo* assembly of FP sequences, as well as all other transcripts, by mapping reads on assembled contigs for each sample. Such read alignment to contigs is necessary to provide support for new transcript identification as well as for determining gene expression levels [49,50]. In order to measure the Reads Per Kilobase of exon model per Million mapped reads (RPKM) [50], a sub-fasta cDNA region, corresponding to each ORF, within each contig was generated. Reads were aligned to these annotated cDNA regions. Coverage (RPKM) measurements were determined using a Perl script (Script 3). The results are reported (Additional file 22: File S16, Additional file 23: File S17). The mapping of all the reads onto the annotated faviids transcript showed that the number of reads corresponding to each transcript ranged from 10 to 47,189, with an average of 850 reads per transcript in *Fav1*, and 10 to 29,222, with an average of 766.37 reads per transcript in *Fav2*, indicating a wide range of expression level of faviids transcripts. It also indicates that very low expressed annotated faviids transcripts were also represented in our assembly. The minimum coverage (RPKM) of an annotated *Fav1* transcript was 3.89 and maximum of 6,919.20 with an average of 68.61. The RPKM ranged from 3.60 to 8,576, with an average of 72.64 in *Fav2.* The average and the range of RPKM per transcript is similar and somewhat higher (25.7) than other whole transcriptome studies [26].

All the cDNA regions annotated for fluorescent property had reasonable coverage, including the long candidate cDNA sequence (*Fav1* s23Contig16657-5) (Additional file 21: Figure C2. S5). Based on the calculated RPKMs for each of the identified fluorescent protein in both samples*,* s23Contig19691-3 in *Fav1*, and s62Contig57475-7 in *Fav2* had the highest

C2.4: Conclusions

In this study, we demonstrate a gene clustering strategy and utilize this in conjunction with NGS contig assembly, sequence conservation measurements, annotation and expression quantification for *de novo* assembled transcriptomic data. Working with two uncharacterized f*aviid* corals, we report 120,000 non-redundant transcripts to a genus whose sequence data was previously limited to 496 in public databases. These results provided greatly enhanced access to the expressed genes in Faviidae reef building corals, a potentially valuable resource of genetic/functional markers for population structure and functional genomic studies. We also took advantage of the optical properties of these corals expressed FPs to validate our annotation methods to show that these sequences were indeed *bonafide* fluorescent protein genes. These methods reported in this study are available via Open Source software programs as well as our provided scripts.

C2.5: Methods

Coral collection and total RNA isolation

This study was conducted during May–June 2009 on a coral reef on the northern tip of the Gulf of Eilat, in the northern Red Sea (29º30′N, 34°55′E).

Samples were collected at 65 m, using closed-circuit trimix rebreather system (Megalodon™). The organisms were identified under water to the family level, Faviidae, and brought to the surface in a black mesh bag to avoid sun exposure. The organisms were immediately photographed and vouchered with white light and fluorescent photography as described in [51] and stored in a shaded running-seawater facility. Within 1–2 hours of collection, samples were rinsed in sterile-filtered artificial seawater and processed for RNA and DNA. The tissue of the coral was extracted from the skeleton using QiaShredder

(Qiagen). For RNA, the TriZol method was used and stored as an ethanol precipitate for travel back to the US. DNA was extracted using Qiagen DNAeasy kit according to manufacturer's protocol and stored in at 4 °C. The specimens have been photo vouchered and their genomic and transcriptomic raw materials are stored in the American Museum of Natural History Ambrose Monell Cryo Collection.

Preparation and screening of cDNA library

Illumina sequencing using the GAII platform was performed at the Yale University W.M. Keck Biotechnology Resource Laboratory according to manufacturer's instructions (Illumina, San Diego, CA) (Additional file 24: File S18) and using high quality RNA with a 28S rRNA band at 4.5 kb that is at least twice the intensity of the 18 s rRNA band at 1.9 kb. The cDNA library contained 77,804,306, 75-mer length reads. The sequencing data are deposited in NCBI Sequencing Read Archive [52]. (The BiosampleIDs = SAMN01761696, SAMN01761695).

***De novo* assembly**

De novo assembly was carried out using ABySS with default settings across multiple k-mer values [18]. After assessing different k-mer values, the three best k-mer assemblies (35-mer, 39-mer, 45-mer for *Fav1* and 31-mer, 35-mer, 39-mer for *Fav2*) were selected and concatenated for the second step of assembly. To evaluate the N50 length and the number of assembled contigs using different k-mer values, we used a Perl script. CAP3 [22] was used to remove redundancy across ABySS assemblies and to merge contigs into longer sequences. All assembled contigs were subjected to annotation and further protein homology searches. Trans-ABySS [20] and Trinity [45] were used to confirm the long ORFs, homologous to fluorescent proteins, which were identified with ABySS and CAP3.

Gene annotation and analysis

A set of possible Open Reading Frames (ORF), stop to stop from assembled sequences, was generated using EMBOSS [24]. To annotate the *de novo* assembled sequences, a similarity search against *Nematostella vectensis* proteome was conducted using BLASTP with two E values of $2e^{-10}$ and $2e^{-30}$. The resulting data (E-value of $2e^{-30}$) was filtered and clustered using TRIBE-MCL [27]. Each homologous group was annotated using GO and KOG annotated *Nematostella vectensis* data (http://genome.jgi-psf.org/Nemve1/Nemve1.download.html). For *Symbiodinium* peptide annotation, a homology search using BLASTN with E values of $2e^{-80}$ against the *Symbiodinium* transcriptome (http://medinalab.org/zoox/) was carried out. The final non-symbiont FASTA cDNA fragments were reported.

Completeness measurement

The BlastP (E-value of $2e^{-30}$) output list generated from homology search of both samples against *Nematostella vectensis* [41] and *Acropora digitifera* [53] was organized for completeness measurements. The completeness formula according to [28] was implemented into a Perl script (*In prep*) to determine the percentage of the reference proteome that is covered by each of our sets of assembled transcripts. Length coverage of each of these reference ORFs by a hit from our data set had to be at least 80%.

Phylogenetic analysis of FPs

The maximum likelihood tree of identified fluorescent protein was generated using RaXML [34] under PROTGAMMAWAGF amino acid substitution model, selected based on the results from ProtTest [44]. The alignment was generated using MAFFT [54] and CLUSTALW2 [32] with minor adjustment at the N-terminus region, when long gaps were inconsistent with other isoforms. Bootstrap values were estimated based on 1,000 replicates

and were given for all presented branches. The variant sites were visualized with geneious (http://www.geneious.com). Dendroscope was used for visualization [55].

Phylogenetic assessment

Molecular barcodes for all the *Favia* related sequences were downloaded from NCBI. A similarity search with sequences from our annotation was carried out against this *Favia* dataset. Cytb, COI and 28S sequences were identified and extracted from the cDNA contig files in both samples. DNA alignments for each locus were generated using ClustalW2 with default parameters [32]. Consequently, a matrix of these three loci was generated using FASconCAT [33]. A Maximum likelihood phylogenetic analysis (RaxML) was carried out [34]. Bootstrap values were estimated based on 10,000 replicates and were given for all presented branches. Dendroscope was used for visualization [55].

Cloning of fluorescent proteins

Three cDNA sequences (*Fav1* s23Contig16657, *Fav2* s62Contig19888-6 and *Fav2* s62Contig41210-3) were synthesized and propagated in pUC57 (GenScript USA Inc.). Kozak [48] analysis was used to determine the location of the potential start codon. The genes were subcloned from pUC57 into the NotI-BamHI site of the mammalian expression vector pcDNA 3.1 (Invitrogen, Inc.) using standard recombinant techniques [56].

***In Silico* gene coverage measurements**

Gene coverage levels were determined using a Perl script (Script 3). This script implements Bowtie [57] to map reads to an annotated reference cDNA, and calculates the RPKM according the formula used in [50]. For visualization, BWA [58] was used to generate the read-to-contig alignment. The annotated cDNA from individual samples were used as the

reference contig, and SAMtools [59] was used to generate binary files to be visualized in the IGV [60] genome viewer (For commands, see Additional file 1: File S1).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

D.F.G., R.D. V.A.P. and S.F.P.M. designed the study. S.F.P.M. carried out the molecular genetic studies. S.F.P.M. and A.N. designed the bioinformatics pipeline and data handling scripts. R.D coordinated the statistical analysis. H-T. K., V.A.P and D.F.G. designed synthetic genes for expression in mammalian cells. Z.H., V.A.P, D.T. and D.F.G. participated in sample collection and Illumina sequencing. D.F.G., R.D. V.A.P. and S.F.P.M drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We thank Timor Katz and Tali Mass of the Interuniversity Institute of Eilat for their technical diving assistance in the collection of coral samples. Coral sample collections in this study have complied with the current laws of Israeli Natural Parks Authority; permit 2010/38008. Funding was provided by NSF grant # 0920572 and via a Baruch College Travel Grant to DFG.

Additional files description

Additional_file C2_1 as DOCX

Additional file 1: File S1 Parameters and commands used in this manuscript.

Additional_file C2_2 as TXT

Additional file 2: Files S2 BlastP parsed output files against *Nematostella vectensis* proteome for sample *Fav1* and *Fav2* with 2e-30.

Additional_file C2_3 as TXT

Additional file 3: Files S3 BlastP parsed output files against *Nematostella vectensis* proteome for sample *Fav1* and *Fav2* with 2e-30.

Additional_file C2_4 as TXT

Additional file 4: Files S4 TRIBE-MCL input files.

Additional_file C2_5 as TXT

Additional file 5: File S5 TRIBE-MCL input files.

Additional_file C2_6 as TXT

Additional file 6: Files S6 Homologous protein clusters (TRIBE-MCL) output for sample *Fav1* and *Fav2*.

Additional_file C2_7 as TXT

Additional file 7: File S7 Homologous protein clusters (TRIBE-MCL) output for sample *Fav1* and *Fav2*.

Additional_file C2_8 as TIFF

Additional file 8: Table C2. S1 Completeness metrics for two samples compared to *Nematostella ventensis* and *Acropora digitifera.*

Additional_file C2_9 as XLSX

Additional file 9: Files S8 GO,KOG, InterPro annotation for homologous protein clusters in *Fav1*.

Additional_file C2_10 as TIFF

Additional file 10: Figure C2. S1 Distribution of *Fav1* transcript clusters in different GO categories.

Additional_file C2_11 as FAS

Additional file 11: Files S9 FASTA files for cDNA region encoding for non-symbiont annotated ORFs in *Fav1* and *Fav2.*

Additional_file C2_12 as FAS

Additional file 12: File S10 FASTA files for cDNA region encoding for non-symbiont annotated ORFs in *Fav1* and *Fav2*.

Additional_file C2_13 as PHY

Additional file 13: File S11 Alignment of *Fav1* and *Fav2* Cytb nucleotide sequences, including other *Favia* species.

Additional_file C2_14 as PHY

Additional file 14: File S12 Alignment of *Fav1* and *Fav2* COI nucleotide sequences, including other *Favia* species.

Additional_file C2_15 as ALN

Additional file 15: File S13 Alignment of *Fav1* and *Fav2* 28S nucleotide sequences, including other *Favia* species.

Additional_file C2_16 as TIFF

Additional file 16: Figure C2. S2 Maximum likelihood tree of three loci (COI, Cytb, 28S). Data matrix was generated from 15 *Favia* species and *Fav1* and *Fav2*. Nucleotide sequences were aligned using clustalw2 with default parameters, the 3 loci matrix was generated using FASconCAT, and the tree was constructed using RaxML (See methods). *Montastrea cavernosa* is selected as the out-group.

Additional_file C2_17 as TIFF

Additional file 17: Figure C2. S3 Amino acid sequence alignment of full-length fluorescent protein isoforms.

Additional_file C2_18 as TIFF

Additional file 18: Figure C2. S4 Maximum likelihood tree of 156 known fluorescent

proteins, including our 11 newly identified sequences using RaxML. Shows the relationships of the major groups of known fluorescent proteins. Major lineages cluster together, although Ctenophore and Hydrozoa do not form a monophyletic group. Within Anthozoa class, order Ceriantharia (orange); Actinaria (red); Pennatulacea (dark green); and Scleractinia (black); Hydrozoa (purple); Copepoda (light green); Ctenophora (blue); Chordata (turquoise blue), most basal group; Newly identified sequences are colored blue within Scleractinia. The alignment was 1,000 times bootstrapped and *B. floridae* was the out-group.

Additional_file C2_19 as PHYLIP

Additional file 19: File S14 Alignment of 156 known fluorescent proteins, including the 11 newly identified FP sequences.

Additional_file C2_20 as TXT

Additional file 20: File S15 Search result in Trans-ABySS and Trinity assembly output for homologous contig, similar to identified *Fav1* s23Coting16657-5 produced by ABySS and CAP3.

Additional_file C2_21 as TIFF

Additional file 21: Figure C2. S5 Read-to-contig alignment. 75 bp read alignments to the coding region of s23Contig16657-5, 1,377 bp total length.

Additional_file C2_22 as TXT

Additional file 22: Files S16 RPKM measurement for all annotated cDNA regions from *Fav1* and *Fav2*.

Additional_file C2_23 as TXT

Additional file 23: File S17 RPKM measurement for all annotated cDNA regions from *Fav1* and *Fav2*.

Additional_file C2_24 as PDF

Additional file 24: File S18 Protocol for preparing samples for sequencing of mRNA.

Scripts: They are deposited to (**https://github.com/spooyaei/coralAssembly**) and can be downloaded.

Script 1: Perl script for performing blast search. Script 2: Perl script for pre-clustering the blast parsed file. Script 3: Perl script to calculate RPKM for the assembled file. Script S1: Perl script to shuffle short read sequences. Script S2: Perl script to measure the N50 statistics. Script S3: Unix shell script to remove Fasta files shorter than a threshold. Script S4: Generate the sub-Fasta file. Script S5: Extract the cDNA

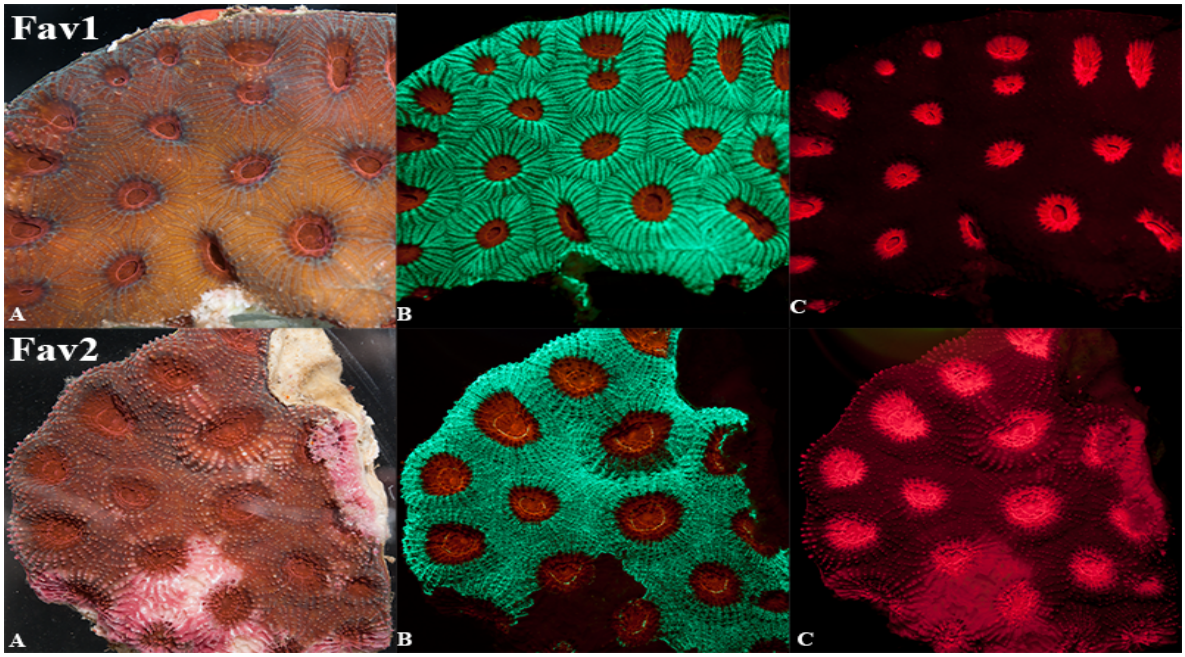

Figure C2. 1: White light and fluorescent macrophotography of scleractinian coral samples. Samples of *Favia* sp. were placed in a narrow photography tank against a thin plate glass front. Fluorescent macro images (13.1 megapixel; Nikon D300S) were produced in a dark room by covering the flash (Vivitar 185) with interference bandpass excitation filters (Semrock, Rochester, NY). Longpass and bandpass emission filters (Semrock) were attached to the front of the camera. A) White light image; B) ex. 450–500 nm; em. 514LP; C) ex. 500–550 nm, em. 555 LP

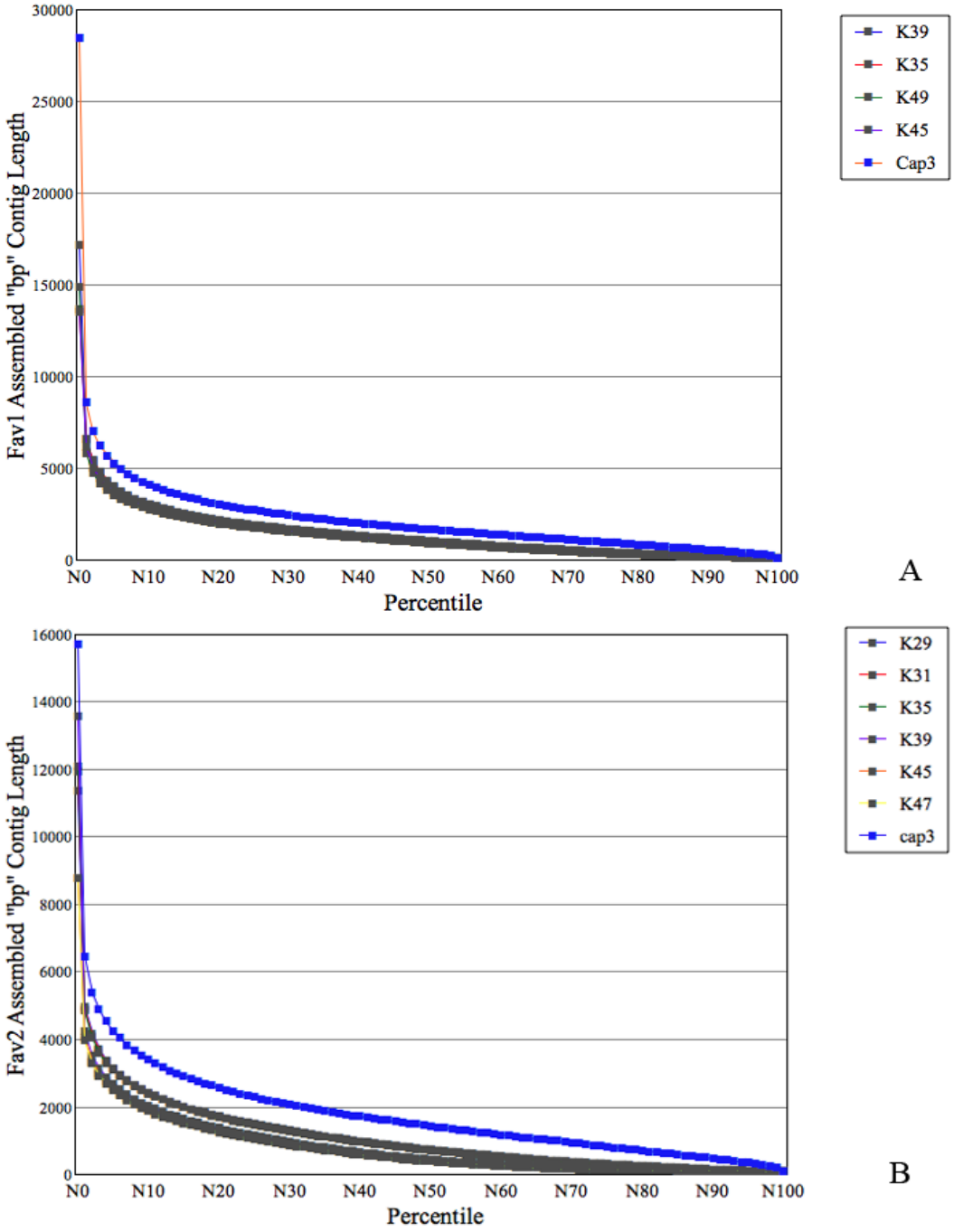

Figure C2. 2: Contig length improvement after using CAP3. N50 (50% of the length of the assembled sequences) is a parameter to assess the contig length distribution (A) Fav1 contig length and N-values relationship. The thin lines represent the values for k-mer 35, 39, 45. The N50 length values were 1027, 1009, 949 bp, respectively. The line with cross represents the N-values after using CAP3, with N50 length of 1665. (B) Fav2 contig length and N-values relationship. The N50 length values for k-mer 39, 45, 49 were 453, 408, 391 bp, respectively. The N50 length values for k-mer 29, 31, 35 were 742, 734, 721 bp, respectively. The line with cross represents the N-values after using CAP3, with the N50 length of 1439 bp.

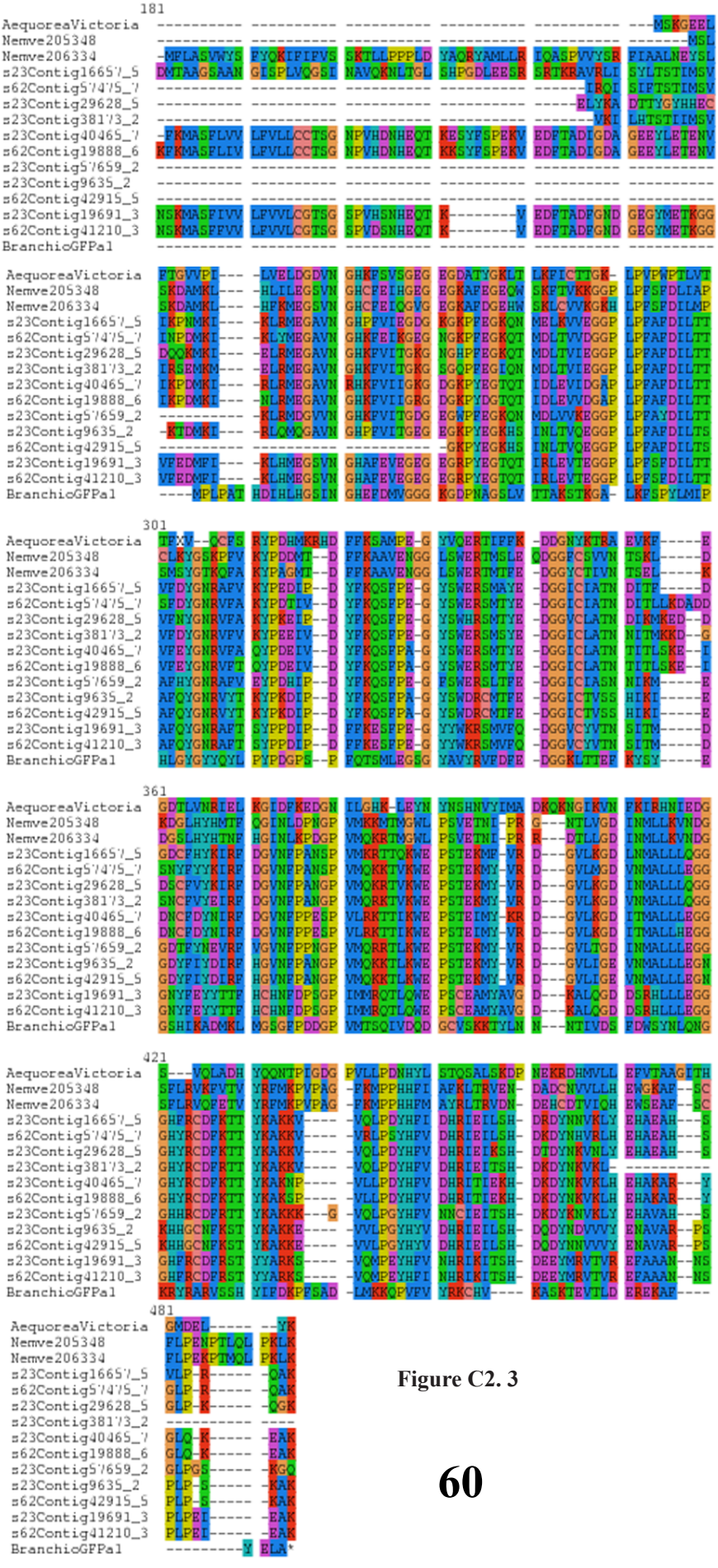

Figure C2. 3

Figure C2. 3: Overlapping region of amino acid sequence alignment of one exemplary cluster of identified homologous protein clusters. This gene family belongs to naturally expressed fluorescent protein. Conserved chromophore region (XYG) is located at the position 303–305. The newly identified sequences with extended N-terminal are s23Contig16657-5, s23Contig40465-7 in Fav1, s62Contig19888-6, and s62Contig41210-3 in Fav2. The full-length alignment is reported in Additional file 17: Figure C2. S3.

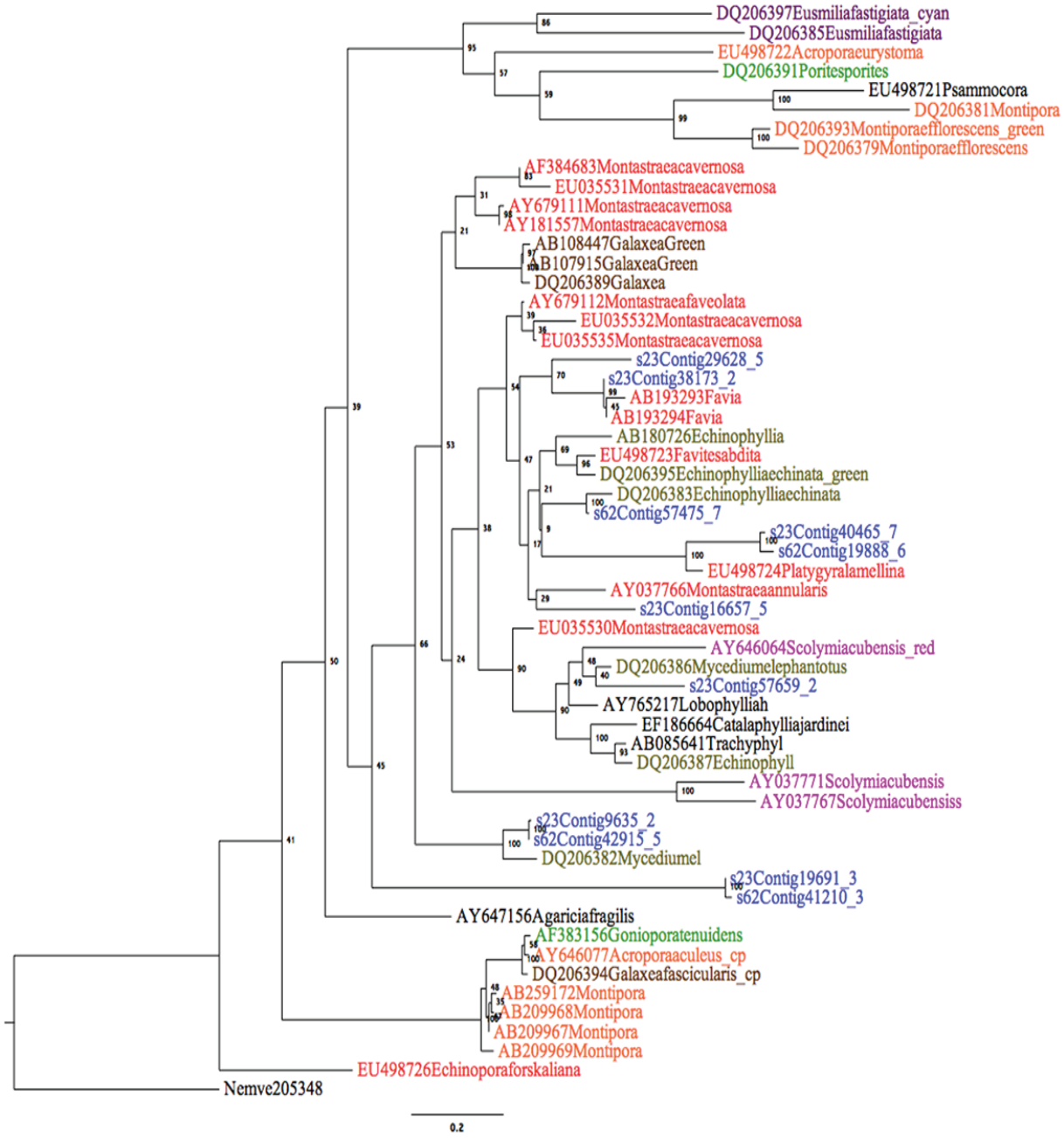

Figure C2. 4: Maximum likelihood tree of 46 known fluorescent proteins and 11 newly identified fluorescent protein sequences using RaxML. The alignment was 1,000 times bootstrapped and one FP sequence from Nematostella vectensis was the out-group. The newly identified FP sequences are colored blue. Other colors represent different coral families; Faviidae, red; Acroporidae, orange; Oculinidae, brown; Pectiniidae, dark green; Meandriniidae, dark purple; Mussidae, pink; Poritidae, green; Node labels are bootstrap supports. See Additional file 19: File S14 for information on alignment.

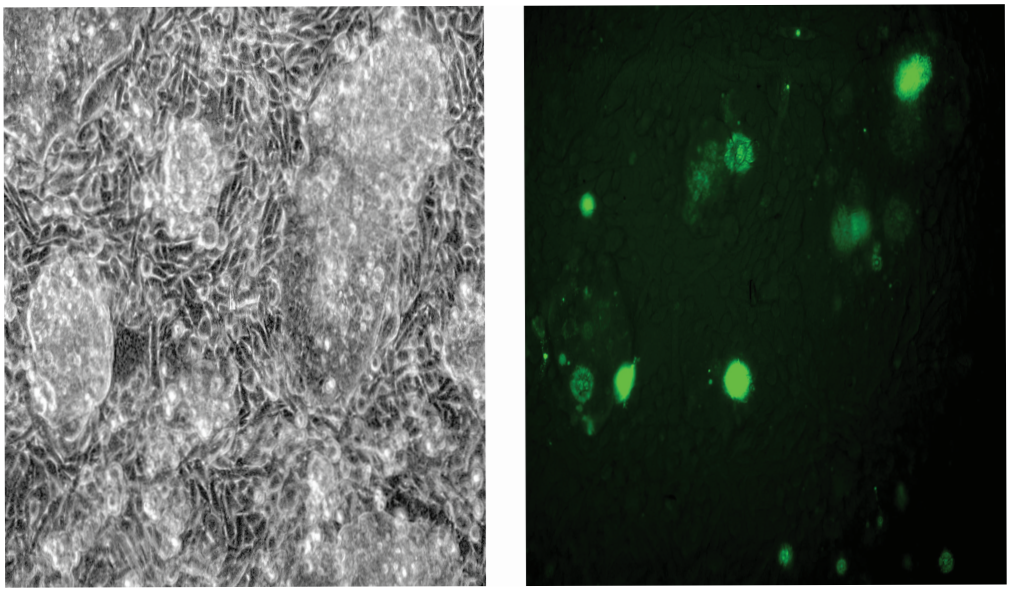

Figure C2. 5: Expression of an assembled contig in HEK293 mammalian cells yields fluorescence. An open reading frame of contig 19888 from Fav2 was synthesized using mammalian preferred codon usage (887 bases of s62Contig19888) and subcloned into pcDNA 3.1, and transfected into HEK293 mammalian cells using Fugene (Boehringer-Mannheim). The left panel depicts a phase contrast image of transfected HEK293 cells, and the right panel depicts fluorescence (using FITC excitation and emission) from the same field. Scale bar = 100 microns.

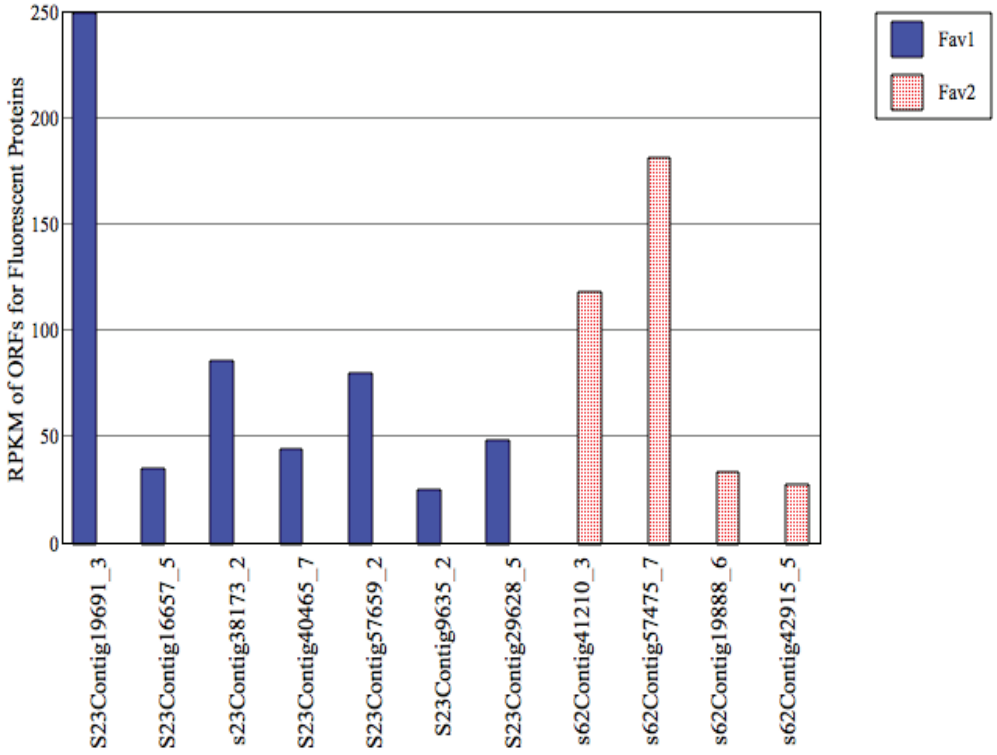

Figure C2. 6 In silico coverage plot of the read-to-contig alignment measurements. The cDNA fragments with annotation for fluorescent protein coverage measurements.

Additional files provided with this submission:

Additional file 1: 8845002068638788_add1.docx, 141K
http://www.biomedcentral.com/imedia/4933499691054589/supp1.docx
Additional file 2: 8845002068638788_add2.txt, 10482K
http://www.biomedcentral.com/imedia/1705647093105458/supp2.txt
Additional file 3: 8845002068638788_add3.txt, 8990K
http://www.biomedcentral.com/imedia/2023516563105458/supp3.txt
Additional file 4: 8845002068638788_add4.txt, 394K
http://www.biomedcentral.com/imedia/5884496921054589/supp4.txt
Additional file 5: 8845002068638788_add5.txt, 402K
http://www.biomedcentral.com/imedia/4512619361054589/supp5.txt
Additional file 6: 8845002068638788_add6.txt, 286K
http://www.biomedcentral.com/imedia/5251572751054589/supp6.txt
Additional file 7: 8845002068638788_add7.txt, 286K
http://www.biomedcentral.com/imedia/5333863201054589/supp7.txt
Additional file 8: 8845002068638788_add8.tiff, 729K
http://www.biomedcentral.com/imedia/1346552651105458/supp8.tiff
Additional file 9: 8845002068638788_add9.xlsx, 522K
http://www.biomedcentral.com/imedia/1714946650105458/supp9.xlsx
Additional file 10: 8845002068638788_add10.tiff, 286K
http://www.biomedcentral.com/imedia/1322518122105458/supp10.tiff
Additional file 11: 8845002068638788_add11.fas, 14382K
http://www.biomedcentral.com/imedia/9039435011054589/supp11.fas
Additional file 12: 8845002068638788_add12.fas, 13352K
http://www.biomedcentral.com/imedia/7686446441054591/supp12.fas
Additional file 13: 8845002068638788_add13.phy, 10K
http://www.biomedcentral.com/imedia/4870527031054591/supp13.phy
Additional file 14: 8845002068638788_add14.phy, 12K
http://www.biomedcentral.com/imedia/1519190342105459/supp14.phy
Additional file 15: 8845002068638788_add15.aln, 10K
http://www.biomedcentral.com/imedia/1305480512105459/supp15.aln
Additional file 16: 8845002068638788_add16.tiff, 3320K
http://www.biomedcentral.com/imedia/1786599653105459/supp16.tiff
Additional file 17: 8845002068638788_add17.tiff, 3256K
http://www.biomedcentral.com/imedia/1510524312105459/supp17.tiff
Additional file 18: 8845002068638788_add18.tiff, 7411K
http://www.biomedcentral.com/imedia/1314032039105459/supp18.tiff
Additional file 19: 8845002068638788_add19.phylip, 56K
http://www.biomedcentral.com/imedia/1163836444105459/supp19.phylip
Additional file 20: 8845002068638788_add20.txt, 5K
http://www.biomedcentral.com/imedia/1788209099105459/supp20.txt
Additional file 21: 8845002068638788_add21.tiff, 1517K
http://www.biomedcentral.com/imedia/1566485031105459/supp21.tiff
Additional file 22: 8845002068638788_add22.txt, 452K
http://www.biomedcentral.com/imedia/8651396731054592/supp22.txt
Additional file 23: 8845002068638788_add23.txt, 461K

http://www.biomedcentral.com/imedia/9133753291054592/supp23.txt
Additional file 24: 8845002068638788_add24.pdf, 136K
http://www.biomedcentral.com/imedia/2056354856105459/supp24.pdf

Chapter 3: An insight into *Hermodice carunculata* (Annelida, Amphinomidae) body segment transcriptome

C3.1: Abstract

The amphinomid polychaete *Hermodice carunculata* is an important omnivore in coral reef ecosystems, known to prey on a diverse suite of reef organisms. The availability of genomic data for this and other species of Amphinomidae, a group with unclear phylogenetic position within Annelida, is particularly scarce. In the specific case of *Hermodice carunculata*, only a few genetic markers are available in public data bases such as NCBI. Application of RNA sequencing provides reliable datasets for marker development and novel protein identification for species with un-sequenced genomes. We obtained 400 million 75-bp reads from Hi-seq Illumina genome analyzer. After removing the low quality reads, the remaining reads were assembled with ABySS using a range of kmer values, followed by BLAT. Assembled transcriptome contained 525,989 >= 200bp contigs, with an N50 of 1095 bp, and mean length of 722.30 bp. Of these *de novo* assembled transcript sequences, we focused on 58,454 predicted Open Reading Frames (ORFs) longer than 200 amino acids for homology search against *Capitella teleta* and *Helobdella robusta*, the most closely related species with available proteome. Of these ORFs, 23,617 (40%) showed significant sequence homology (E^{-15}) to *Capitella teleta* proteome, and 20,468(35%) to *Helobdella robusta*. For functional annotation, Gene Ontology (GO) terms and InterPro IDs were assigned to 32,500 ORFs longer than 200 amino acids (55.59%). In addition, four molecular marker transcripts encoding EF-1α,

H3, CytB, and U2 snRNA previously unavailable for *Hermodice carunculata* were annotated. Furthermore, these ORFs were searched against bioluminescent proteins available in NCBI, and consequently eight full-length sequences were found to be homologous to *Renilla reniformis* (Cnidaria) luciferase. This is the first report of sequences homologous to luminescent proteins in an annelid. Also, eight Attractin-like proteins, a species-specific sex pheromone useful for species delineation, were annotated. The data presented here potentially allows to identify new molecular markers with phylogenetic signal which can help to more accurately build the natural history and phylogeography of *Hermodice carunculata* and other amphinomid populations, to unravel the evolutionary relationships within the family and to elucidate the phylogenetic position of amphinomidae within Annelida. This data provides an 11,548.74 % increase (279 to 32,500) of available genetic data for this species in NCBI.

`

C3.2: Introduction

Fast growing interest in further exploring the complex landscape and dynamic of the transcriptome in non-model organisms has led to the application of Next Generation Sequencing (NGS) methods for generating cDNA libraries. These libraries have an enormous sequencing depth and better reproducibility, producing at least 100 to 10,000 times higher throughput than classical Sanger sequencing [1]. This level of sensitivity in examination of thousands of transcripts from species with no available genome renders it easy to perform a wide range of biological studies including phylogenomics [2],

regulatory gene discovery [3-6], molecular marker development [7], single nucleotide polymorphism (SNP), identification for trait adaptation [8, 9], haplotype detection [10, 11], and differential gene expression profiling [3, 10].

The amphinomid polychaete *Hermodice carunculata* (Annelida, Amphinomidae) is an important omnivore inhabiting coral reefs throughout the Atlantic Ocean, including the Gulf of Mexico and the Caribbean Sea, as well as the Mediterranean and Red seas [12], known to prey on a diverse suite of reef organisms such as zoanthids [13, 14], scleractinian corals [15-18], milleporid hydrocorals [16, 19], anemones [20], and gorgonians [16]. *Hermodice carunculata* is also a winter reservoir and spring-summer vector for the coral-bleaching pathogen *Vibrio shiloi* [21], hence playing an ecologically important and complex role in the health of coral reef ecosystems. Designing functional genomics experiments, such as microarray, for host pathogen interaction, requires a functionally annotated EST data from this species.

Amphinomidae is a well-delineated clade within aciculate polychaetes and it comprises approximately 200 described species from 25 genera [22-24]. Amphinomids are distributed worldwide and are known to inhabit intertidal, continental shelf and shallow reef communities, with a few species also recorded from the deep sea [24]. The clade is primarily identified by a series of morphological apomorphies including nuchal organs situated on a caruncle, a ventral muscular eversible proboscis with thickened cuticle on circular lamellae, and calcareous chaetae [23, 25]. Due to the lack of knowledge regarding their morphological variability (particularly within closely related genera),

previous studies based mainly on morphology have failed to clarify the evolutionary history of the group, leading to taxonomic problems. In fact, several nominal species have been regarded as congeners, often without evaluation of relevant molecular data. This may explain the occurrence of several species with cosmopolitan distributions within the clade [26]. Consequently, detailed revisions of species and even genera are needed [24], which incorporate molecular phylogenetic studies to clarify the affinities within the family [22]. Additionally, amphinomids have been regarded as morphologically primitive and are considered of prime interest for determining the root of the annelid Tree of Life [27]. Incrementing the available genetic data on *Hermodice carunculata* represents an invaluable resource to understand the natural history of this group of polychaetes in particular, and the evolution of annelids in general.

The nuclear genes 18S, 28S and the mitochondrial genes 16S and COI (cytochrome oxidase c subunit 1) are amongst the most popular molecular markers for phylogenetic studies and have proven to be very valuable to clarify evolutionary relationships at different levels within annelids [28], including amphinomids [12, 22, 24, 29], and many other phyla [30-32]. Other genes less commonly used in annelid phylogenetic studies such as the nuclear genes EF-1α (elongation factor 1α) [33], H3 (histone H3) [34] and U2 snRNA (U2 spliceosomal RNA) [34] as well as the mitochondrial CytB (cytochrome oxidase b) [35], have been also proven to be useful to recover evolutionary relationships among annelids [34]. However, the representation of the aforementioned markers in NCBI for amphinomids is very limited. Previous to this study, the only available sequences were EF-1α for *Paramphinome jeffreysii* and *Chloeia pinnata* and H3 and U2 snRNA for *Eurythoe* sp.

Additionally, strong geographical population structure is common among polychaetes and many complexes of cryptic species have been documented [36-38]. On the other hand, studies that confirm the large geographic range of putative cosmopolitan species or reveal weak population structure are scarce [39, 40]. Among amphinomid polychaetes in particular, there are studies that suggest both the presence of cryptic speciation and significant population structure [29] as well as genetic homogeneity among populations of cosmopolitan species [12]. Studies that take advantage of newly developed methodologies and recent advances in molecular techniques provide great resources to further investigate species boundaries and conflictive evolutionary biology questions. For example, a recent study in Hormogastridae (Annelida, Oligochaeta) transcriptomics reported the precursor protein sequence of the sex pheromone Attractin as a reliable phylogenetic marker recovering deep metazoan relationships as well as putatively differentiate between closely related species of morphologically indistinguishable earthworms [41].

There are several annelid lineages with bioluminescent and fluorescent species. Based on the phylogenetic distribution of bioluminescent property, it is estimated that it has evolved independently as many as 30 times [42]. Bioluminescence properties in some terrestrial oligochaetes [43], and several species of marine fireworms have been thoroughly studied [44, 45]. The benthic species *Odontosyllis enopla* (Annelida, Syllidae) for example, uses bioluminescence mainly for reproduction. Females produce a luminescent secretion that attracts males, which in turn gives off short flashes of light while swimming towards the females. In addition to reproduction purposes, like most

other luminescent organisms, annelids produce light in response to physical disturbances. *Hermodice carunculata* displays a fluorescent pattern that consists of orange bands across the dorsum, and a bright green or blue outline [46]. However, no fluorescent or bioluminescent sequences have been isolated for this or any other annelid species to date. Consequently, the genetic and biochemical basis of light production as well as the evolution of this feature with the phylum are poorly understood.

In this study we used Illumina Hi-seq platform to generate a cDNA library of *Hermodice carunculata* muscle tissue. Following a *de novo* assembly method, 400 million 75-bp reads were assembled into 525,989 >= 200bp contigs. Of these we focused on 58,828 Open Reading Frames (ORFs) longer than 200 amino acids for sequence homology search against non-redundant protein databank of *Capitella teleta* (Annelida, Capitellidae) and *Helobdella robusta* (Annelida, Glossiphoniidae), the most closely related species with available proteome and functional annotations. Our primary goal was to produce a reference set of mRNA sequences for *Hermodice carunculata,* which will facilitate annotation of the genome and future studies of polychaete phylogeny, systematics and functional genomics.

C3.3: Results

C3.3.1: Sequencing and *de novo* assembly

To cover the *Hermodice carunculata* body segment transcriptome, total RNA was extracted from three body-segment tissues. The $(A)^+$ RNA was isolated, sheered to

smaller fragments, reverse transcribed to make cDNA for sequencing with Hi-seq Illumina 2000. Four hundred million paired-end reads were obtained from one lane of one plate, generating 97 gigabase pairs (Gbp) of raw data. Reads were checked for Phred-like quality scores above the Q30 level with FastQC [47]. We used the proposed pipeline proposed in [48] for *de novo* assembly. Hi-seq Illumina read sequences were assembled into 525,989 contigs longer than 200bp, with an N50 of 1,095 and mean length of 722.30 bp, using ABySS [49] followed by Blat (with default parameters) [50] for redundancy removal. A range of 8 kmers (21-55) were used for ABySS runs, with the parameter q = 3 to trim low-quality bases from the ends of reads for each run. The final data set was filtered for contigs longer than 200 bp. Summary statistics for each kmer assembly, as well as for the merged and redundant-removed set of contigs is outlined in Table C3.1. The assembled transcripts are available in the supplementary material (File C3. S1). Paired-end reads and assembled contigs that do not contain ambiguous bases have been deposited into NCBI and can be downloaded through the NCBI Sequence Read Archive (SRA) and Transcriptome Shotgun assembly (TSA) website.

Assemblies at higher kmers (e.g. 41-55) had lower mean length and N50, than assemblies at lower kmers (21-35) (Table C3.1). This is in agreement with other summary statistics of NGS reported *de novo* assembly data [51]. The lower N50 and mean in final merged dataset compared with kmer 51 and kmer 55 is due to addition of shorter sequences from lower kmer assemblies. As outlined in Table C3.1, the N50 has changed from 584 in kmer 21 to 1095 bp in the merged set of contigs, indicating an improvement in the assembly contig length. Although the majority of the contig length is between 200-600

bp, we obtained 20,828 contigs, with length greater than 3,563 bp (Figure C3. 2). This result indicates that the data has a very high quality for further annotation. Lastly, the assembled sequences were deposited in Transcriptome Shotgun Assembly (http://www.ncbi.nlm.nih.gov/subs/tsa/) at the NCBI.

A six frame translation (ORFs) from stop to stop for each assembled contig was generated, using EMBOSS package [52]. This file contained 58,454 predicted ORFs longer than 200 AA, with the N50 of 490 AA, and mean length of 443.92 AA.

C3.3.2: Comparative sequence similarity with other annelids

For comparative annotation, all ORFs longer than 200 AA (58,454) were searched against two existing annelid proteome datasets *Capitella teleta* (http://genome.jgi-psf.org/Capca1/Capca1.home.html), and *Helobdella robusta* (http://genome.jgi-psf.org/Helro1/Helro1.home.html) using BlastP [53] with a significant E-value of $2e^{-15}$. Similarity search showed that 23,617 (40.5%) ORFs have similarity higher than 70% against *Capitella teleta,* while 20,468 (35%) ORFs have similarity higher than 70% against *Helobdella robusta* (Figure C3.3). This indicates that the proportion of sequences with matches in the proteome of *Capitella teleta* is greater than the proportion of matches for *Helobdella robusta*. This is expected, as *Capitella teleta* and *Hermodice carunculata* are both polychaete annelids, as oppose to *Helobdella robusta*, a leech (Clitellata). In total, 15,841 transcripts had a significant hit (70% length homology) in both datasets. These sequences are annelid-restricted transcripts, which are shared between polychaetes and clitellates. These

conserved sequences can be used for future genome annotation of any annelid species (File C3.S2).

C3.3.3: Functional annotation and characterization

One of the important aspects of mining the transcriptomic data is assigning function to individual transcript sequences. Functional annotation is an effective way to categorize genes into functional classes. This is useful for understanding the physiological meaning of large amounts of transcripts and evaluating functional differences between subgroups of sequences. These data provides a tool for designing custom microarray experiments related to annotated functions [54]. The gene ontology (GO, http://www.geneontology.org) [55] [56] is an extensive scheme for this purpose. This framework covers wide biological scopes, and with its directed acyclic graph (DAG) structure, it accounts for biological dependencies. In addition, programs such as InterProScan [57] [58] provide an integrated platform for domain-based searches against databases such as PROSITE [59], PRINTS [60], Pfam [61], and SMART [62], in addition to others. Over the past few years, resources have been developed for automatic GO term and InterPro ID assignment to unknown sequences. Here we used Blast2GO [63] for functional annotation, visualization and its associated statistics.

As part of the Blast2GO pipeline, ORFs longer than 200 AA (58,454) were subjected to sequence homology search against the non-redundant protein database (NR) at NCBI, using BlastP (E 10-10, cutoff =55, GO weight=5, HSP coverage=0). Followed by

mapping to collect GO terms, and assigning reliable information to each query sequence. Default values of Blast2GO annotation parameters were chosen to optimize the ratio between annotation accuracy and coverage [64]. This provided a framework for categorizing genes into functional annotation groups, namely biological process (sets of molecular events or operations with a defined beginning and end), molecular function (the primary activities of gene product at the molecular level, such as catalysis or binding), and cellular compartment. Furthermore, InterPro IDs (protein domain IDs) were assigned to sequences by running InterProScan (part of the Blast2Go pipeline).

Out of 58,454 predicted ORFs, excluding 8 sequences with 'unknown', 1,403 with 'NA', 39 with 'unnamed', 4,164 'hypothetical', and 31 with 'putative' terms assigned to them, 55.59 % (32,500) of the data contained definitive functional annotation. Then, sequences were classified into three categories (GOslim): biological process, cellular component and molecular function. Summary of classification of annotation is reported at Level 2. In the molecular function, the clusters relating to "binding" and "catalytic activity" were enriched (21,089 and 12,443, respectively) (Figures C3.4a; C3.4b). In the biological process classification, "metabolic process" with 14,272 sequences, "cellular processes" with 14,254 sequences, "biological regulation" with 8,818 sequences were large compared to "regulation of anatomical structure size" and "cell growth" with about 200 sequences (Figures C3.4c; C3.4d). This is expected, as these data are not collected from developmental stage with high rate of divisions. In cellular component category, the cluster size of "cell" with 20,053 sequences and "organelle" with 11,413 sequences were highly represented compared to "microbody" or "extracellular matrix" with less than 100

sequences (Figure C3.4e; C3.4f). This pattern is very similar to recent analysis of *Lymnae stagnalis* (pond snail) transcriptome functional annotation [4].

In terms of length distribution of annotated sequences, 70% to 90% of the sequences with length ranging from 200 AA to 1,500 AA were functionally annotated, while 100% of the sequences with length between 1,500 AA to 3,500 AA had a function assigned to them (Figure C3.5). This result indicates that longer sequences have higher rate of annotation than shorter sequences. A Fasta file containing all the annotated sequences, and a table representing sequence IDs with their assigned GO terms and InterPro IDs and enzyme codes are reported (Files C3.S3; C3.S4).

C3.3.4: Identification of candidate genes and potential phylogenetic markers

Using reciprocal BLAST searches between the *Hermodice carunculata* transcriptome and publicly available sequences, we have identified putative *Hermodice carunculata* homologues of genes that have been previously used as phylogenetic markers in Annelida but were unavailable for *Hermodice carunculata* and amphinomids in general, with a few exceptions. We identified 900 homologous sequences of EF-1α, 101 homologous to H3, 7 homologous to CytB, and 400 homologous to U2 snRNA. We chose the longest sequence in each category for downstream phylogenetic analysis. The alignment of each of these sequences, along with the five best hits retrieved by BLAST from the NCBI database, are available in the supplementary materials (Files C3.S5; C3.S6; C3.S7; C3.S8). Sequences were deposited in GeneBank.

Additionally, as part of our annotation pipeline, seven homologous sequences to the sex pheromone Attractin have been identified in the transcriptome of *Hermodice carunculata*. A phylogenetic analysis was performed to evaluate the potential of the *Hermodice carunculata* Attractin protein as a reliable phylogenetic marker for polychaete systematics and evolutionary studies. Our analysis corroborates results by previous authors [41] suggesting that Attractin represents an effective phylogenetic marker, recovering deep metazoan relationships (Figure C3.6; File C3.S9) and important clades such as Bilateria, its split into Deuterostomia and Protostomia, and the subdivision of the latter in Ecdysozoa and Spiralia (Lohpotrochozoa). It also recovers Annelida as a monophyletic group, finding *Hermodice carunculata* as a basal taxon within annelids (Figure C3.6).

Furthermore, a search for sequence homology in the transcriptome of *Hermodice carunculata* against 182 known bioluminescence related proteins including Obelin, Aequorin, and other luciferases, found eight sequence transcripts showing an average of 44.9% homology to the luciferase protein of the phylogenetically distant sea pansy *Renilla reniformis* (Cnidaria, Renillidae). An alignment of the *Hermodice carunculata* putative luciferase with *Renilla* luciferase is generated (Figure C3.7). The corresponding cDNA sequences are stored in a Fasta file (File C3.S10).

C3.3.5: *In silico* quantification of *Hermodice carunculata* transcript

In order to identify poor quality and potentially misassembled transcripts, reads were

mapped back onto the non-redundant set of transcripts. The number of reads corresponding to each transcript ranged from 2 to 9000 with an average of 1,644 reads, indicating a wide range of expression (File C3.S11). This indicates that very low expressed transcripts were represented in our dataset. Furthermore, we analyzed the coverage of the functionally annotated transcripts. The minimum coverage was 2 FPKM and maximum was 20,000 FPKM. Among these, 400 transcripts had a mean coverage less than 3, or gaps were removed from dataset (Table C3.2).

C3.4: Discussion and conclusion

Several so-called cosmopolitan species within amphinomids have proven to consist of various cryptic species [12]. *Hermodice carunculata* has a widespread distribution and has been reported throughout the Atlantic Ocean, Mediterranean and Red seas [29, 65]. Despite its widespread distribution, there have been only a handful of studies related to *Hermodice carunculata* and prior to this study, its representation in NCBI consisted of only 279 nucleotide sequences. In a recent species delineation study, two mitochondrial genes (COI and 16S rDNA) and the internal transcribed spacer 1 (ITS1) were used to find evidence for cryptic speciation in *Hermodice carunculata* [12]. This analysis showed that the genetic divergence is low among samples across the Atlantic and these particular three genes do not reflect any genetic basis for the observed morphological differences (variable filament abundance) among populations. Therefore, identification of adaptive loci for phylogeographic application is necessary. However, a different study using COI molecular data has found that *Eurythoe complanata* represents a complex of three

genetically different and morphologically indistinguishable lineages inhabiting the Atlantic and Pacific oceans. In the same line, the deep sea genus *Archinome* consists of four genetically distinct lineages with no apparent morphological differences. Therefore, the *de novo* assembled transcriptome presented herein for *Hermodice carunculata*, can also be used to develop additional molecular phylogenetic markers to aid forthcoming studies of species boundaries and evolutionary relationships within Amphinomidae. Furthermore, amphinomids are a morphologically primitive basal group of annelids, considered as a highly important taxon for defining the root of the annelid tree [27]. Thus, the vast amount of molecular data provided herein can also help to elucidate the basal relationships of Annelida.

Additionally, sex pheromones have been postulated to drive cryptic speciation in oligochaetes [41]. Within polychaetes, there are several species known to use pheromones to attract the opposite sex and to control the release of the gametes such as the scale worm *Harmothoe imbricata* [66], the rag worms *Nereis succinea* and *Platynereis dumerilii* [67], and the lugworm *Arenicola marina* [67].
Fluorescent and bioluminescent proteins have been reported coupled together in *Aequorea macrodactyla [68]*. Furthermore, since fluorescence was observed in our sample (Figure C3.1), the presence of bioluminescent photoproteins can be hypothesized. While within annelid polychaetes there are a number of bioluminescent distributed in various families such as Acrocirridae (*Swima*), Chaetopteridae (*Chaetopterus*), Flabelligeridae (*Poeobius*, *Flota*), Polynoidae (*Harmothoe*, *Polynoe*), Syllidae (*Odontosyllis*, *Eusyllis*, *Pionosyllis*), Terebellidae (*Polycirrus*, *Thelepus*) and

Tomopteridae (*Tomopteris*), so far no bioluminescent protein sequence has been reported from this phylum (Shimomura, 2006). Thus, this is the first report of a homologous sequence of a bioluminescent protein in Annelida. The fact that the putative *Hermodice carunculata* luciferase shows highest homology to the luciferase of a phylogenetically distant cnidarian (*Renilla reniformis*) can probably be attributable to the lack of publicly available luciferase sequences from more closely related organisms. The transcriptomic dataset presented herein can greatly help identify and characterize this putative photoprotein and facilitate future studies about the genetic and biochemical basis of light production in annelids.

An additional recent approach in estimating more accurate intraspecific and intrageneric level relationships is using conserved blocks of homologous sequences shared between genomic regions of multiple species [69]. Our data provides a complementary resource for this kind of application in the future. Also, the annotation of the genomes is reliant on transcriptome data for the exon intron boundary delimitation. Our data provides a great database for future annotation of the genome in this species. These data provide a valuable base for future ecological research in *Hermodice carunculata.*

C3.5: Materials and Methods

Sample collection

Scientific divers (D. Gruber, J. Sparks, and M. Lombardi) explored the cavern zone of Norman's Pond Cay Cave, Norman's Pond Cay, Exumas, Bahamas (GPS N 23 47.181,

W 076 08.428). The cavern zone can be explored to approximately 50 meters linear penetration from the entrance, and to a depth of 40 m, where the passage narrows and ambient light is lost.

Divers explored the walls of the cavern zone using compact LED lights for cryptic invertebrate specimens. Two polychaete specimens were collected for identification and further analysis.

RNA Extraction and Transcriptome Sequencing

Specimen of *Hermodice carunculata* was frozen in liquid nitrogen. Total RNA was extracted from dissected tail muscles of one specimen. The muscle tissue was homogenized in TriZol reagent (Life Technologies, NY) and the total RNA was precipitated with isopropanol and dissolved in ddH_2O. The quality of RNA was assessed on a 2100 Bioanalyzer and with agarose gel electrophoresis. The total RNA was pooled for Library preparation. Libraries were prepared using Hi-seq RNA sample preparation kit (Illumina Inc, San Diego, CA) according to manufacturer's instructions. One lane was multiplexed for four samples and was sequenced as 80-bp PE reads. FASTQ file generation was performed by CASAVA version 1.8.2 (llumina).

***De novo* assembly**

All the assemblies were performed on a server with 50 cores and 250 GB random access memory. Obtained reads were *de novo* assembled, using ABySS [49] followed by Blat [50], according to the proposed pipeline for merge and redundancy removal [48] in conitgs generated by ABySS. In order to recover high and low expressed transcripts, a range of kmers (21- 55) was used prior to merge with Blat.

Phylogenetic analysis

Sequences for the sex pheromone Attractin were downloaded from GenBank (accession number generation is in progress) and aligned with the *Hermodice carunculata* translated sequence using MUSCLE in SEAVIEW 4.3.0 [70]. A phylogenetic analysis using amino acid sequences was conducted with RAxML ver. 7.7.1 [71] using maximum likelihood optimality criterion with a JTT amino acid substitution model. Support values were estimated using rapid bootstrap algorithm with 1,000 replicates. The protozoan symbiont *Capsaspora owczarzaki* was specified as outgroup.

Acknowledgement

Fieldwork made possible by National Geographic Society/Waitt Grants Program Award W140-10 to ML and W101-10 to DG and funds from Ocean Opportunity Inc. Project hosted at the John H Perry Caribbean Research Center. We would like to thank Ana Riesgo for her helpful comments.

Table C3.1 Summary statistics for individual and merged assemblies

Assembly	Number of trancripts >200 bp	N50 bp	Mean length bp	Max length bp	Total number of bases
Kmer 21	143,194	584	505.54	7,342	72,390,913
Kmer 25	160,583	771	605.87	13,382	97,292,569
Kmer 29	188,890	631	523.05	8,878	98,798,757
Kmer 35	225,756	689	551.61	11,724	124,529,844
Kmer 41	179,143	891	633.86	18,825	113,552,250
Kmer 45	171,154	983	667.66	24,711	114,273,429
Kmer 51	156,387	1,096	713.03	17,800	111,509,378
Kmer 55	144,565	1,160	740.32	14,922	107,023,822
Final	525,989	1,095	722.30	24,711	379,922,870
ORFs >200AA	58,454(ORFs)	490(AA)	443.92(AA)	8,167(AA)	25,948,636(AA)

For each kmer, data from ABySS assembly are shown. The "Final" assembly is the result of merging the ABySS kmer assemblies and using BLAT to remove the redundancies. Predicted ORFs longer than 200AA from this final contig set were subjected to annotation. Kmer= required length of overlap match between two reads in ABySS; N50= length-weigthed median contig length; bp= base pair; ORF= Open Reading Frame

Table C3. 2 Summary statistics of read counts and coverage

Total number of reads	426,555,924
Number of read used reads for assembly	141,684,860 (33.22%)
Number of unused reads	28,4871,064 (66.78%)
Number of non-redundant transcripts (>200 bp)	525,989
Number of non-redundant trasncripts with back-aligned reads (>200 bp)	525,939
Number of transcripts with coverage fpkm >1 (Filterd data 1)	176,412
Number of transcripts with coverage fpkm >5 (Filtered data 2)	49,690
Average coverage for contigs from filtered dataset 2 (fpmk)	15.279
Average number of reads mapped per contig for filterd data 2	1644

bp = base pair; fpkm = paired-reads per kilo base per million ; contig= contiguous, overlapping sequence read resulting from the assembly

A

B

Figure C3.1: White light and fluorescent macrophotography of *Hermodice caranculata*. It was placed in a narrow photography tank against a thin plate glass front. Fluorescent macro images (13.1 megapixel; Nikon D300S) were produced in a dark room by covering the flash (Vivitar 185) with interference bandpass excitation filters (Semrock, Rochester, NY). Longpass and band-pass emission filters (Semrock) were attached to the front of the camera. A) White light image; B) ex. 450–500nm; em. 514LP

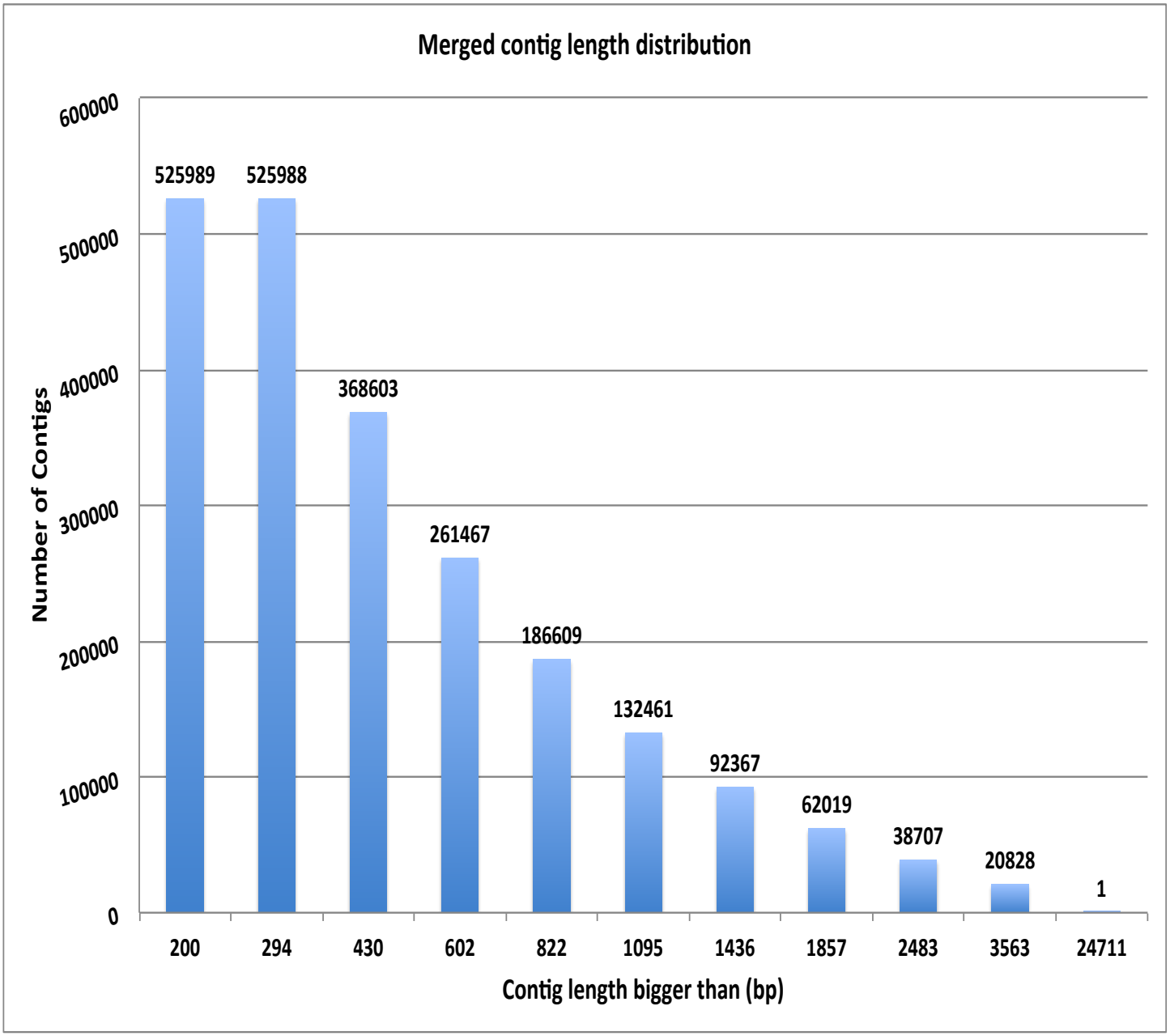

Figure C3.2: Assembled contig length distribution. Each number on top of each bar represents number of assembled contigs per length category.

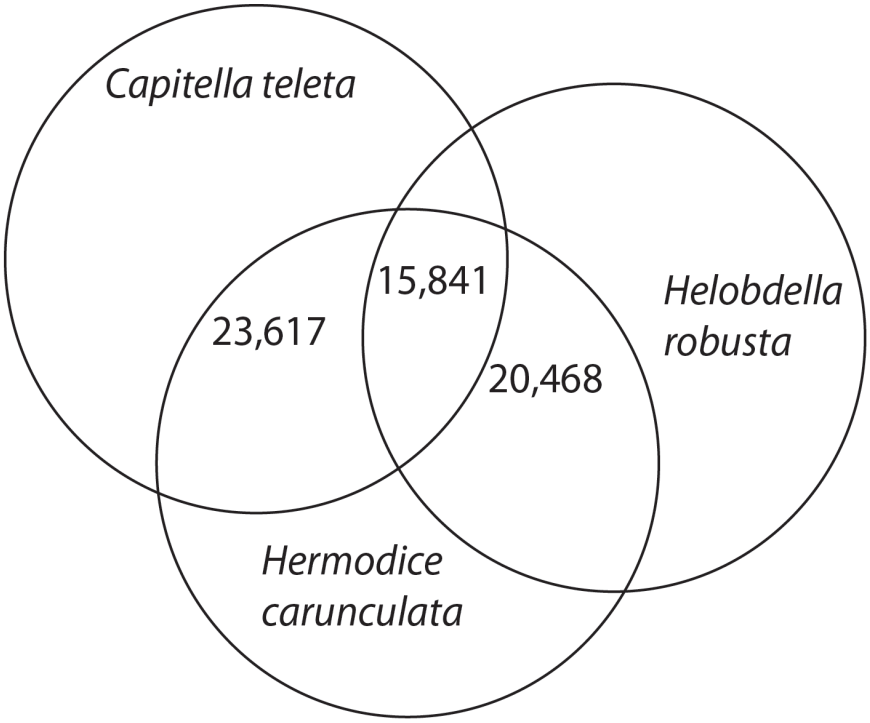

Figure C3.3: Venn diagram distribution of similarity search results. The number of unique sequence-based annotation is the best sum of unique best blastP hits from *Capitella teleta*, and *Helobdella robusta* proteome, respectively.

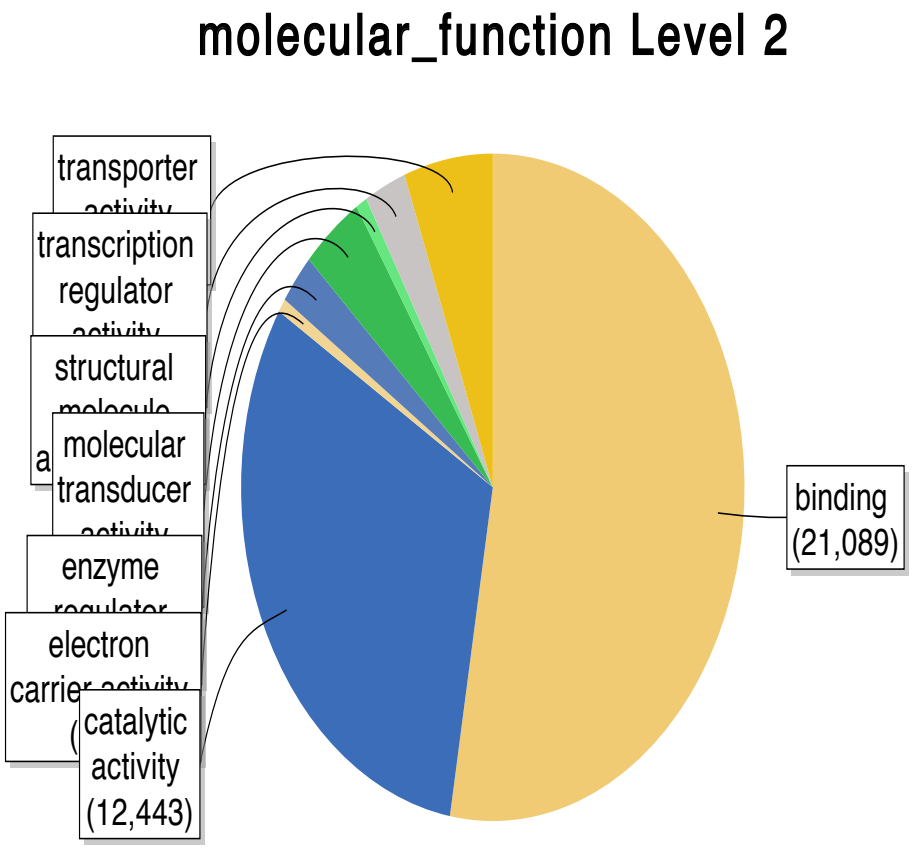

Figure C3.4a: Functional annotation of *Hermodice carunculata* transcripts. GOslim term asseingment under molecular function category.

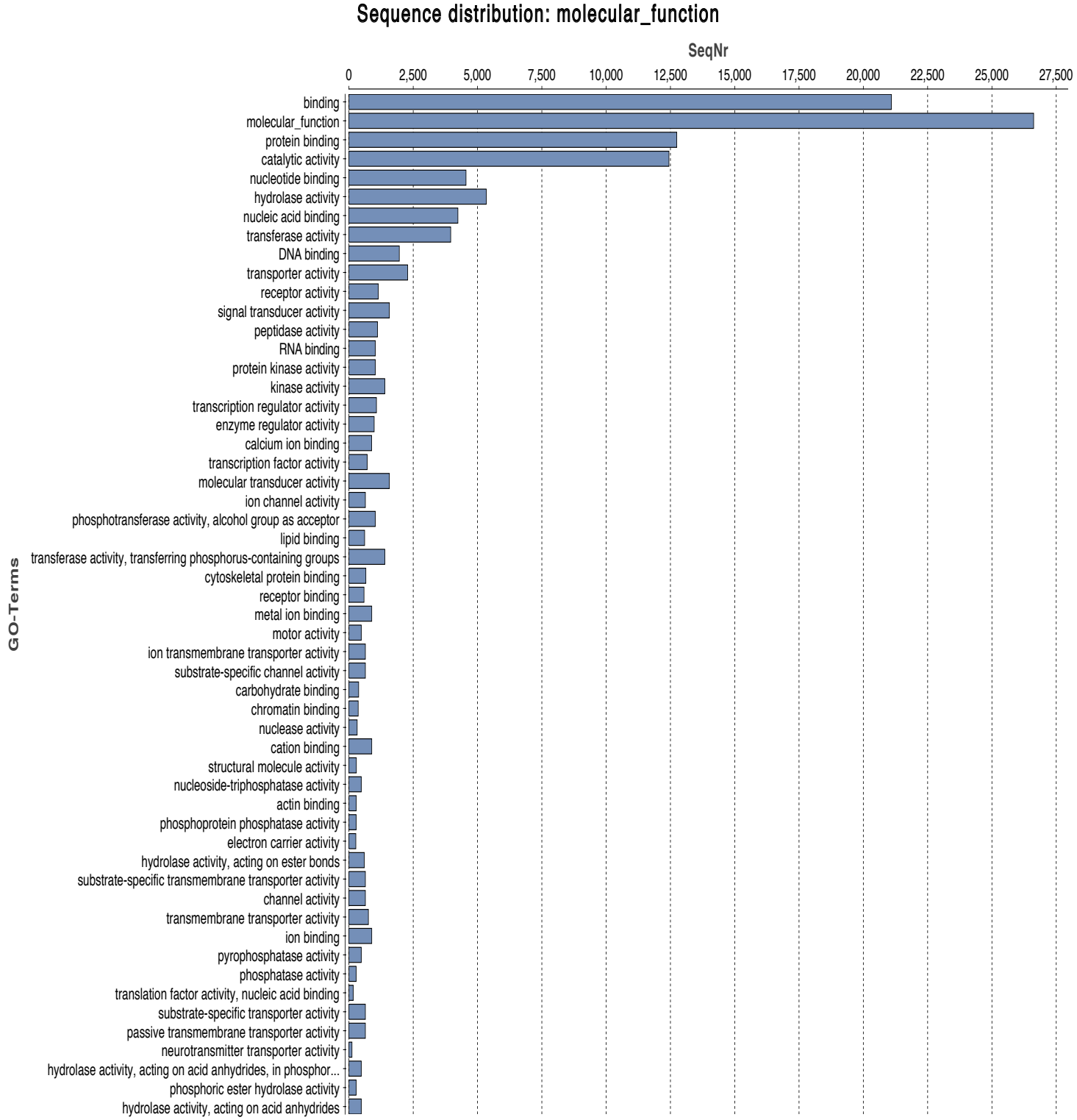

Figure C3.4b: Functional annotation of *Hermodice carunculata* transcripts. 55 most frequent GOslim term asseingment under molecular function category.

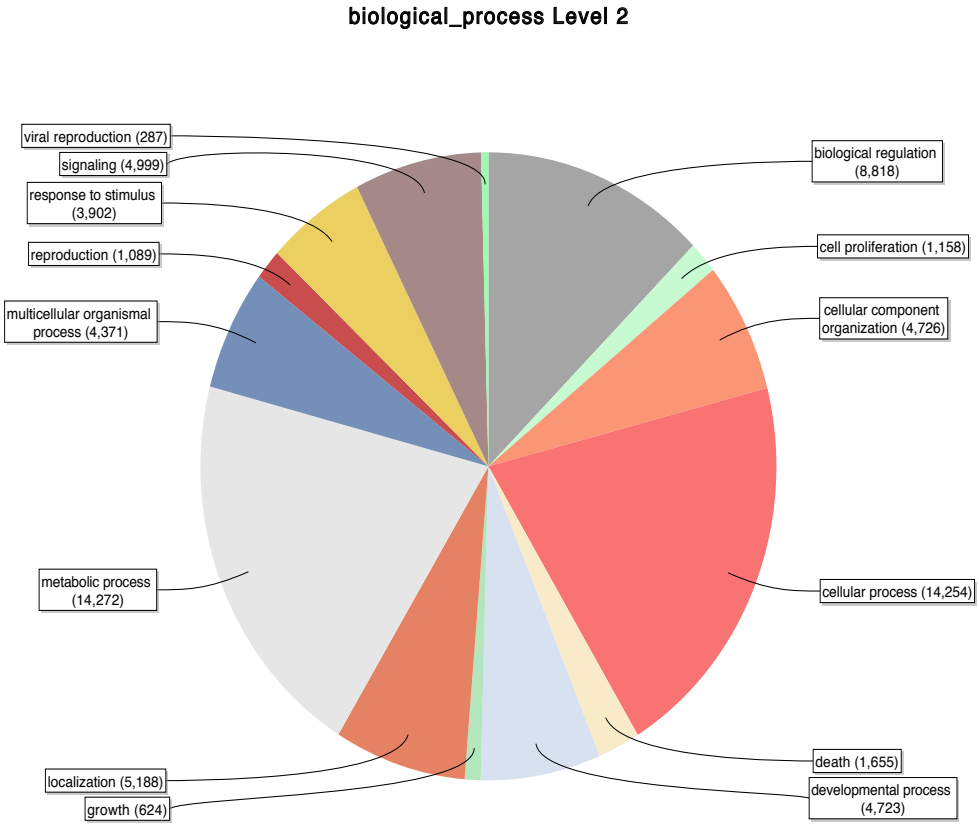

Figure C3.4c: Functional annotation of *Hermodice carunculata* transcripts. GOslim term asseingment under biological process category.

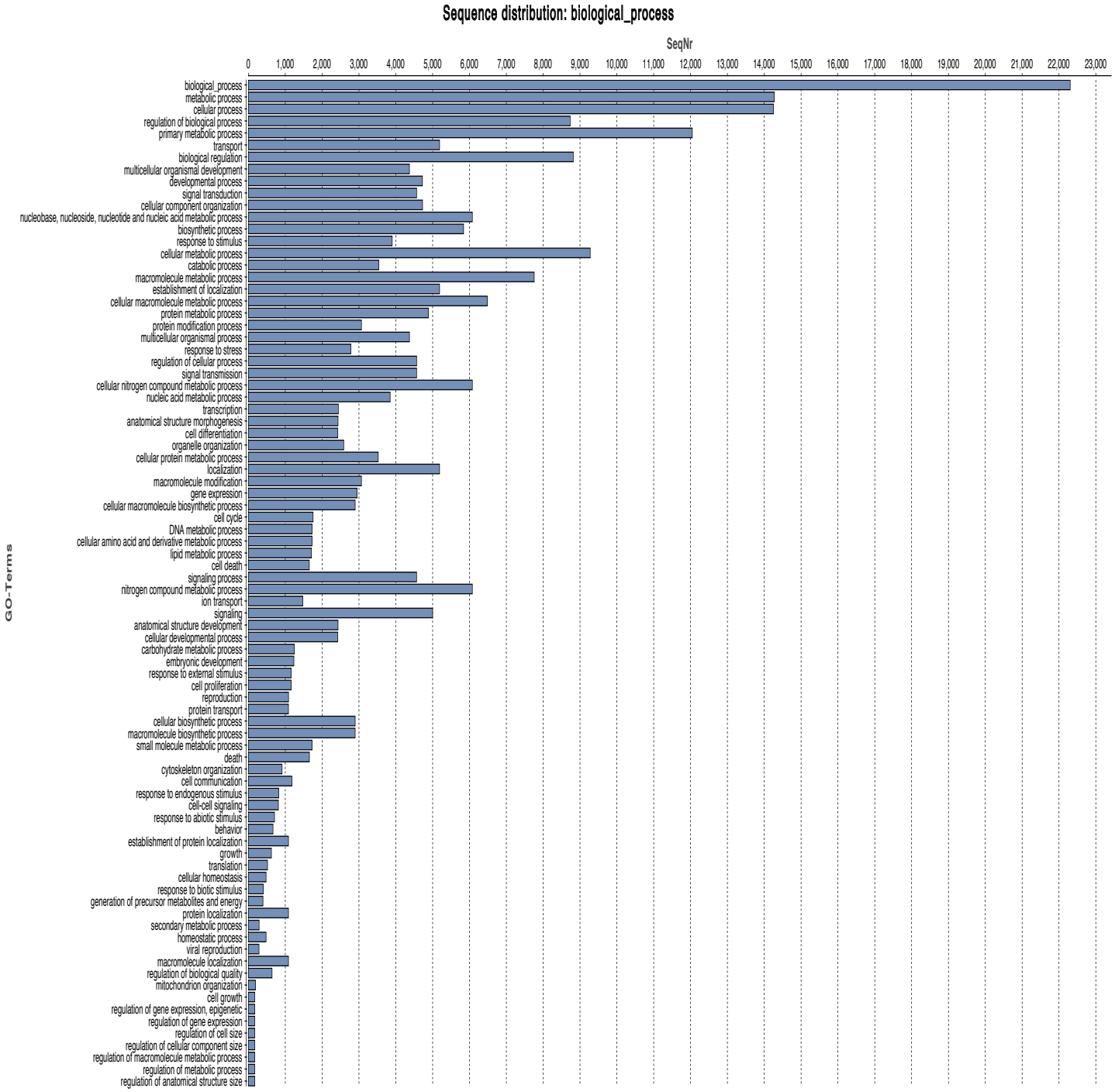

Figure C3.4d: Functional annotation of *Hermodice carunculata* transcripts. 55 most frequent GOslim term asseingment under biological proces category.

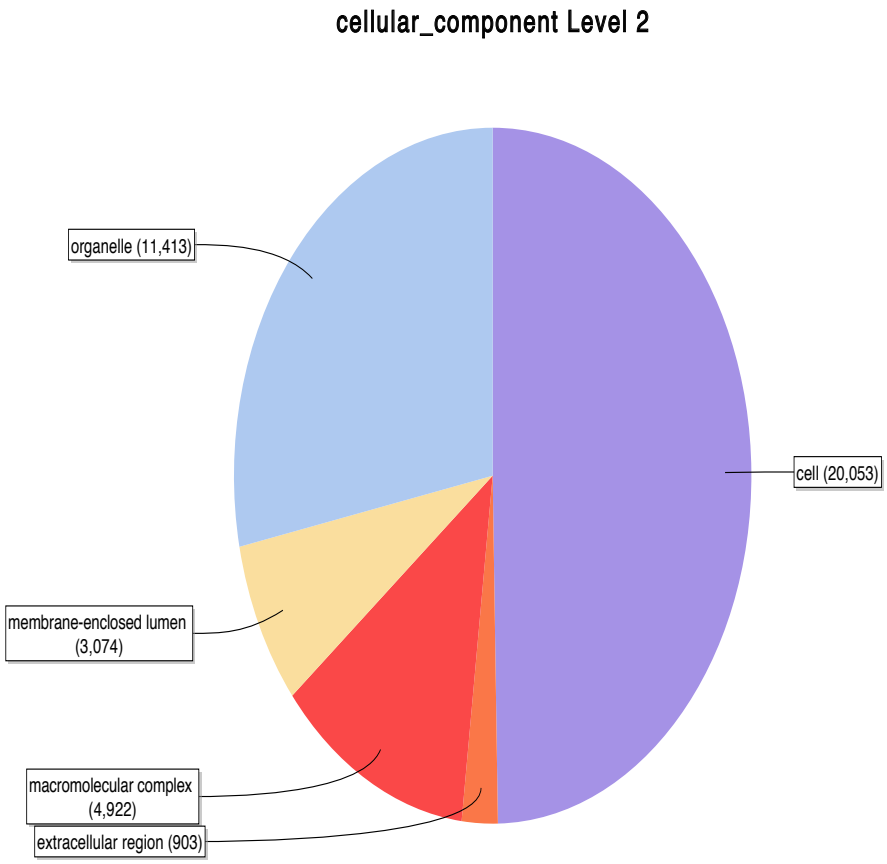

Figure C3.4e: Functional annotation of *Hermodice carunculata* transcripts. GOslim term asseingment under cellular component category.

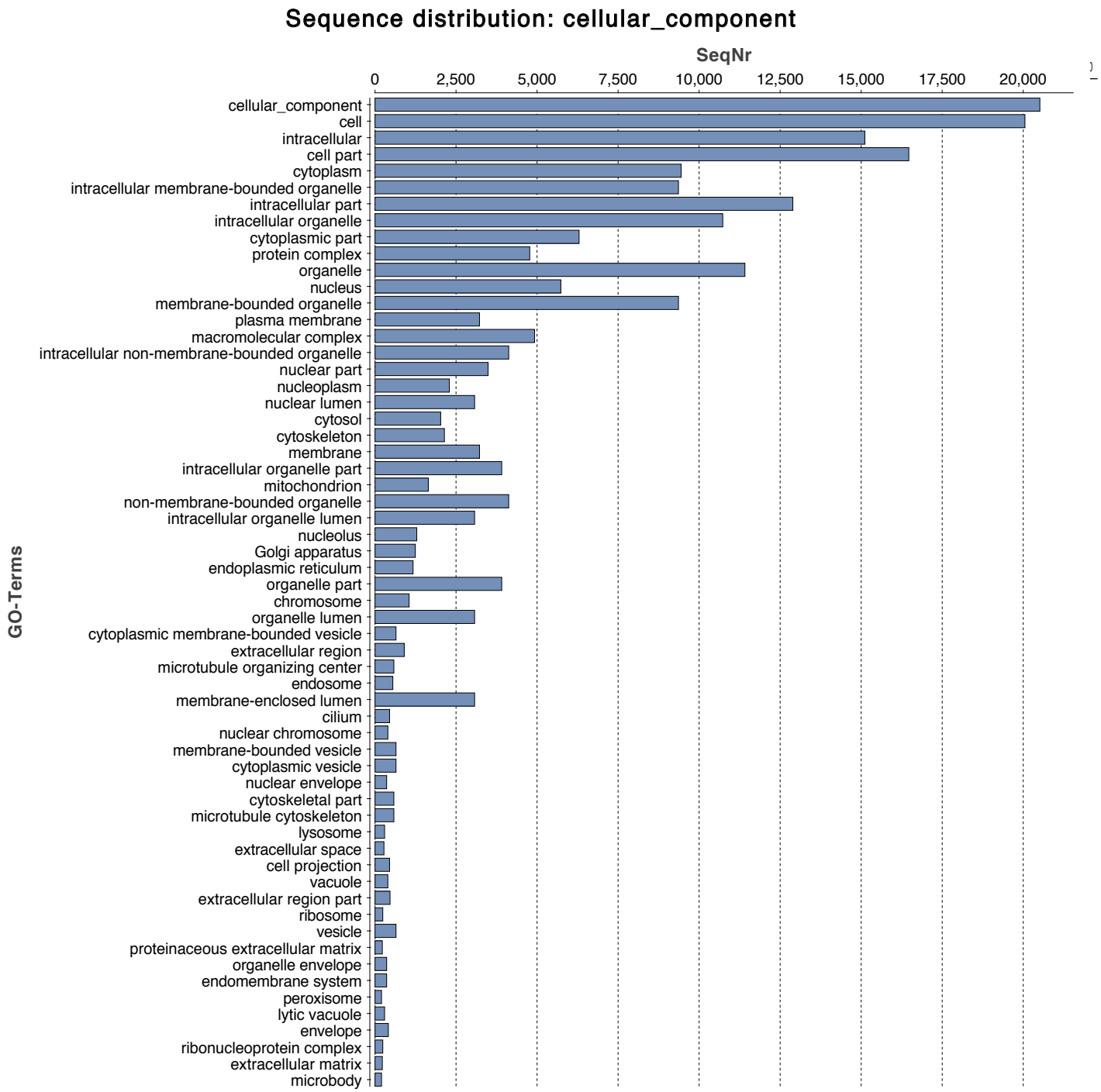

Figure C3.4f: Functional annotation of *Hermodice carunculata* transcripts. 55 most frequent GOslim term asseingment under cellular component category.

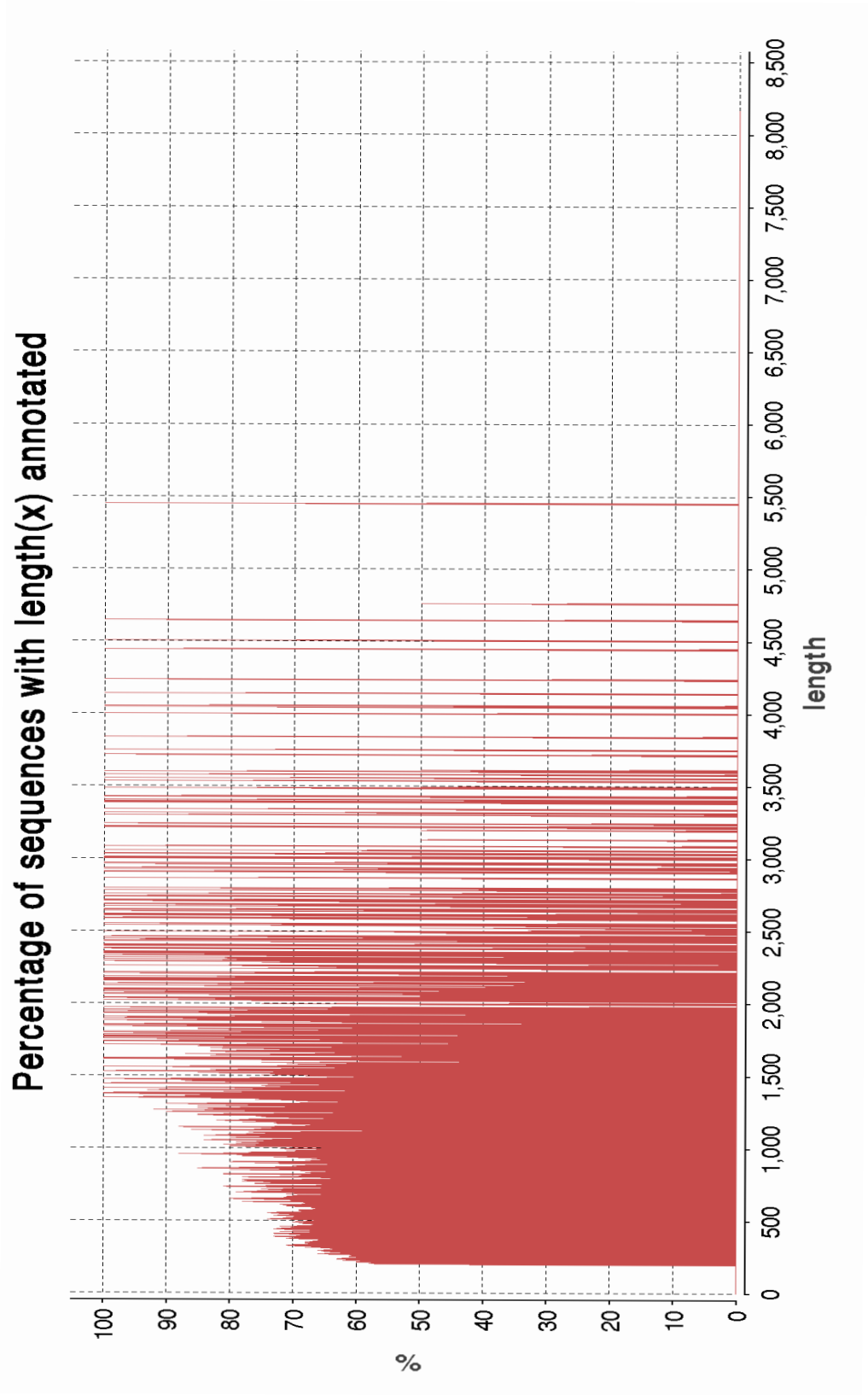

Figure C3.5: Percentage of functionally annotated transcripts relative to their length

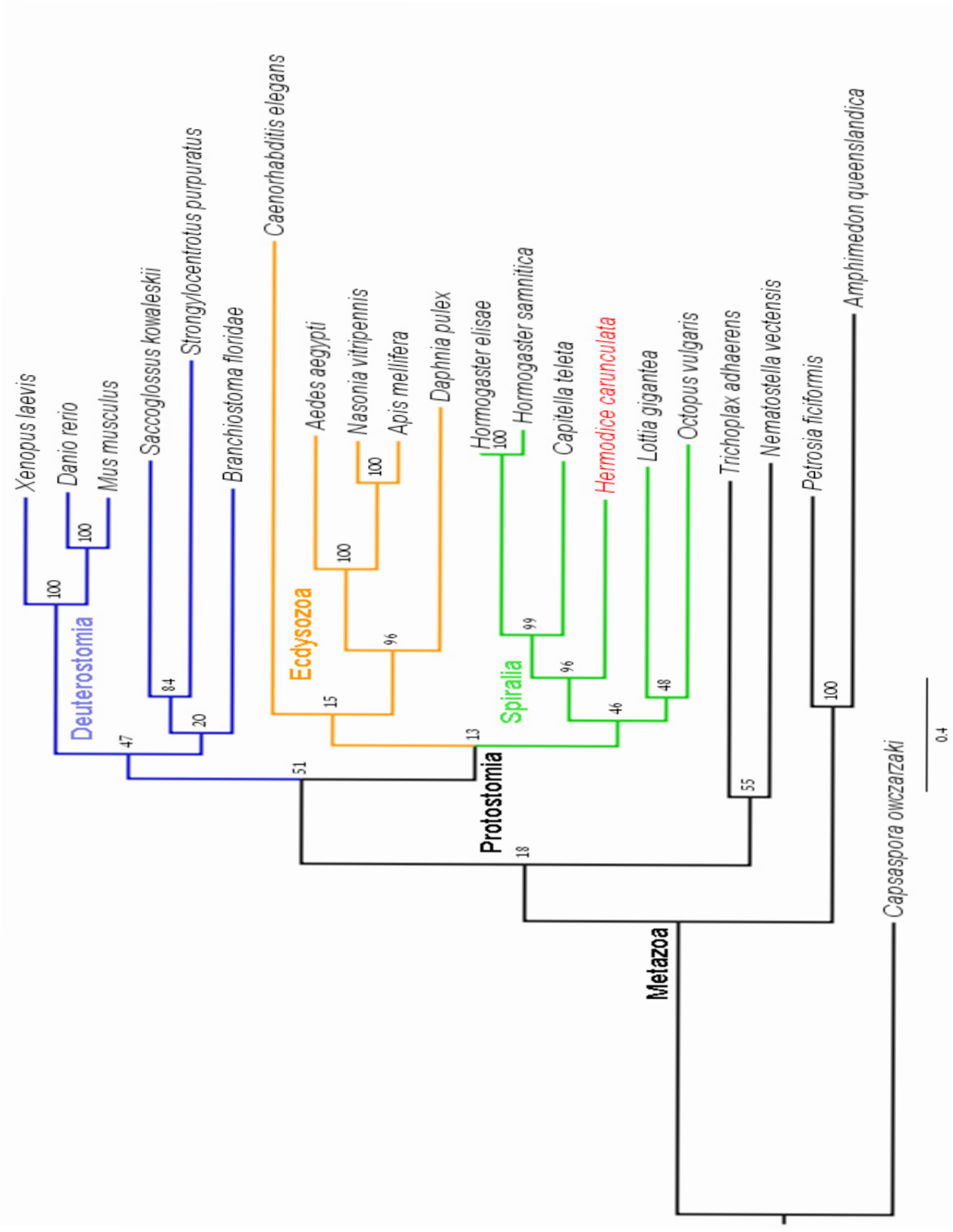

Figure C3.6: Maximum likelihood tree of 21 Attractin proteins and one of the newly identified sequences from *H.carunculata*. The newly identified Attractin is colored red. See additional file S9 for informations on alignment.

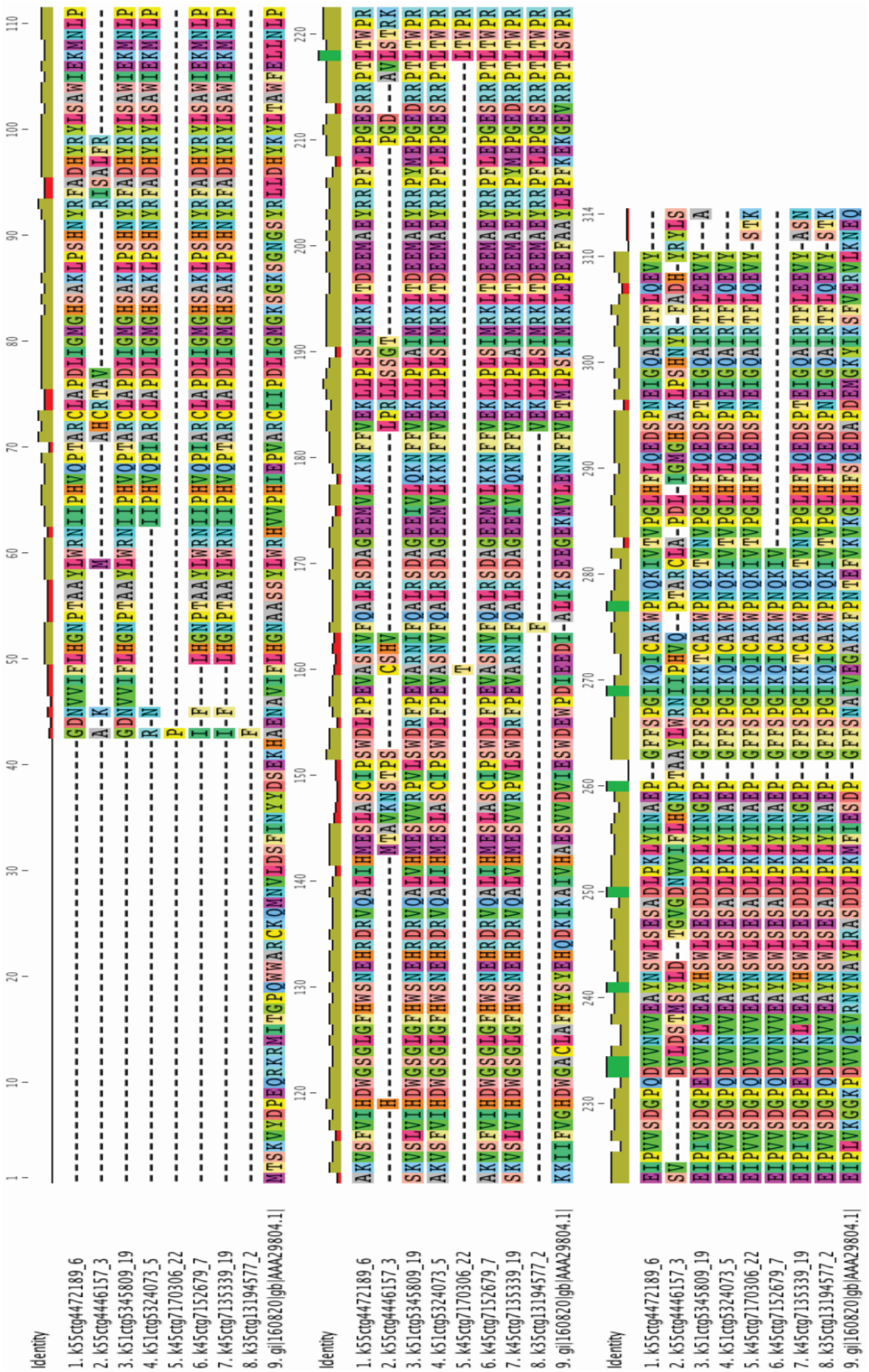

Figure C3.7: Overlapping region of amino acid sequence alignment of identified homologous proteins sequences to bioluminescent related protein (luciferase) from sea pansy *Renilla.*

Chapter 4: Fluorescent Proteins in chlopsid eels

C4.1: Abstract

Discovery of the naturally expressed fluorescent proteins has expanded the utility of fluorescent protein toolbox, useful for biomedical research. Using muscle transcriptome deep sequencing, the target transcripts were predicted and supported by protein mass spectrometry approach. Here, we report twelve new chromogenic ligand-inducible fluorescent proteins from a false moray eel, *Kaupichthys hyoproroides*. These sequences are homologous to the protein UnaG from *Anguilla japonica*, identified as a fluorescent reporter in hypoxic or anoxic conditions, and a bilirubin quantifier biomarker in human health clinical assays. In addition, we report for the first time deep sequencing of muscle transcriptome from *Kaupichthys hyoproroides*, and its functional annotation. These data provide a useful resource for future molecular ecology, functional genomics of this species. These data brings a 26,628 % increase (52 to 13,904) of these species available genetic data in NCBI.

C4.2: Background

The mechanism of light production through absorbing photons and releasing the excited electrons at a higher energy states, distinguishes fluorescence from other natural optical phenomena such as bioluminescence and phosphorescence. As the excited electrons relax to their basal state, they release energy at the longer wavelength. This excitation and

relaxation procedure causes the emission of a visible light. Conventional Green Fluorescent Proteins (GFPs) have constitutive fluorescence. Their chromophore is located near the center of the protein. Spontaneous formation of chromophore, within the folded beta-can, includes cyclization, loss of water, and oxidation. In contrast to this archetype, near-infrared proteins such as IFP1.4 [1], iRFP [2], and Wi-Phy [3] require external covalently bound (phytochrome based) biliverdin for chromophore formation to reveal the fluorescence property. In addition to these fluorescent proteins, recent discovery of a protein with non-covalent, highly specific bilirubin binding utility, from *Anguilla japonica* has expanded the conception of variation in mechanism of activation for this family of proteins.

GFPs, with covalently bound and autocatalytic chromophore, are phylogenetically diverse. They were originally discovered co-localized with the luminescent photocytes of the jellyfish, *Aequorea victoria* [4]. Fluorescent Proteins (FPs) have been found to be widespread in non-bioluminescent anthozoans, especially scleractinian corals [5-9]. In some anthozoans, pigments homologous to the FPs constitute up to 14% of the soluble protein content [10], likely constituting an important protein for these organisms, although the natural function is still unclear. In addition, homologues proteins to FPs have recently been discovered in non-luminous planktonic copepods [11], lancelets [12], and ctenophore [13]. All together, these constitute four major phyla including cnidarians, arthropods, ctenophora, and chordates.

In 2009, multiple peptides contributing to a novel GFP-like protein from *Anguilla japonica* were partially purified [14]. This protein, UnaG, has a molecular weight of ~ 16 kDa [14], and has been characterized as a fatty-acid-binding protein (FABP), and a ligand-inducible FP. It is reported producing oxygen-independent green fluorescent protein when bound to bilirubin, a membrane-permeable heme metabolite [15]. Although quite different from traditional GFP, UnaG is the first reported photoprotein in vertebrates, with a unique clinical and bioengineering use, as a bilirubin quantifying biomarker.

Anguilla is one of 111 genera within the order Anguilliformes [16]. This order consists of four major suborders, and one of these, Anguilloidei consisting seven families, includes Anguillidae (freshwater eels), Chlopsidae (false morays). Anguillidae is famous for its migratory behavior between growth habitats in freshwater and spawning areas, in tropical and subtropical areas [17]. In contrast, its sister clade Chlopsidae, lives on coral reefs with over 100 species occupying tropical oceans worldwide [18].

Identification of homologous sequences in cDNA libraries is reliant on some level of sequence or pattern similarity. In the absence of full-length sequence similarity, identification of reliable full-length transcripts encoding for the specific trait (i.e. fluorescent property) is a challenging job. The accuracy of the predicted target transcripts can be supported by identical, though fragmented, small peptides identifiable by protein purification approaches such as Mass Spectrometry (MS). MS can provide insights into the ingredient molecules of the sample and their fragmentation. We hypothesized here

that highly expressed transcripts should share sequence identity with highly abundant small peptide sequences from MS methods.

In this study, we report the identification of the twelve homologous proteins to UnaG FPs from a false moray eel, *Kaupichthys hyoproroides*. We used deep sequencing cDNA library to identify transcripts homologous to UnaG. Furthermore, we used Liquid Chromotography-Mas Spectrometry (LC/MS) to find high confident homologous peptide fragments for these proteins. In addition, we report, for the first time the *de novo* assembled transcriptome from muscle tissue in this species, and its functional annotation. This will provide 1) a useful genomic resource for future functional genomics studies in this species, 2) an expanded toolbox for bilirubin clinical biomarker.

C4.3: Results

C4.3.1: Sample identification

The eel was identified as *Kaupichthys hyoproroides* (Order: Anguilliformes) based on the wide, duck-like snout and pectoral placement (J. Sparks, AMNH), a shy and rare animal.

C4.3.2: Checking the fluorescent property of the animal

Kaupichthys hyoproroides exhibits bright green fluorescence with an emission peak of ~535nm (Figure C4.1) in the animals trunk musculature skin and muscle tissue, with the

skin being almost entirely green fluorescent (Figure C4.2) under blue (450-500nm) illumination. This peak is very similar to UnaG spectra isolated from from *Anguilla japonica*, with 535nm.

C4.3.3: *De novo* assembly

The RNA samples were prepared from muscle tissue of the tail of *Kaupichthys hyoproroides*, and sequencing was performed in a multiplexed lane of a flow cell using an Illumina Hi-seq 2000. To obtain better assembly, we examined two *de novo* assembly tools such as Trinity [19], Trans-ABySS [20], to assemble 142, 526,414 M reads.

First, to capture both low and high expressed transcripts, we applied 14 different k-mer lengths (21 to 59) for Trans-ABySS (with default parameters) [20]. We also performed assemblies with Trinity and compared its performance criteria such as contig number, N50 length, and maximum contig length with Trans-AByss. While the number of contigs from Trans-ABySS seemed larger (206,683 for Trans-ABySS; 84,610 for Trinity), the N50 (877pb for Trans-ABySS; 880pb for Trinity) and maximum contig length (8,547 bp for Trans-ABySS; 13,309bp for Trinity) were larger in Trinity compared to Trans-ABySS. These results indicated that Trans-ABySS produces greater number of shorter contigs than Trinity. As previously reported results [21]. Because of higher N50 and maximum contig length generated by Trinity, we chose to use the *de novo* assembled sequences generated by this program for the downstream analysis. These data contained 84,610 sequences, with N50 of 880pb (Table C4.1). The mean length of assembled

transcripts in this study was somewhat higher than that obtained in trascriptome studies in eel *Anguilla Anguilla*, 531 bp [22]. A Fasta file of these transcripts is available as an additional file (File C4.S1).

C4.3.4: *In Silico* quantification of transcripts

In order to identify potentially poor quality transcripts, reads were mapped back onto the non-redundant set of assembled transcripts using Bowtie [23]. A total of 109,268,961 (76.67%) of reads had at least one reported alignment. The minimum coverage of a transcript was 0.03 FPKM, and the maximum was 62,622 with the average of 9.44, indicating a wide range of gene expression (Table C4.2). Among these, 65,877 (77.85%) transcripts had FPKM >1, with the average of 11.93. Also, two transcripts had FPKM bigger than 20,000, with homology to Parvalbumin and muscle related Actin, with FPKM of 24,006 and 62,622, respectively.

C4.3.5: Functional annotation of the assembled transcripts

Using EMBOSS package [24], all possible open reading frames (ORFs), from stop to stop for each assembled contig was generated. ORFs longer than 150 AA (23,768) were searched for similarities against NCBI non-redundant (NR) protein database, using BlastP (E-value $\leq 2e^{-10}$) [25]. We used Blast2GO [26] for running BlastP, mapping, and GO term assignment to each transcript query sequence (default values of annotation

parameters were chosen). Furthermore, InterPro (domain annotation) IDs assigned to sequences using InterProScan (part of Blast2GO pipeline).

Out of 23,768 predicted ORFs, excluding 9,864 with 'NA', 13,904 (58.50%) sequences were annotated with either GO terms or InterPro IDs. This provides a rough estimate of the number of genes expressed in muscle tissue of *Kaupichthys hyoproroides*. A Fasta file, and a table file of assigned functions of these annotated ORFs, is available as an additional file. (File C4.S2; File C4.S3)

Based on assigned GO terms, sequences were classified into three categories (GOslim): biological process, cellular component and molecular function. In the molecular function, the terms relating to "binding", "catalytic activity" and "transporter activity" were enriched (654, 674 and 192, respectively) (Figures C4.3A). These terms represented 88% of the total. In cellular component category, the cluster size of "cell" with 727 sequences and "organelle" with 215 sequences were highly represented compared to " "extracellular matrix" with less than 12 sequences (Figure C4.3B). In the biological process classification, "metabolic process" with 685 sequences, "cellular processes" with 444 sequences, "localization" with 259 sequences were large compared to "reproduction" and "growth" (Figures C4.3C). This is somewhat expected, as these data are not collected from developmental stages with a high rate of growth or reproduction. This pattern is very similar to a recent analysis of *Lymnae stagnalis* (pond snail) transcriptome functional annotation [27].

C4.3.6: Identification of Bilirubin-Inducible Fluorescent (BIF) transcripts (UnaG homologs)

Since we observed fluorescent property in muscle and skin (Figure C4.2B; C4.2C), we hypothesized that the deep sequenced cDNA library should provide a useful resource for GFP homology search. In order to identify putative green fluorescent transcript isoforms, 156 naturally expressed GFP sequences from NCBI (based on accession numbers from [28]), and nine partial novel fluorescent peptides from *Anguilla japonica* (reported in table 2 of [14]), were downloaded. After using EMBOSS package [24] to generate all possible open reading frames (ORFs) from stop to stop for each assembled contig, ORFs were searched for sequence similarity against the GFP sequences and partial peptides, using BlastP (E-value $\leq 2e^{-10}$). For the 434,342 predicted ORFs longer than 150bp, 12 unique ORFs showed considerable sequence similarity to the eel polypeptide [14] composed of peptide 5 (LVYVQK), 6 (WDGKETTVR) and 7 (ELSDGGDATTPTL). Parallel to these findings, the cDNA encoding for full-length protein (139 AA) similar to this peptide was recently characterized as a Fatty-acid-binding protein (FABP) [14]. According to this characterization, this protein shows fluorescent property. It holds bilirubin-inducible activation property and is oxygen-independent (UnaG). In order to examine the homology of newly characterized BIF protein (UnaG) to *Kaupichthys hyoproroides* 12 transcript isoforms, we included this sequence into the alignment, as well as a FABP sequence from a cod fish, *Notothenia coriiceps* (ID= AAC60358.1). The alignment of these sequences with UnaG from *Anguilla japonica* showed a considerable homology (Figure C4.4). The homology score between UnaG and these ORFs ranges

from 55.4% to 24.1%. Two of the sequences (comp92260_c0_seq1, and comp92260_c1_seq1) were 100% identical at the amino acid level, but their nucleotide sequence was different. Therefore, we considered them as separate transcript isoforms and kept both of them. These 12 transcripts share 5 out of 7 residues critical for bilirubin binding in UnaG (marked red in Figure C4.4), indicating that most likely they have Bilirubin-Inducible Fluorescent (BIF) property, making good candidates for further validation.

C4.3.7: Identification of purified muscle peptides with 100% identity to BIF transcripts

In order to give more support to predicted transcript isoforms, we carried out protein purification followed by mass spectrometry. Non-denatured extract of the muscle tissue was clarified (see material and methods) and loaded the soluble fraction onto a native gel. Two fluorescent bands were detected (Figure C4.5), excised and subjected to trypsin digestion and LC-MS analysis. A trypsin peptide library of molecular weight was led to identification of about 6,680 short peptides from each band. A fasta file was generated using these peptides. These sequences were searched against putative *Kaupichthys hyoproroides* Bilirubin-Inducible Fluorescent (BIF) transcripts, using BlastP with special parameters for short sequence search (options = -W 2 -P 0 -e $2e^{-01}$). Several matches were found to align with 100% identity to each of these transcripts, suggesting that the transcripts are likely to encode for putative BIF proteins.

C4.4: Discussion and conclusion

Over the past decade, the discovery and application of fluorescent proteins in biomedical research has brought improvements in many fields [29, 30]. Mechanism of chromophore formation in commonly known fluorescent proteins (GFPs) is autocatalytic, and oxygen dependent [31, 32]. This utility limits the application of these proteins in an aerobic environment. Although other proteins such as Flavin mononucleotide (FMN)-based fluorescent proteins have been engineered to show constitutive emission in aerobic and anaerobic systems, the generated fluorescent property is very weak [32]. Therefore, it is important to have more oxygen-independent fluorescent proteins with strong emission property, useful as a reporter sensor.

In most cases, tags with fluorogenic chromophores offer a unique inducible fluorescent optical property. They remain non-fluorescent until covalently bound to their exogenous chemical ligand. Therefore, they can be used as switch-like fluorogenic probes. In the case of UnaG and its homologous proteins, bilirubin chromogenic compound is located inside in the beta-barrel and it gives the protein a similar inducible property. Because it is highly specific to bilirubin, its switch-like application is limited in eukaryotic cell line with no endogenous expression of bilirubin, or bacterial systems. Therefore, some further genetic modification might be required to make the bilirubin replaceable with its chemical variant. These newly identified proteins from *Kaupichthys hyoproroides* will expand the versatility of this toolbox and its further application.

Bilirubin is known to have physiological antioxidant and cell-protectant property [33] [34]. Considering this possibility, the interaction of UnaG with bilirubin may play an important role in regulating the free or transient pools of this metabolite for reducing the cellular oxidative stress [15]. This non-covalent binding of bilirubin to UnaG has been attributed to its function in maintaining muscle oxidative metabolism during long distance migration. Our findings support the bilirubin contribution in muscle metabolism, as it is expressed in the muscle of *Kaupichthys hyoproroides* as well. However its implication seems to be broader than just migration purposes, as *Kaupichthys hyoproroides* is not a migratory fish.

Altogether, identification of BIF in non-migratory moray eels shows that this protein has been under convergent evolution in two families of Anguilloidei, both migratory and non-migratory eel species. Although these results require further follow-up studies to shed light on physiology of these proteins, they bring a broad functional perspective for the role of this protein in these species.

C4.5: Material and Methods

Sample Collection

The *Kaupichthys hyoproroides* specimen were collected via SCUBA in the Caribbean (off Lee Stocking Island, Bahamas). Rotenone was dispersed over a confined area on the

reef slope or floor. The specimen was collected using small hand nets, and bagged for transport back to the surface.

Emission spectra were collected using an Ocean Optics USB2000+ miniature spectrometer (Dunedin, FL) equipped with a hand-held fiber optic probe (Ocean Optics ZFQ-12135). Excitation spectra were achieved by the incorporation of particular band-pass filters (Chroma Optics, Inc., VT). Emission spectra were recorded by applying the fiber optic probe to specific anatomical parts of the individual fish specimen exhibiting biofluorescence. This was repeated several times for each specimen to ensure the accuracy of these measurements.

Individual eel specimens were placed in a narrow photographic tank. They were held flat against a thin plate glass front for photography. Fluorescent macro images [4928 x 3264 (Nikon D7000); 2180 x 1800 pixel (Nikon D300S)] were produced in a dark room by covering the flash (Nikon SB 600 and SB 800) with interference bandpass excitation filters (Omega Optical, Brattleboro, VT; Semrock, Rochester, NY). Longpass (LP) and bandpass (BP) emission filters (Semrock) were attached to the front of the camera lens. A variety of excitation/emission filter pairs was tested on each sample to elicit the strongest fluorescence emission: excitation 450–500nm, emission 514 LP; excitation 500–550nm, emission 555 and 561 LP.

RNA Extraction and deep cDNA Sequencing

Specimens of *Kaupichthys hyoproroides* were frozen in liquid nitrogen. Total RNA extracted from dissected tail muscles of a specimen. Muscle tissue was in TriZol reagent (Life Technologies, NY) and the total RNA was precipitated with isopropanol and dissolved in ddH2O. The quality of RNA was assessed on a 2100 Bioanalyzer and with agarose gel electrophoresis-. The total RNA was pooled for Library preparation. Libraries were prepared with Hi-seq RNA sample preparation kit (Illumina Inc, San Diego, CA) according to manufacturer's instructions. One lane was multiplexed for four samples and was as 80-bp PE reads. FASTQ file generation was performed by CASAVA version 1.8.2 (llumina).

***De novo* Assembly**

All the assemblies were performed on a server with 50 cores and 250 GB random access memory. Obtained reads were de novo assembled using Trinity [19] and trans-ABySS [20]. Trinity was used with the default parameter. Trans-ABySS was executed using different k-mer length from 21 to 59, along with other default parameters.

Assembly validation

Gene coverage levels were determined using a Perl script to calculate the RPKM [28]. Contigs with RPKM smaller than 1 were removed for functional annotation.

BIF protein annotation

A set of possible ORFs, stop to stop from assembled transcripts by Trinity, was generated using EMBOSS-ref-. To annotate the BIF homologs, a similarity search against the small polypeptide reported from [14] was conducted using BlastP with E value of $2e^{-10}$.

Functional annotation

Using the 23,000 ORFs sequences longer than 150 AA, assignment of gene ontology (GO) terms was performed by running BlastP, mapping and annotation with Blast2GO.

LC/MS

A non-denatured muscle was extract clarified by high-speed centrifugation, (50,000Xg for 30 min) and the soluble fraction loaded onto a native gel (non-denaturing gel). Bands were excised and subjected to trypsin digestion and LC-MS analysis. A trypsin peptide library of molecular weight generated from the gel and searched against the fish database. List of abundant peptides were downloaded and searched against putative BIF transcripts, using BlastP (options = -W 2 -P 0 -e $2e^{-01}$).

Table C4.1 Summary statistics for individual assemblies

Assembly	Number of trancripts >200 bp	N50 bp	Mean length bp	Max length bp	Total number of bases
TransABySS	206,683	877	672.11	8,547	138,913,931
Trinity	84,610	880	641.00	13,309	54,235,411

N50= length-weigthed median contig length; bp= base pair; ORF= Open Reading Frame

Table C4.2 Summary statistics of read counts and coverage

Total number of reads	142,526,414
Number of read used reads for assembly	109,268,961 (76.67%)
Number of unused reads	33,257,453 (23.33%)
Number of non-redundant transcripts (>200 bp)	84,610
Number of transcripts with coverage fpkm >1 (dataset 1)	65,877
Average coverage for contigs with coverage fpkm >1	11.93
Average number of reads mapped per contigs in dataset 1	1,649.71

bp = base pair; fpkm = paired-reads per kilo base per million ; contig= contiguous, overlapping sequence read resulting from the assembly

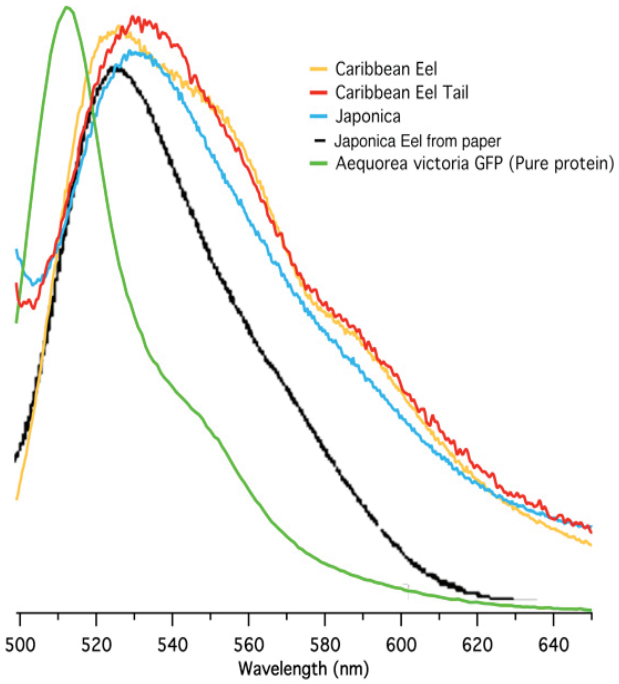

Figure C4.1: Comparative spectra of *Kaupichthys hyoproroides*, *Anguilla japonica* (UnaG), and purified *Aequorea victoria* GFP

Figure C4. 2. Fluorescent a) head; b) skin and c) muscle of *Kaupichthys hyoproroides (*Left). White light (top right) and fluorescent illumination (bottom right) image of eel (Lee Stocking Island, BWI).

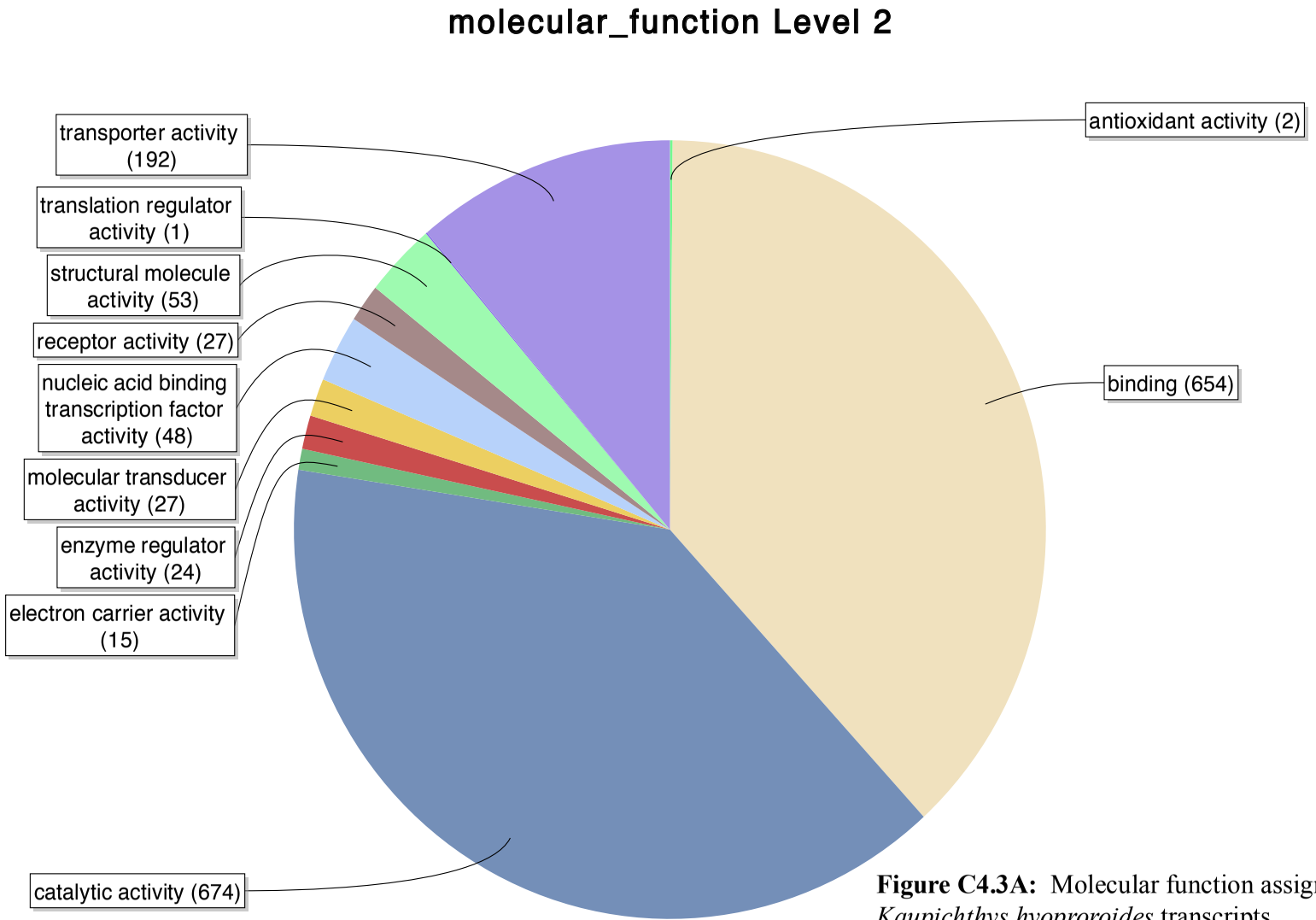

Figure C4.3A: Molecular function assignment for *Kaupichthys hyoproroides* transcripts.

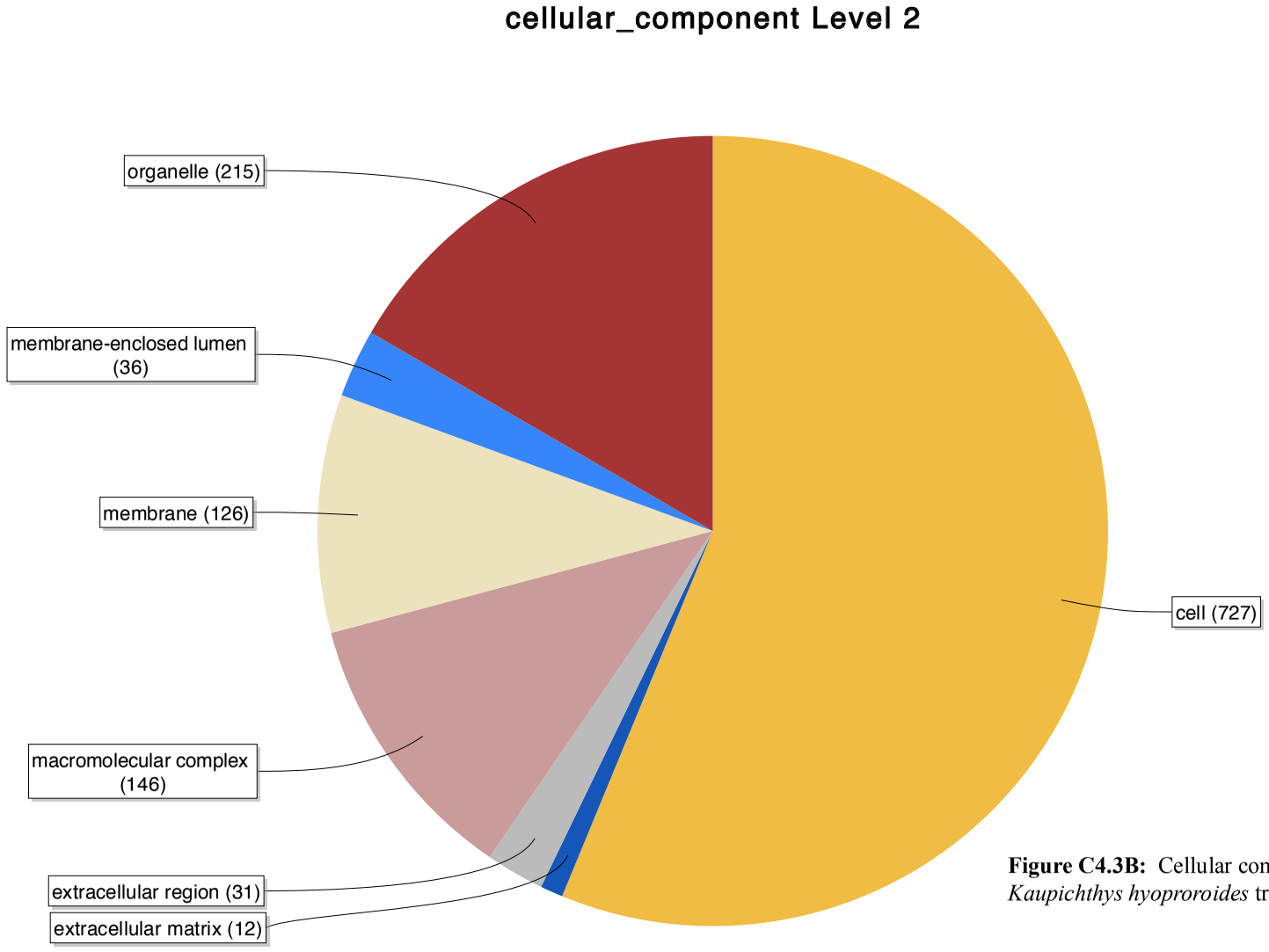

Figure C4.3B: Cellular component assignment for *Kaupichthys hyoproroides* transcripts.

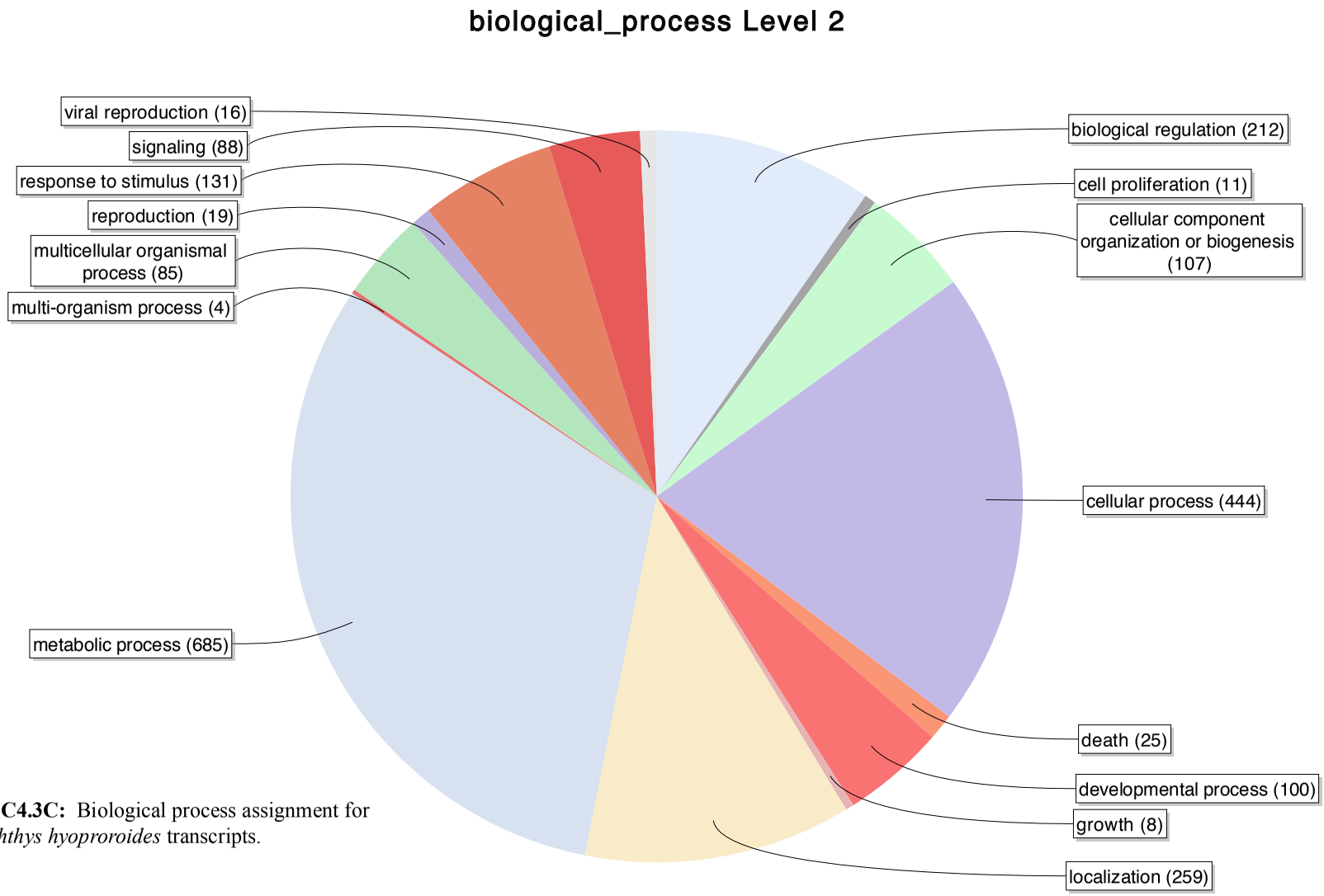

Figure C4.3C: Biological process assignment for *Kaupichthys hyoproroides* transcripts.

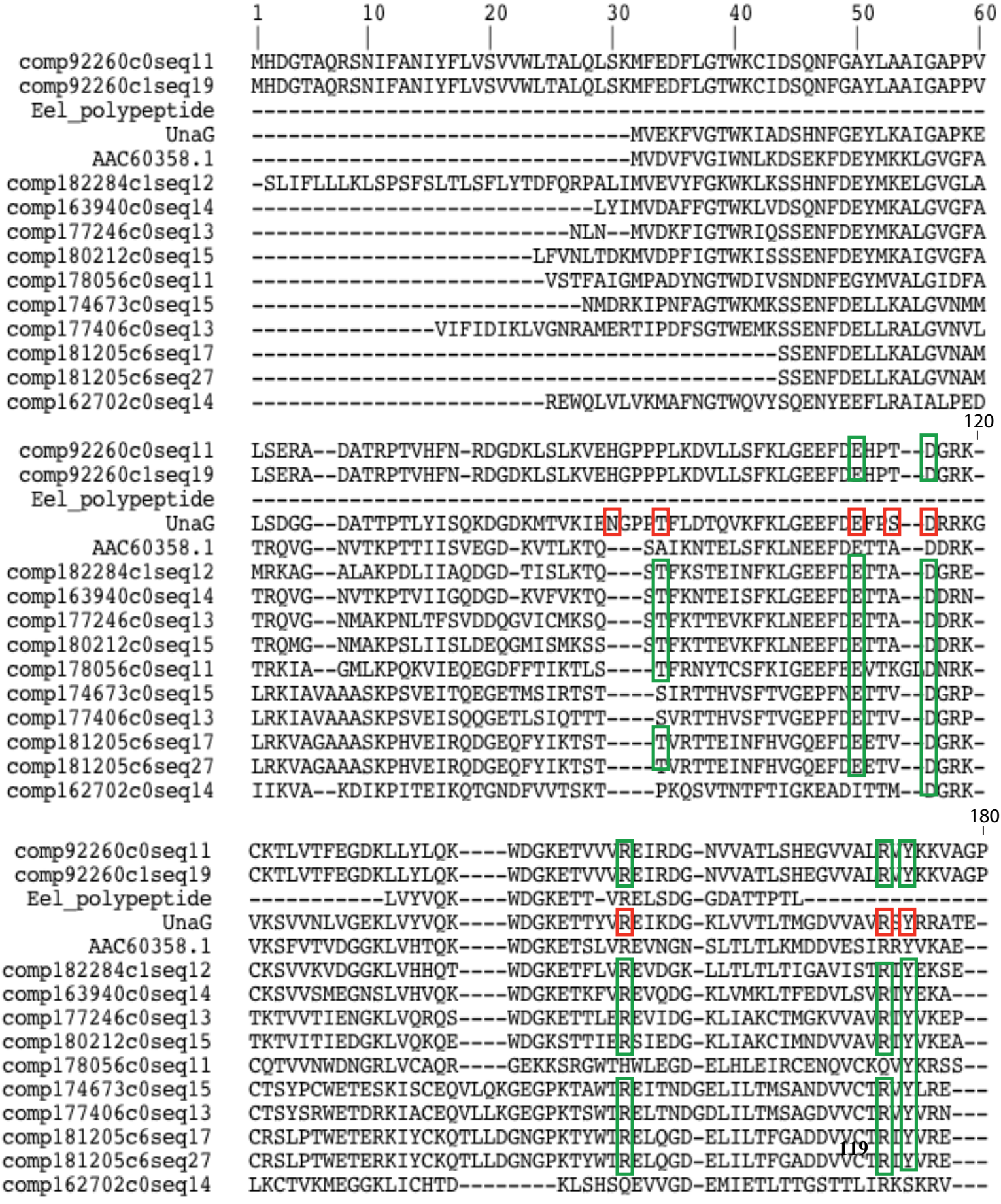
1 10 20 30 40 50 60
comp92260c0seq11 MHDGTAQRSNIFANIYFLVSVVWLTALQLSKMFEDFLGTWKCIDSQNFGAYLAAIGAPPV
comp92260c1seq19 MHDGTAQRSNIFANIYFLVSVVWLTALQLSKMFEDFLGTWKCIDSQNFGAYLAAIGAPPV
Eel_polypeptide
UnaG MVEKFVGTWKIADSHNFGEYLKAIGAPKE
AAC60358.1 MVDVFVGIWNLKDSEKFDEYMKKLGVGFA
comp182284c1seq12 -SLIFLLLKLSPSFSLTLSFLYTDFQRPALIMVEVYFGKWKLKSSHNFDEYMKELGVGLA
comp163940c0seq14 LYIMVDAFFGTWKLVDSQNFDEYMKALGVGFA
comp177246c0seq13 NLN--MVDKFIGTWRIQSSENFDEYMKALGVGFA
comp180212c0seq15 LFVNLTDKMVDPFIGTWKISSSENFDEYMKAIGVGFA
comp178056c0seq11 VSTFAIGMPADYNGTWDIVSNDNFEGYMVALGIDFA
comp174673c0seq15 NMDRKIPNFAGTWKMKSSENFDELLKALGVNMM
comp177406c0seq13 VIFIDIKLVGNRAMERTIPDFSGTWEMKSSENFDELLRALGVNVL
comp181205c6seq17 SSENFDELLKALGVNAM
comp181205c6seq27 SSENFDELLKALGVNAM
comp162702c0seq14 REWQLVLVKMAFNGTWQVYSQENYEEFLRAIALPED
120
comp92260c0seq11 LSERA--DATRPTVHFN-RDGDKLSLKVEHGPPPLKDVLLSFKLGEEFDEHPT--DGRK-
comp92260c1seq19 LSERA--DATRPTVHFN-RDGDKLSLKVEHGPPPLKDVLLSFKLGEEFDEHPT--DGRK-
Eel_polypeptide
UnaG LSDGG--DATTPTLYISQKDGDKMTVKIENGPPTFLDTQVKFKLGEEFDEFPS--DRRKG
AAC60358.1 TRQVG--NVTKPTTIISVEGD-KVTLKTQ---SAIKNTELSFKLNEEFDETTA--DDRK-
comp182284c1seq12 MRKAG--ALAKPDLIIAQDGD-TISLKTQ---STFKSTEINFKLGEEFDETTA--DGRE-
comp163940c0seq14 TRQVG--NVTKPTVIIGQDGD-KVFVKTQ---STFKNTEISFKLGEEFDETTA--DDRN-
comp177246c0seq13 TRQVG--NMAKPNLTFSVDDQGVICMKSQ---STFKTTEVKFKLNEEFDETTA--DDRK-
comp180212c0seq15 TRQMG--NMAKPSLIISLDEQGMISMKSS---STFKTTEVKFKLNEEFDETTA--DDRK-
comp178056c0seq11 TRKIA--GMLKPQKVIEQEGDFFTIKTLS----TFRNYTCSFKIGEEFEEVTKGLDNRK-
comp174673c0seq15 LRKIAVAAASKPSVEITQEGETMSIRTST----SIRTTHVSFTVGEPFNETTV--DGRP-
comp177406c0seq13 LRKIAVAAASKPSVEISQQGETLSIQTTT----SVRTTHVSFTVGEPFDETTV--DGRP-
comp181205c6seq17 LRKVAGAAASKPHVEIRQDGEQFYIKTST----TVRTTEINFHVGQEFDEETV--DGRK-
comp181205c6seq27 LRKVAGAAASKPHVEIRQDGEQFYIKTST----TVRTTEINFHVGQEFDEETV--DGRK-
comp162702c0seq14 IIKVA--KDIKPITEIKQTGNDFVVTSKT----PKQSVTNTFTIGKEADITTM--DGRK-
180
comp92260c0seq11 CKTLVTFEGDKLLYLQK----WDGKETVVVREIRDG-NVVATLSHEGVVALRVYKKVAGP
comp92260c1seq19 CKTLVTFEGDKLLYLQK----WDGKETVVVREIRDG-NVVATLSHEGVVALRVYKKVAGP
Eel_polypeptide LVYVQK----WDGKETT-VRELSDG-GDATTPTL
UnaG VKSVVNLVGEKLVYVQK----WDGKETTYVREIKDG-KLVVTLTMGDVVAVRSYRRATE-
AAC60358.1 VKSFVTVDGGKLVHTQK----WDGKETSLVREVNGN-SLTLTLKMDDVESIRRYVKAE--
comp182284c1seq12 CKSVVKVDGGKLVHHQT----WDGKETFLVREVDGK-LLTLTLTIGAVISTRIYEKSE--
comp163940c0seq14 CKSVVSMEGNSLVHVQK----WDGKETKFVREVQDG-KLVMKLTFEDVLSVRIYEKA---
comp177246c0seq13 TKTVVTIENGKLVQRQS----WDGKETTLEREVIDG-KLIAKCTMGKVVAVRIYVKEP--
comp180212c0seq15 TKTVITIEDGKLVQKQE----WDGKSTTIERSIEDG-KLIAKCIMNDVVAVRIYVKEA--
comp178056c0seq11 CQTVVNWDNGRLVCAQR----GEKKSRGWTHWLEGD-ELHLEIRCENQVCKQVYKRSS--
comp174673c0seq15 CTSYPCWETESKISCEQVLQKGEGPKTAWTREITNDGELILTMSANDVVCTRVYLRE---
comp177406c0seq13 CTSYSRWETDRKIACEQVLLKGEGPKTSWTRELTNDGDLILTMSAGDVVCTRVYVRN---
comp181205c6seq17 CRSLPTWETERKIYCKQTLLDGNGPKTYWTRELQGD-ELILTFGADDVVCTRIYVRE---
comp181205c6seq27 CRSLPTWETERKIYCKQTLLDGNGPKTYWTRELQGD-ELILTFGADDVVCTRIYVRE---
comp162702c0seq14 LKCTVKMEGGKLICHTD--------KLSHSQEVVGD-EMIETLTTGSTTLIRKSKRV---
119

Fig. C4.4. Twelve isoforms alignment, along with short peptide previously identified from Anguilla japonica, and one other fish (Notothenia coriiceps) fatty acid binding protein.

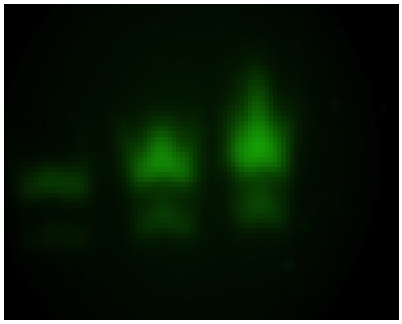

Figure C4.5: Fluorescing bands, non-denaturing gel

Chapter 1 Bibliography

1. Shimomura O: **The discovery of aequorin and green fluorescent protein**. *Journal of microscopy* 2005, **217**(1):3-15.
2. Tsien RY: **The green fluorescent protein**. *Annu. Rev. Biochem.* 1998, **67**(1):509-544.
3. Morin JG, Hastings J: **Energy transfer in a bioluminescent system**. *J. Cell. Physiol.* 1971, **77**(3):313-318.
4. Wampler JE, Hori K, Lee JW, Cormier MJ: **Structured bioluminescence. Two emitters during both the in vitro and the in vivo bioluminescence of the sea pansy, *Renilla***. *Biochemistry (Mosc.)* 1971, **10**(15):2903-2909.
5. Markova SV, Burakova LP, Frank LA, Golz S, Korostileva KA, Vysotski ES: **Green-fluorescent protein from the bioluminescent jellyfish *Clytia gregaria*: cDNA cloning, expression, and characterization of novel recombinant protein.** *Photochemical & photobiological sciences : Official journal of the European Photochemistry Association and the European Society for Photobiology* 2010, **9**(6):757-765.
6. Kumagai A, Ando R, Miyatake H, Greimel P, Kobayashi T, Hirabayashi Y, Shimogori T, Miyawaki A: **A Bilirubin-Inducible fluorescent protein from eel muscle**. *Cell* 2013.
7. Petersen J, Wilmann PG, Beddoe T, Oakley AJ, Devenish RJ, Prescott M, Rossjohn J: **The 2.0-Å crystal structure of eqFP611, a far red fluorescent**

protein from the sea anemone *Entacmaea quadricolor*. *J. Biol. Chem.* 2003, **278**(45):44626-44631.

8. Haddock SHD, Mastroianni N, Christianson LM: **A photoactivatable green-fluorescent protein from the phylum Ctenophora**. *Proc. R. Soc., B.* 2010, **277**(1685):1155-1160.

9. Shaner NC, Lambert GG, Chammas A, Ni Y, Cranfill PJ, Baird MA, Sell BR, Allen JR, Day RN, Israelsson M: **A bright monomeric green fluorescent protein derived from *Branchiostoma lanceolatum***. *Nat. Methods* 2013, **10**(5):407-409.

10. Deheyn DD, Kubokawa K, McCarthy JK, Murakami A, Porrachia M, Rouse GW, Holland ND: **Endogenous green fluorescent protein (GFP) in amphioxus**. *Biol. Bull.* 2007, **213**(2):95-100.

11. Szöllosi J, Damjanovich S, Mátyus L: **Application of fluorescence resonance energy transfer in the clinical laboratory: routine and research**. *Cytometry* 1998, **34**(4):159-179.

12. Ward WW, Cormier MJ: **Energy transfer via protein‐protein interaction in *Renilla* bioluminescence**. *Photochem. Photobiol.* 1978, **27**(4):389-396.

13. Morise H, Shimomura O, Johnson FH, Winant J: **Intermolecular energy transfer in the bioluminescent system of *Aequorea***. *Biochemistry (Mosc.)* 1974, **13**(12):2656-2662.

14. Kenaley CP, Hartel KE: **A revision of Atlantic species of Photostomias (Teleostei: Stomiidae: Malacosteinae), with a description of a new species**. *Ichthyol. Res.* 2005, **52**(3):251-263.

15. Haddock SH, Case JF: **A bioluminescent chaetognath**. *Nature* 1994, **367**:225-226.

16. Shimomura O: **Bioluminescence: chemical principles and methods**: World Scientific Publishing Company; 2012.

17. Haddock SH, Rivers TJ, Robison BH: **Can coelenterates make coelenterazine? Dietary requirement for luciferin in cnidarian bioluminescence**. *PNAS* 2001, **98**(20):11148-11151.

18. Simon S, Narechania A, DeSalle R, Hadrys H: **Insect phylogenomics: Exploring the source of incongruence using new transcriptomic data**. *Genome Biol. Evol.* 2012, **4**(12):1295-1309.

19. Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung H-JG, Vance CP, Gronwald JW: **Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems**. *BMC Genomics* 2011, **12**(1):199.

20. Mehr SF, DeSalle R, Kao H-T, Narechania A, Han Z, Tchernov D, Pieribone V, Gruber DF: **Transcriptome deep-sequencing and clustering of expressed isoforms from *Favia* corals**. *BMC Genomics* 2013, **14**:546.

21. Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T *et al*: **Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds**. *BMC Genomics* 2011, **12**:131.

22. Feng C, Chen M, Xu CJ, Bai L, Yin XR, Li X, Allan AC, Ferguson IB, Chen KS: **Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq**. *BMC Genomics* 2012, **13**:19.

23. Elder RT, Szabo P, Uhlenbeck OC: **In situ hybridization of three transfer RNAs to the polytene chromosomes of *Drosophila melanogaster***. *J. Mol. Biol.* 1980, **142**(1):1-17.

24. McNicol P, Guijon F, Wayne S, Hidajat R, Paraskevas M: **Expression of human papillomavirus type 16 E6-E7 open reading frame varies quantitatively in biopsy tissue from different grades of cervical intraepithelial neoplasia**. *J. Clin. Microbiol.* 1995, **33**(5):1169-1173.

25. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**(5235):467-470.

26. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF**. *Nature* 2001, **409**(6819):533-538.

27. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C)**. *Nat. Genet.* 2006, **38**(11):1348-1354.

28. Wilhelm BT, Landry JR: **RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing**. *Methods* 2009, **48**(3):249-257.

29. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat. Meth.* 2008, **5**(7):621-628.

30. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M: **The transcriptional landscape of the yeast genome defined by RNA sequencing**. *Science* 2008, **320**(5881):1344-1349.

31. Metzker ML: **Sequencing technologies - the next generation**. *Nat. Rev. Genet.* 2010, **11**(1):31-46.

32. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life**. *Nature* 2008, **452**(7188):745-749.

33. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y: **De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing**. *PLoS One* 2012, **7**(8):e42546.

34. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP: **De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology**. *PLoS One* 2010, **5**(12):e14202.

35. Garg R, Patel RK, Tyagi AK, Jain M: **De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification**. *DNA Res.* 2011, **18**(1):53-63.

36. Franchini P, Van der Merwe M, Roodt-Wilding R: **Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis**. *BMC Res. Notes* 2011, **4**(1):59.

37. Salem M, Vallejo RL, Leeds TD, Palti Y, Liu S, Sabbagh A, Rexroad CE, 3rd, Yao J: **RNA-Seq identifies SNP markers for growth traits in rainbow trout**. *PLoS One* 2012, **7**(5):e36264.

38. Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L: **SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus spp.)***. *Mol. Ecol.* 2011, **20**(3):545-559.

39. Canovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF: **SNP discovery in the bovine milk transcriptome using RNA-Seq technology**. *Mamm. Genome* 2010, **21**(11-12):592-598.

40. Hong LZ, Li J, Schmidt-Küntzel A, Warren WC, Barsh GS: **Digital gene expression for non-model organisms**. *Genome Res.* 2011, **21**(11):1905-1915.

41. Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S: **De novo characterization of a whitefly transcriptome and analysis of its gene expression during development**. *BMC Genomics* 2010, **11**(1):400.

42. Shendure J, Ji H: **Next-generation DNA sequencing**. *Nat. Biotechnol.* 2008, **26**(10):1135-1145.

43. Li P, Ponnala L, Gandotra N, Wang L, Si Y, Tausta SL, Kebrom TH, Provart N, Patel R, Myers CR *et al*: **The developmental dynamics of the maize leaf transcriptome**. *Nat. Genet.* 2010, **42**(12):1060-1067.

44. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW *et al*: **The developmental transcriptome of *Drosophila melanogaster***. *Nature* 2011, **471**(7339):473-479.

45. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. *Nat. Rev. Genet.* 2011, **12**(2):87-98.

46. Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH: **Transcriptomic analysis of autistic brain reveals convergent molecular pathology**. *Nature* 2011, **474**(7351):380-384.

47. Lin M, Pedrosa E, Shah A, Hrabovsky A, Maqbool S, Zheng D, Lachman HM: **RNA-Seq of human neurons derived from iPS cells reveals candidate long non-coding RNAs involved in neurogenesis and neuropsychiatric disorders**. *PLoS One* 2011, **6**(9):e23356.

48. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nature reviews. Genetics* 2009, **10**(1):57-63.

49. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer**. *Nature* 2009, **458**(7234):97-101.

50. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH: **Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans***. *Genome Res.* 2009, **19**(4):657-666.

51. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**(7221):470-476.

52. Agarwal A, Koppstein D, Rozowsky J, Sboner A, Habegger L, Hillier LW, Sasidharan R, Reinke V, Waterston RH, Gerstein M: **Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays**. *BMC Genomics* 2010, **11**:383.

53. Wyman SK, Parkin RK, Mitchell PS, Fritz BR, O'Briant K, Godwin AK, Urban N, Drescher CW, Knudsen BS, Tewari M: **Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries**. *PLoS One* 2009, **4**(4):e5311.

54. Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK: **Phylotranscriptomics to Bring the Understudied into the Fold: Monophyletic Ostracoda, Fossil Placement, and Pancrustacean Phylogeny**. *Mol. Biol. Evol.* 2013, **30**(1):215-233.

55. Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung HJ, Vance CP, Gronwald JW: **Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems**. *BMC Genomics* 2011, **12**:199.

56. Miller HC, Biggs PJ, Voelckel C, Nelson NJ: **De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*)**. *BMC Genomics* 2012, **13**:439.

57. Andrews S: **A quality control tool for high throughput sequence data.** 2010.

58. Patel RK, Jain M: **NGS QC Toolkit: a toolkit for quality control of next generation sequencing data**. *PLoS One* 2012, **7**(2):e30619.

59. Lindgreen S: **AdapterRemoval: easy cleaning of next-generation sequencing reads**. *BMC Res. Notes* 2012, **5**:337.

60. Martin JA, Wang Z: **Next-generation transcriptome assembly**. *Nat. Rev. Genet.* 2011, **12**(10):671-682.

61. Blanca JM, Pascual L, Ziarsolo P, Nuez F, Cañizares J: **Ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence**. *BMC Genomics* 2011, **12**(1):285.

62. Blanquer IB, Brasche G, Cala J, Gagliardi F, Gannon D, Hiden H, Soncu H, Takeda K, Tomás A, Woodman S: **Supporting NGS pipelines in the cloud**. *EMBnet. journal* 2013, **19**(A):pp. 14-16.

63. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C *et al*: **Annotating genomes with massive-scale RNA sequencing**. *Genome Biol.* 2008, **9**(12):R175.

64. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C *et al*: **Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs.** *Nat Biotech* 2010, **28**(5):503-510.

65. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat. Biotechnol.* 2010, **28**(5):511-515.

66. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome Res.* 2009, **19**(6):1117-1123.

67. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res.* 2008, **18**(5):821-829.

68. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y: **SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler**. *GigaScience* 2012, **1**(1):18.

69. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels**. *Bioinformatics* 2012, **28**(8):1086-1092.

70. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al*: **De novo assembly and analysis of RNA-seq data**. *Nat. Meth.* 2010, **7**(11):909-912.

71. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S: **SOAPdenovo-Trans: De novo transcriptome assembly with short RNA-Seq reads**. *arXiv preprint arXiv:1305.6760* 2013.

72. Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P: **Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study**. *BMC Bioinformatics* 2011, **12 Suppl 14**:S2.

73. Rismani-Yazdi H, Haznedaroglu B, Bibby K, Peccia J: **Transcriptome sequencing and annotation of the microalgae Dunaliella tertiolecta: Pathway**

description and gene discovery for production of next-generation biofuels. *BMC Genomics* 2011, **12**(1):148.

74. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotech* 2011, **29**(7):644-652.

75. Iyer MK, Chinnaiyan AM: **RNA-Seq unleashed**. *Nat. Biotechnol.* 2011, **29**(7):599-600.

76. Jeffrey Martin, M V, Zhide B, Meng FX, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z: **Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads**. *BMC Genomics* 2010, **11**:663.

77. Sacomoto GA, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, Peterlongo P, Lacroix V: **KISSPLICE: de-novo calling alternative splicing events from RNA-seq data**. *BMC Bioinformatics* 2012, **13**(Suppl 6):S5.

78. Fu L, Niu B, Zhu Z, Wu S, Li W: **CD-HIT: accelerated for clustering the next-generation sequencing data**. *Bioinformatics* 2012, **28**(23):3150-3152.

79. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17**(suppl 1):S140-S148.

80. Zhaxybayeva O, Gogarten JP: **Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses**. *BMC Genomics* 2002, **3**:4.

81. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet.* 2000, **16**(6):276-277.

82. Enright A, Van Dongen S, Ouzounis C: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res.* 2002, **30**(7):1575-1584.

83. Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM, DeSalle R: **OrthologID: automation of genome-scale ortholog identification within a parsimony framework**. *Bioinformatics* 2006, **22**(6):699-707.

84. Harris M, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res.* 2004, **32**(Database issue):D258-261.

85. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res.* 2000, **28**(1):45-48.

86. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res.* 2007, **35**(suppl 1):D61-D65.

87. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**(9):847-848.

88. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

89. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2010, **26**(5):589-595.

90. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25.

91. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat. Meth.* 2012, **9**(4):357-359.

92. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores**. *Genome Res.* 2008, **18**(11):1851-1858.

93. Ning Z, Cox AJ, Mullikin JC: **SSAHA: a fast search method for large DNA databases**. *Genome Res.* 2001, **11**(10):1725-1729.

94. Ramirez-Gonzalez RH, Bonnal R, Caccamo M, Maclean D: **Bio-samtools: Ruby bindings for SAMtools, a library for accessing BAM files containing high-throughput sequence alignments**. *Source Code Biol. Med.* 2012, **7**(1):6.

95. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Brief. Bioinform.* 2013, **14**(2):178-192.

96. Carver T, Böhme U, Otto TD, Parkhill J, Berriman M: **BamView: viewing mapped read alignment data in the context of the reference sequence**. *Bioinformatics* 2010, **26**(5):676-677.

97. Karolchik D, Hinrichs AS, Kent WJ: **The UCSC Genome Browser**. *Curr Protoc Hum Genet* 2011, **Chapter 18**:Unit18 16.

98. Hatem A, Bozda D, Toland AE, Çatalyürek ÜV: **Benchmarking short sequence mapping tools**. *BMC Bioinformatics* 2013, **14**(1):184.

Chapter 2 Bibliography

1. Metzker ML: **Sequencing technologies—the next generation.** *Nat Rev Gen* 2009, **11**(1)**:**31–46.

2. Koepke T, Schaeffer S, Krishnan V, Jiwan D, Harper A, Whiting M, Oraguzie N, Dhingra A: **Rapid gene-based SNP and haplotype marker development in non-model eukaryotes using 3'UTR sequencing.** *BMC Genomics* 2012, **13:**18.

3. Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S, Meyer A: **Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes.** *Mole Ecol* 2010, **19**(Suppl 1)**:**197–211.

4. Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung HJ, Vance CP, Gronwald JW: **Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems.** *BMC Genomics* 2011, **12:**199.

5. Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T, *et al*: **Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds.** *BMC Genomics* 2011, **12:**131.

6. Garg R, Patel RK, Tyagi AK, Jain M: **De novo assembly of chickpea transcriptome using short reads for gene discovery and marker identification.** *DNA Res* 2011, **18**(1)**:**53–63.

7. Miller HC, Biggs PJ, Voelckel C, Nelson NJ: **De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*).** *BMC Genomics* 2012, **13:**439.

8. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP: **De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology.** *PloS One* 2010, **5**(12)**:**e14202.

9. Pandolfi JM, Bradbury RH, Sala E, Hughes TP, Bjorndal KA, Cooke RG, McArdle D, McClenachan L, Newman MJ, Paredes G, *et al*: **Global trajectories of the long-term decline of coral reef ecosystems.** *Science* 2003, **301**(5635)**:**955–958.

10. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10:**219.

11. Meyer E, Aglyamova GV, Matz MV: **Profiling gene expression responses of coral larvae (*Acropora millepora*) to elevated temperature and settlement inducers using a novel RNA-Seq procedure.** *Mol Ecol* 2011, **20**(17)**:**3599–3616.

12. Traylor-Knowles N, Granger BR, Lubinski TJ, Parikh JR, Garamszegi S, Xia Y, Marto JA, Kaufman L, Finnerty JR: **Production of a reference transcriptome and transcriptomic database (PocilloporaBase) for the cauliflower coral. Pocillopora damicornis.** *BMC Genomics* 2011, **12:**585.

13. Vidal-Dupiol J, Zoccola D, Tambutte E, Grunau C, Cosseau C, Smith KM, Freitag M, Dheilly NM, Allemand D, Tambutte S: **Genes related to ion-transport and energy production are upregulated in response to CO2-driven pH decrease in corals: new insights from transcriptome analysis.** *PloS One* 2013, **8**(3)**:**e58652.

14. Moya A, Huisman L, Ball EE, Hayward DC, Grasso LC, Chua CM, Woo HN, Gattuso JP, Foret S, Miller DJ: **Whole transcriptome analysis of the coral Acropora millepora reveals complex responses to CO(2)-driven acidification during the initiation of calcification.** *Mol Ecol* 2012, **21**(10)**:**2440–2454.

15. Kitahara MV, Cairns SD, Stolarski J, Blair D, Miller DJ: **A comprehensive phylogenetic analysis of the Scleractinia (Cnidaria, Anthozoa) based on mitochondrial CO1 sequence data.** *PloS One* 2010, **5**(7)**:**e11490.

16. Huang D, Licuanan WY, Baird AH, Fukami H: **Cleaning up the 'Bigmessidae': molecular phylogeny of scleractinian corals from Faviidae, Merulinidae, Pectiniidae and Trachyphylliidae.** *BMC Evol Biol* 2011, **11:**37.

17. Veron J: *Corals World* 2000, **3:**100–131.

18. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19**(6)**:**1117–1123.

19. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N: **Reference-free transcriptome assembly in**

non-model animals from next-generation sequencing data. *Mol Ecol Resour* 2012, **12**(5)**:**834–845.

20. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, *et al*: **De novo assembly and analysis of RNA-seq data.** *Nat Methods* 2010, **7**(11)**:**909–912.

21. Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW: **Scaffolding a Caenorhabditis nematode genome with RNA-seq.** *Genome Res* 2010, **20**(12)**:**1740–1747.

22. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**(9)**:**868–877.

23. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y: **De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing.** *PloS one* 2012, **7**(8)**:**e42546.

24. Rice P, Longden I, Bleasby A: **EMBOSS: the European molecular biology open software suite.** *Trends Genet* 2000, **16**(6)**:**276–277.

25. Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, *et al*: **Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization.** *Science* 2007, **317**(5834)**:**86–94.

26. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL: **NCBI BLAST: a better web interface.** *Nucleic Acids Res* 2008, **36:**W5–W9. Web Server issue).

27. Enright A, Van Dongen S, Ouzounis C: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30**(7)**:**1575–1584.

28. Martin JA, Wang Z: **Next-generation transcriptome assembly.** *Nat Rev Gen* 2011, **12**(10)**:**671–682.

29. Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S: **De novo characterization of a whitefly transcriptome and analysis of its gene expression during development.** *BMC Genomics* 2010, **11**(1)**:**400.

30. Finney JC, Pettay DT, Sampayo EM, Warner ME, Oxenford HA, LaJeunesse TC: **The relative significance of host-habitat, depth, and geography on the ecology, endemism, and speciation of coral endosymbionts in the genus *Symbiodinium*.** *Microbial Ecol* 2010, **60**(1)**:**250–263.

31. Hagedorn M, Carter VL, Leong JC, Kleinhans FW: **Physiology and cryosensitivity of coral endosymbiotic algae (*Symbiodinium*).** *Cryobiology* 2010, **60**(2)**:**147–158.

32. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al*: **Clustal W and clustal X version 2.0.** *Bioinformatics* 2007, **23**(21)**:**2947–2948.

33. Patrick Kück KM: **FASconCAT: convenient handling of data matrices.** *Mole Phylogen Evol* 2010, **56**(2010)**:**1115–1118.

34. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.** *Bioinformatics* 2005, **21**(4)**:**456–463.

35. Budd AF, Romano SL, Smith ND, Barbeitos MS: **Rethinking the phylogeny of scleractinian corals: a review of morphological and molecular data.** *Integ Comp Biol* 2010, **50**(3)**:**411–427.

36. Fukami H, Budd AF, Paulay G, Solé-Cava A, Allen Chen C, Iwao K, Knowlton N: **Conventional taxonomy obscures deep divergence between Pacific and Atlantic corals.** *Nat Geosci* 2004, **427**(6977)**:**832–835.

37. Ilagan RP, Rhoades E, Gruber DF, Kao HT, Pieribone VA, Regan L: **A new bright green-emitting fluorescent protein–engineered monomeric and dimeric forms.** *FEBS J* 2010, **277**(8)**:**1967–1978.

38. Labas YA, Gurskaya NG, Yanushevich YG, Fradkov AF, Lukyanov KA, Lukyanov SA, Matz MV: **Diversity and evolution of the green fluorescent protein family.** *Proc Natl Acad Sci U S A* 2002, **99**(7)**:**4256–4261.

39. Matz MV, Fradkov AF, Labas YA, Savitsky AP, Zaraisky AG, Markelov ML, Lukyanov SA: **Fluorescent proteins from nonbioluminescent Anthozoa species.** *Nat Biotech* 1999, **17**(10)**:**969–973.

40. Bogdanov AM, Mishin AS, Yampolsky IV, Belousov VV, Chudakov DM, Subach FV, Verkhusha VV, Lukyanov S, Lukyanov KA: **Green fluorescent proteins are light-induced electron donors.** *Nature chemical biology* 2009, **5**(7)**:**459–461.

41. Sullivan JC, Ryan JF, Watson JA, Webb J, Mullikin JC, Rokhsar D, Finnerty JR: **StellaBase: the *Nematostella vectensis* genomics database.** *Nucleic acids res* 2006, **34**(Database issue)**:**D495–D499.

42. Bomati EK, Manning G, Deheyn DD: **Amphioxus encodes the largest known family of green fluorescent proteins, which have diversified into distinct functional classes.** *BMC evol bio* 2009, **9:**77.

43. Haddock SHD, Mastroianni N, Christianson LM: **A photoactivaTable C2. green-fluorescent protein from the phylum Ctenophora.** *Proc Royal Soc B Biol Sci* 2010, **277**(1685)**:**1155–1160.

44. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution.** *Bioinformatics* 2011, **27**(8)**:**1164–1165.

45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nature biotechnology* 2011, **29**(7)**:**644–652.

46. Yang F, Moss LG, Phillips GN Jr: *The molecular structure of green fluorescent protein.* Dept of biochemistry and cell biology: Rice University; 1997.

47. Ormö M, Cubitt AB, Kallio K, Gross LA, Tsien RY, Remington SJ: **Crystal structure of the *Aequorea victoria* green fluorescent protein.** *Science* 1996, **273**(5280)**:**1392–1395.

48. Kozak M: **An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.** *Nucleic Acids Res* 1987, **15**(20)**:**8125–8148.

49. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotech* 2010, **28**(5)**:**511–515.

50. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5**(7)**:**621–628.

51. Gruber D, Kao H, Janoschka S, Tsai J, Pieribone V: **Patterns of fluorescent protein expression in scleractinian corals.** *Biol Bull* 2008, **215**(2)**:**143.

52. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Federhen S, *et al*: **Database resources of the national center for biotechnology information.** *Nucleic Acids Res* 2010, **38**(Database issue)**:**D5–D16.

53. Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, *et al*: **Using the *Acropora digitifera* genome to**

understand coral responses to environmental change. *Nat Geosci* 2011, **476**(7360)**:**320–323.

54. Katoh K, Asimenos G, Toh H: **Multiple alignment of DNA sequences with MAFFT.** *Methods Mole Biol (Clifton, NJ)* 2009, **537:**39–64.

55. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R: **Dendroscope: An interactive viewer for large phylogenetic trees.** *BMC Bioinforma* 2007, **8**(1)**:**460.

56. Sambrook J, Fritsch EF, Maniatis T: *Molecular cloning, vol. 2.* New York: Cold spring harbor laboratory press; 1989.

57. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3)**:**R25.

58. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14)**:**1754–1760.

59. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools.** *Bioinformatics* 2009, **25**(16)**:**2078–2079.

60. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer.** *Nat Biotech* 2011, **29**(1)**:**24–26.

Chapter 3 Bibliography

1. Metzker ML: **Sequencing technologies—the next generation**. *Nat. Rev. Genet.* 2009, **11**(1):31-46.
2. Dunn CW, Hejnol A, Matus DQ, Pang K, Browne WE, Smith SA, Seaver E, Rouse GW, Obst M, Edgecombe GD *et al*: **Broad phylogenomic sampling improves resolution of the animal tree of life**. *Nature* 2008, **452**(7188):745-749.
3. Feng C, Chen M, Xu CJ, Bai L, Yin XR, Li X, Allan AC, Ferguson IB, Chen KS: **Transcriptomic analysis of Chinese bayberry (*Myrica rubra*) fruit development and ripening using RNA-Seq**. *BMC Genomics* 2012, **13**:19.
4. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y: **De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing**. *PLoS One* 2012, **7**(8):e42546.
5. Shi CY, Yang H, Wei CL, Yu O, Zhang ZZ, Jiang CJ, Sun J, Li YY, Chen Q, Xia T *et al*: **Deep sequencing of the *Camellia sinensis* transcriptome revealed candidate genes for major metabolic pathways of tea-specific compounds**. *BMC Genomics* 2011, **12**:131.
6. Crawford JE, Guelbeogo WM, Sanou A, Traore A, Vernick KD, Sagnon N, Lazzaro BP: **De novo transcriptome sequencing in *Anopheles funestus* using Illumina RNA-seq technology**. *PLoS One* 2010, **5**(12):e14202.

7. Franchini P, Van der Merwe M, Roodt-Wilding R: **Transcriptome characterization of the South African abalone *Haliotis midae* using sequencing-by-synthesis**. *BMC Res. Notes* 2011, **4**(1):59.

8. Salem M, Vallejo RL, Leeds TD, Palti Y, Liu S, Sabbagh A, Rexroad CE, 3rd, Yao J: **RNA-Seq identifies SNP markers for growth traits in rainbow trout**. *PLoS One* 2012, **7**(5):e36264.

9. Renaut S, Nolte AW, Rogers SM, Derome N, Bernatchez L: **SNP signatures of selection on standing genetic variation and their association with adaptive phenotypes along gradients of ecological speciation in lake whitefish species pairs (*Coregonus spp.)***. *Mol. Ecol.* 2011, **20**(3):545-559.

10. Yang SS, Tu ZJ, Cheung F, Xu WW, Lamb JF, Jung H-JG, Vance CP, Gronwald JW: **Using RNA-Seq for gene identification, polymorphism detection and transcript profiling in two alfalfa genotypes with divergent cell wall composition in stems**. *BMC Genomics* 2011, **12**(1):199.

11. Canovas A, Rincon G, Islas-Trejo A, Wickramasinghe S, Medrano JF: **SNP discovery in the bovine milk transcriptome using RNA-Seq technology**. *Mamm. Genome* 2010, **21**(11-12):592-598.

12. Ahrens JB, Borda E, Barroso R, Paiva PC, Campbell AM, Wolf A, Nugues MM, Rouse GW, Schulze A: **The curious case of *Hermodice carunculata* (Annelida: Amphinomidae): evidence for genetic homogeneity throughout the Atlantic Ocean and adjacent basins**. *Mol. Ecol.* 2013, **22**(8):2280-2291.

13. Sebens KP: **Intertidal distribution of zoanthids on the Caribbean coast of Panama: effects of predation and desiccation**. *Bull. Mar. Sci.* 1982, **32**(1):316-335.

14. Karlson RH: **Disturbance and monopolization of a spatial resource by *Zoanthus sociatus* (Coelenterata, Anthozoa)**. *Bull. Mar. Sci.* 1983, **33**(1):118-131.

15. Ott B, Lewis JB: **The importance of the gastropod *Coralliophila abbreviata* (Lamarck) and the polychaete *Hermodice carunculata* (Pallas) as coral reef predators**. *Can. J. Zool.* 1972, **50**(12):1651-1656.

16. Rylaarsdam KW: **Life histories and abundance patterns of colonial corals on Jamaican reefs**. *Marine ecology progress series. Oldendorf* 1983, **13**(2):249-260.

17. Wolf AT, Nugues MM: **Predation on coral settlers by the corallivorous fireworm *Hermodice carunculata***. *Coral Reefs* 2012.

18. Marsden JR: **The digestive tract of *Hermodice carunculata* (Pallas). Polychaeta: Amphinomidae**. *Can. J. Zool.* 1963, **41**(2):165-184.

19. B.J. L, R.E. C: **Foraging Cycles of the Amphinomid Polychaete *Hermodice Carunculata* Preying on the Calcareous Hydrozoan *Millepora Complanata***. *Bull. Mar. Sci.* 1996, **53**(3):853-856.

20. Fauchald K, Jumars PA: **The diet of worms: a study of polychaete feeding guilds**. 1979.

21. Sussman M, Loya Y, Fine M, Rosenberg E: **The marine fireworm *Hermodice carunculata* is a winter reservoir and spring‐summer vector for the coral‐bleaching pathogen *Vibrio shiloi***. *Environ. Microbiol.* 2003, **5**(4):250-255.

22. Wiklund H, Nygren A, Pleijel F, Sundberg P: **The phylogenetic relationships between Amphinomidae, Archinomidae and Euphrosinidae (Amphinomida: Aciculata: Polychaeta), inferred from molecular data**. *Journal of the Marine Biological Association of the UK* 2008, **88**(03):509-513.

23. Rouse G, Pleijel F: **Polychaetes**: Oxford University Press; 2001.

24. Borda E, Kudenov JD, Bienhold C, Rouse GW: **Towards a revised Amphinomidae (Annelida, Amphinomida): description and affinities of a new genus and species from the Nile Deep‐sea Fan, Mediterranean Sea**. *Zool. Scr.* 2012, **41**(3):307-325.

25. Rouse GW, Fauchald K: **Cladistics and polychaetes**. *Zool. Scr.* 1997, **26**(2):139-204.

26. Yáñez-Rivera B, Salazar-Vallejo SI: **Revision of *Hermodice Kinberg*, 1857 (Polychaeta: Amphinomidae)**. *Sci. Mar.* 2011, **75**(2):251-262.

27. Colgan DJ, Hutchings PA, Beacham E: **Multi-gene analyses of the phylogenetic relationships among the Mollusca, Annelida, and Arthropoda**. *Zool Sci* 2008, **47**:338-351.

28. Halanych KM, Janosik AM: **A review of molecular markers used for Annelid phylogenetics**. *Integrative and Comparative Biology* 2006, **46**(4):533-543.

29. Barroso R, Paiva PC: **Amphinomidae (Annelida: Polychaeta) from Rocas Atoll, Northeastern Brazil**. *Arq. Mus. Nac* 2007, **65**(3):357-362.

30. Kitahara MV, Cairns SD, Stolarski J, Blair D, Miller DJ: **A comprehensive phylogenetic analysis of the Scleractinia (Cnidaria, Anthozoa) based on mitochondrial CO1 sequence data**. *PLoS One* 2010, **5**(7):e11490.

31. Koenemann S, Jenner RA, Hoenemann M, Stemme T, von Reumont BM: **Arthropod phylogeny revisited, with a focus on crustacean relationships**. *Arthropod. Struct. Dev.* 2010, **39**(2):88-110.

32. Dayrat B, Conrad M, Balayan S, White TR, Albrecht C, Golding R, Gomes SR, Harasewych M, de Frias Martins AM: **Phylogenetic relationships and evolution of pulmonate gastropods (Mollusca): new insights from increased taxon sampling**. *Mol. Phylogenet. Evol.* 2011, **59**(2):425-437.

33. Kojima S: **Paraphyletic status of Polychaeta suggested by phylogenetic analysis based on the amino acid sequences of elongation factor-1 alpha**. *Mol. Phylogenet. Evol.* 1998, **9**(2):255-261.

34. Brown S, Rouse G, Hutchings P, Colgan D: **Assessing the usefulness of histone H3, U2 snRNA and 28S rDNA in analyses of polychaete relationships**. *Aust. J. Zool.* 1999, **47**(5):499-516.

35. Burnette AB, Struck TH, Halanych KM: **Holopelagic Poeobius meseres ("Poeobiidae," Annelida) is derived from benthic flabelligerid worms**. *Biol. Bull.* 2005, **208**(3):213-220.

36. Carr CM, Hardy SM, Brown TM, Macdonald TA, Hebert PD: **A tri-oceanic perspective: DNA barcoding reveals geographic structure and cryptic diversity in Canadian polychaetes**. *PLoS One* 2011, **6**(7):e22232.

37. Nygren A, Eklöf J, Pleijel F: **Arctic-boreal sibling species of Paranaitis (Polychaeta, Phyllodocidae)**. *Mar. Biol. Res.* 2009, **5**(4):315-327.

38. Sampertegui S, Rozbaczylo N, Canales-Aguirre CB, Carrasco F, Hernandez CE, Rodriguez-Serrano E: **Morphological and molecular characterization of**

Perinereis gualpensis (Polychaeta: Nereididae) and its phylogenetic relationships with other species of the genus off the Chilean coast, Southeast Pacific. *Cah. Biol. Mar* 2013, **54**:27-40.

39. Westheide W, Schmidt H: **Cosmopolitan versus cryptic meiofaunal polychaete species: an approach to a molecular taxonomy**. *Helgol. Mar. Res.* 2003, **57**(1):1-6.

40. Schmidt H, Westheide W: **Are the meiofaunal polychaetes Hesionides arenaria and Stygocapitella subterranea true cosmopolitan species?−results of RAPD‐PCR investigations**. *Zool. Scr.* 2000, **29**(1):17-27.

41. Novo M, Riesgo A, Fernández-Guerra A, Giribet G: **Pheromone evolution, reproductive genes, and comparative transcriptomics in Mediterranean earthworms (Annelida, Oligochaeta, Hormogastridae)**. *Mol. Biol. Evol.* 2013, **30**(7):1614-1629.

42. Hastings J: **Biological diversity, chemical mechanisms, and the evolutionary origins of bioluminescent systems**. *J. Mol. Evol.* 1983, **19**(5):309-321.

43. Petushkov V, Rodionova N, Purtov K, Bondar' V: **The Luminescence System of Soil Enchytraeids, Henlea sp.,(Annelida: Clitellata: Oligochaeta: Enchytraeidae)**. In: *Doklady Biological Sciences: 2002*: Springer; 2002: 310-312.

44. Zörner S, Fischer A: **The spatial pattern of bioluminescent flashes in the polychaete *Eusyllis blomstrandi* (Annelida)**. *Helgol. Mar. Res.* 2007, **61**(1):55-66.

45. Fischer A, Fischer U: **On the life-style and life-cycle of the luminescent polychaete *Odontosyllis enopla* (Annelida: Polychaeta)**. *Invertebr. Biol.* 1995:236-247.

46. Trauth K: **Night ecology and fluorescence of the fireworm, *Hermodice carunculata***. *Physis* 2007, **2**:3-8.

47. Andrews S: **A quality control tool for high throughput sequence data.** 2010.

48. Swaminathan K, Chae WB, Mitros T, Varala K, Xie L, Barling A, Glowacka K, Hall M, Jezowski S, Ming R: **A framework genetic map for Miscanthus sinensis from RNAseq-based markers shows recent tetraploidy**. *BMC Genomics* 2012, **13**(1):142.

49. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data**. *Genome Res.* 2009, **19**(6):1117-1123.

50. Kent WJ: **BLAT--the BLAST-like alignment tool**. *Genome Res.* 2002, **12**(4):656-664.

51. Surget-Groba Y, Montoya-Burgos JI: **Optimization of de novo transcriptome assembly from next-generation sequencing data**. *Genome Res.* 2010, **20**(10):1432-1440.

52. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet.* 2000, **16**(6):276-277.

53. Altschul S: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res.* 1997, **25**(17):3389-3402.

54. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis**. *Brief. Bioinform.* 2007, **8**(1):6-21.

55. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: **Gene Ontology: tool for the unification of biology**. *Nat. Genet.* 2000, **25**(1):25-29.

56. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res.* 2004, **32**(Database issue):D258-261.

57. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**(9):847-848.

58. Mulder NJ, Apweiler R: **The InterPro database and tools for protein domain analysis**. *Curr Protoc Bioinformatics* 2008, **Chapter 2**:Unit 2 7.

59. Hofmann K, Bucher P, Falquet L, Bairoch A: **The PROSITE database, its status in 1999**. *Nucleic Acids Res.* 1999, **27**(1):215-219.

60. Attwood TK, Croning MDR, Flower DR, Lewis A, Mabey J, Scordis P, Selley J, Wright W: **PRINTS-S: the database formerly known as PRINTS**. *Nucleic Acids Res.* 2000, **28**(1):225-227.

61. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths‐Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res.* 2004, **32**(suppl 1):D138-D141.

62. Letunic I, Copley RR, Schmidt S, Ciccarelli FD, Doerks T, Schultz J, Ponting CP, Bork P: **SMART 4.0: towards genomic data integration**. *Nucleic Acids Res.* 2004, **32**(suppl 1):D142-D144.

63. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

64. Conesa A, Götz S: **Blast2GO: A comprehensive suite for functional analysis in plant genomics**. *Int J Plant Genomics* 2008, **2008**.

65. Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, Hoeksema BW, Poore GC, van Soest RW, Stohr S, Walter TC *et al*: **Global Coordination and Standardisation in Marine Biodiversity through the World Register of Marine Species (WoRMS) and Related Databases**. *PLoS One* 2013, **8**(1):e51629.

66. Watson G, Langford F, Gaudron S, Bentley M: **Factors influencing spawning and pairing in the scale worm Harmothoe imbricata (Annelida: Polychaeta)**. *The Biological Bulletin* 2000, **199**(1):50-58.

67. Zeeck E, Hardege J, Bartels-Hardege H: **Pla tyn ereis d urn erilii**. *Mar. Ecol. Prog. Ser* 1990, **67**:183-188.

68. Xia N-S, Luo W-X, Zhang J, Xie X-Y, Yang H-J, Li S-W, Chen M, Ng M-H: **Bioluminescence of Aequorea macrodactyla, a common jellyfish species in the East China Sea**. *Mar. Biotechnol.* 2002, **4**(2):155-162.

69. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol. Biol. Evol.* 2000, **17**(4):540-552.

70. Gouy M, Guindon S, Gascuel O: **SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building**. *Mol. Biol. Evol.* 2010, **27**(2):221-224.

71. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models**. *Bioinformatics* 2006, **22**(21):2688-2690.

Chapter 4 Bibliography

1. Shu X, Royant A, Lin MZ, Aguilera TA, Lev-Ram V, Steinbach PA, Tsien RY: **Mammalian expression of infrared fluorescent proteins engineered from a bacterial phytochrome**. *Science* 2009, **324**(5928):804-807.
2. Filonov GS, Piatkevich KD, Ting L-M, Zhang J, Kim K, Verkhusha VV: **Bright and stable near-infrared fluorescent protein for in vivo imaging**. *Nat. Biotechnol.* 2011, **29**(8):757-761.
3. Auldridge ME, Satyshur KA, Anstrom DM, Forest KT: **Structure-guided engineering enhances a phytochrome-based infrared fluorescent protein**. *J. Biol. Chem.* 2012, **287**(10):7000-7009.
4. Shimomura M, Kunitake T: **Fluorescence and photoisomerization of azobenzene-containing bilayer membranes**. *J. Am. Chem. Soc.* 1987, **109**(17):5175-5183.
5. Pieribone VA, Gruber DF: **Aglow in the dark: the revolutionary science of biofluorescence**: Harvard University Press; 2005.
6. Alieva NO, Konzen KA, Field SF, Meleshkevitch EA, Hunt ME, Beltran-Ramirez V, Miller DJ, Wiedenmann J, Salih A, Matz MV: **Diversity and evolution of coral fluorescent proteins**. *PLoS One* 2008, **3**(7):e2680.
7. Gruber DF, Desalle R, Lienau EK, Tchernov D, Pieribone VA, Kao H-T: **Novel internal regions of fluorescent proteins undergo divergent evolutionary patterns.** *Mol. Biol. Evol.* 2009, **26**(12):2841-2848.

8. Gruber D, Kao H, Janoschka S, Tsai J, Pieribone V: **Patterns of fluorescent protein expression in scleractinian corals**. *The Biological Bulletin* 2008, **215**(2):143.

9. Matz MV, Fradkov AF, Labas YA, Savitsky AP, Zaraisky AG, Markelov ML, Lukyanov SA: **Fluorescent proteins from nonbioluminescent Anthozoa species**. *Nat Biotech* 1999, **17**(10):969-973.

10. Leutenegger A, D'Angelo C, Matz MV, Denzel A, Oswald F, Salih A, Nienhaus GU, Wiedenmann J: **It's cheap to be colorful**. *FEBS J.* 2007, **274**(10):2496-2505.

11. Hunt ME, Scherrer MP, Ferrari FD, Matz MV: **Very Bright Green Fluorescent Proteins from the Pontellid Copepod *Pontella mimocerami***. *PLoS One* 2010, **5**(7):e11517.

12. Bomati EK, Manning G, Deheyn DD: **Amphioxus encodes the largest known family of green fluorescent proteins, which have diversified into distinct functional classes**. *BMC Evol. Biol.* 2009, **9**:77.

13. Haddock SHD, Mastroianni N, Christianson LM: **A photoactivatable green-fluorescent protein from the phylum Ctenophora**. *Proceedings of the Royal Society B: Biological Sciences* 2010, **277**(1685):1155-1160.

14. Hayashi S, Toda Y: **A novel fluorescent protein purified from eel muscle**. *Fish. Sci.* 2009, **75**(6):1461-1469.

15. Kumagai A, Ando R, Miyatake H, Greimel P, Kobayashi T, Hirabayashi Y, Shimogori T, Miyawaki A: **A Bilirubin-Inducible fluorescent protein from eel muscle**. *Cell* 2013.

16. WEITZMAN SH, Myers G: **Phyletic studies of teleostean fishes, with a provisional classification of living forms**. *Bull. Amer. Mus. Nat. Hist* 1966, **131**:341-455.

17. Tesch FW, White RJ: **The eel**: Wiley. com; 2008.

18. Inoue JG, Miya M, Miller MJ, Sado T, Hanel R, Hatooka K, Aoyama J, Minegishi Y, Nishida M, Tsukamoto K: **Deep-ocean origin of the freshwater eels**. *Biol. Lett.* 2010, **6**(3):363-366.

19. Garber M, Grabherr MG, Guttman M, Trapnell C: **Computational methods for transcriptome annotation and quantification using RNA-seq**. *Nat. Methods* 2011, **8**(6):469-477.

20. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al*: **De novo assembly and analysis of RNA-seq data**. *Nat. Meth.* 2010, **7**(11):909-912.

21. Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U: **454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases.** *The New phytologist* 2010, **188**(1):291-301.

22. Coppe A, Pujolar J, Maes G, Larsen P, Hansen M, Bernatchez L, Zane L, Bortoluzzi S: **Sequencing, de novo annotation and analysis of the first *Anguilla anguilla* transcriptome: EeelBase opens new perspectives for the study of the critically endangered European eel**. *BMC Genomics* 2010, **11**(1):635.

23. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat. Meth.* 2012, **9**(4):357-359.

24. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet.* 2000, **16**(6):276-277.

25. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL: **NCBI BLAST: a better web interface**. *Nucleic Acids Res.* 2008, **36**(Web Server issue):W5-9.

26. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.

27. Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y: **De novo sequencing and transcriptome analysis of the central nervous system of mollusc *Lymnaea stagnalis* by deep RNA sequencing**. *PLoS One* 2012, **7**(8):e42546.

28. Mehr SF, DeSalle R, Kao H-T, Narechania A, Han Z, Tchernov D, Pieribone V, Gruber DF: **Transcriptome deep-sequencing and clustering of expressed isoforms from *Favia* corals**. *BMC Genomics* 2013, **14**:546.

29. Gao J, Gu H, Xu B: **Multifunctional magnetic nanoparticles: design, synthesis, and biomedical applications**. *Acc. Chem. Res.* 2009, **42**(8):1097-1107.

30. Shaner NC, Steinbach PA, Tsien RY: **A guide to choosing fluorescent proteins**. *Nat. Methods* 2005, **2**(12):905-909.

31. Tsien RY: **The green fluorescent protein**. *Annu. Rev. Biochem.* 1998, **67**(1):509-544.

32. Karasawa S, Araki T, Yamamoto-Hino M, Miyawaki A: **A green-emitting fluorescent protein from Galaxeidae coral and its monomeric version for use in fluorescent labeling**. *J. Biol. Chem.* 2003, **278**(36):34167-34171.

33. Stocker R, Yamamoto Y, McDonagh AF, Glazer AN, Ames BN: **Bilirubin is an antioxidant of possible physiological importance**. *Science* 1987, **235**(4792):1043-1046.

34. Maines MD: **New insights into biliverdin reductase functions: linking heme metabolism to cell signaling**. *Physiology* 2005, **20**(6):382-389.