1-2020

Cloning and Expression of the Green Fluorescent Protein (GvFP) from the Clinging Jellyfish (Gonionemus vertens) Using RNA-Seq

Vasilios A. Orologas
Montclair State University

Abstract

We have discovered a novel GFP in the clinging jellyfish (*Gonionemus vertens*), a protein we have named GvFP (*Gonionemus vertens* Fluorescent Protein). This protein is responsible for the bright green fluorescence seen in the rhopalia along the margin of the bell, the gonads associated with the radial canals, and the manubrium when illuminated with blue light (470 nm). GvFP was first identified by Nextgen sequencing (RNA-Seq) of libraries made from mRNA of mature medusa collected from a salt pond on Martha's Vineyard, MA. The full-length GvFP gene is 690 bp, encoding a protein of 230 AA (25.95 kDa; pI = 5.74). This protein shows low homology (43%) to the classic wtGFP from *Aequorea victoria,* but higher homology (86%) to another GFP (ScSuFP) from the Hydrozoan *Scolionema suvaence*. The tripeptide chromophore in GvFP - MYG - is unique among hydrozoans and has only been reported once before in *Anemonia sulcata* (Class Anthozoa). PHYRE2 analysis of GvFP predicts a β-barrel conformation and a 2° structure of 55% β-sheet, 7% alpha-helix, and 13% disordered structure, modelled at >90% confidence. The dimensions of the protein are 6.8 nm x 5.4 nm x 3.8 nm. Cloning of the full-length GvFP (Contig 6770) into the pET-SUMO expression vector and transformation into *E. coli* (BL21(DE3)) has expressed a recombinant GvFP that is a near perfect match in excitation and emission maxima of the native *in vivo* GvFP. Future work on this novel GFP may shed light on its unique chromophore and further our understanding of the ecological consequences of biological fluorescence in this jellyfish.

CLONING AND EXPRESSION OF THE GREEN FLUORESCENT PROTEIN (GvFP) FROM THE CLINGING JELLYFISH (*GONIONEMUS VERTENS*) USING RNA-SEQ

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTERS OF SCIENCE

BY

VASILIOS A. OROLOGAS

MONTCLAIR STATE UNIVERSITY
MONTCLAIR, NJ
January 2020

College of Science and Mathematics
Biology

Thesis Committee:

John J. Gaynor, Ph.D.
Thesis Sponsor

Paul A.X. Bologna, Ph.D.
Committee Member

Vladislav Snitsarev, Ph.D.
Committee Member

CLONING AND EXPRESSION OF THE GREEN FLUORESCENT PROTEIN (GvFP) FROM THE CLINGING JELLYFISH (*GONIONEMUS VERTENS*) USING RNA-SEQ

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

BY
VASILIOS A. OROLOGAS

MONTCLAIR STATE UNIVERSITY

MONTCLAIR, NJ

2020

Acknowledgements

I would like to thank Dr. John Gaynor for allowing me to work (and make plenty of mistakes) in his lab as a graduate research assistant during my time at Montclair State University. Without his expertise, patience and mentorship; inside and outside of the lab, this thesis would not be possible. Thank you. I would also like to thank my committee members: Dr. Paul Bologna and Dr. Vladislav Snitsarev for their input and assistance throughout this project, as well as my lab mates: Alorah Bliese and Elias Chalet who have made the failures of research as bearable as possible. I must also thank Adam Parker for his help with genetic sequencing for this project and many others. Lastly, I would like to thank my parents who have supported me throughout my academic career and continue to do so in all aspects of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

INTRODUCTION

THE DISCOVERY AND DEVELOPMENT OF THE GREEN FLUORESCENT PROTEIN

The green fluorescent protein (GFP) was discovered by the Japanese scientist Osamu Shimomura in 1961, from the jellyfish *Aequorea victoria* (Figure 1) as a companion protein to aequorin (Shimomura *et al*., 1962; Tsien, 1998). Today the green fluorescent protein serves as an invaluable tool in biomedical research. In the summer of 1961 Shimomura and the head of his lab at Princeton University traveled to Friday Harbor, Washington to collect *Aequorea* samples in order to study their light emitting properties. Upon arrival back to Princeton, multiple attempts were made to extract the luminescent molecule that caused a green glow in *Aequorea.* These attempts utilized the same methods used to extract the bioluminescent photoproteins luciferin and luciferase from fireflies. After various attempts of isolating the substance responsible for the luminescence of *Aequorea,* Shimomura turned away from his lab experiments and began to work independently on his own experiments. Shimomura postulated that a protein may be responsible for *Aequorea* luminescence.

This theory led Shimomura to test the presence of a possible luminescent protein by manipulating pH level in order to inhibit and activate luminescence. Shimomura cut off the bell margin of *Aequorea* and exposed the tissue to various pH levels. At pH 7, 6, and 5 luminescence was visible, however at a pH of 4, luminescence was not visible. When tissue exposed to pH 4 buffer was exposed to sodium bicarbonate (~pH 8.4) luminescence was visible again. After pouring cell-free *Aequorea* extract in a pH 4 buffer, into a sink that had seawater runoff luminescence was visible. Shimomura speculated that some component in seawater was responsible for the production of light. Eventually, in 1971, Shimomura discovered the

chromophore sequence of GFP using papain digestion (Shimomura, 2005). It would take another 21 years - and the advent of molecular cloning - to eventually isolate the gene from this jellyfish.

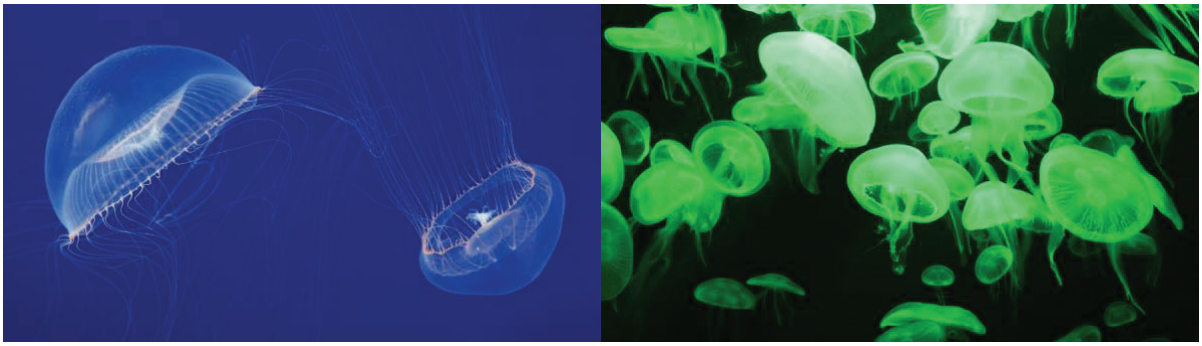


Figure 1. Left: *Aequorea victoria* under normal light conditions. Right: *Aequorea victoria* under blue light illumination displaying green fluorescence. (Left photo source: www.montereybayaquarium.org, right photo source: www.universityofcalifornia.edu)

Douglas Prasher was the first to clone a cDNA encoding the green fluorescent protein from *A. victoria* (Prasher *et al.*, 1992). Using this GFP sequence discovered by Prasher, Martin Chalfie was the first to create transgenic *Escherichia coli* using a construct that would express the intron-free GFP gene, downstream of the highly active T7 promoter, in order to achieve a high GFP production rate (Chalfie, 1994). Chalfie's study made four major claims in regard to the uses of this fluorescent protein in research. First, a strong green fluorescence is visible when GFP is excited by blue or UV light in cells. Second, GFP synthesis did not impact cell growth and/or death. Third, the excitation and emission wavelengths of crude GFP and GFP synthesized in *E. coli* are nearly identical. And fourth, GFP fluorescence does not require the addition of jellyfish specific substrates or cofactors (Chalfie, 1994). Chalfie's lab took it a step further and attempted to use wild-type GFP (wtGFP) to identify the localization of the *Mec7* gene product in the

transparent nematode, *C. elegans*, making his labs attempt the first to use wtGFP as a fluorescent marker to create a fusion wtGFP-Mec7 protein that would display fluorescence only in cells that expressed Mec7. The transgenic *C. elegans* created by Chalfie's lab can be seen in Figure 2.

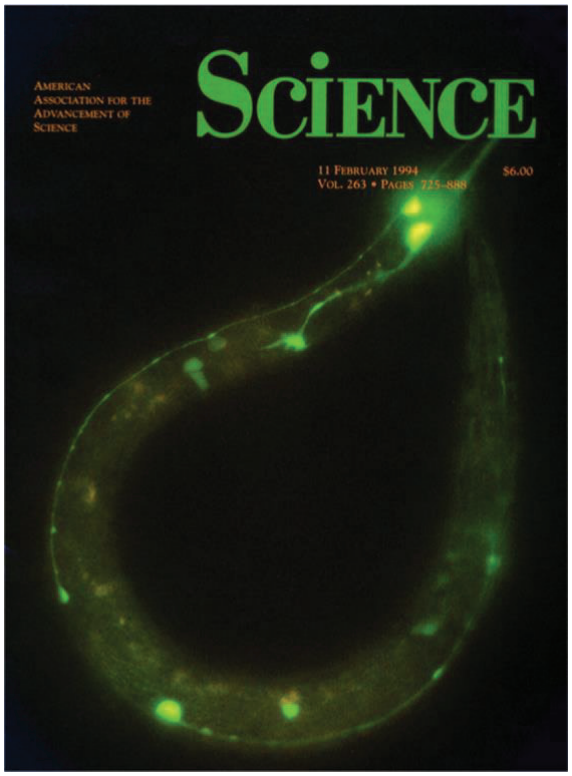


Figure 2: The February 11th, 1994 cover of *Science* displaying Chalfie's transgenic *C. elegans* expressing a fusion of wild-type *Aequorea victoria* GFP protein with the Mec-7 protein only expressed in neurons that also express the *Mec-7* gene.

Also using Prasher's GFP cDNA, the first ever GFP mutants were engineered. BFP (blue fluorescent protein) was created by Heim, Tsien and Prasher (Heim *et al.*, 1994) using random mutagenesis, transfection into *E. coli*, followed by colony screening and DNA sequencing of selected colonies. Random mutagenesis gave rise to a blue emitting colony that was subsequently

sequenced. Upon sequencing, it was determined that the only difference between the wtGFP and mutant GFP was its chromophore (65S, 66Y, 67G). The mutant chromophore had a single substitution of tyrosine (Y) at position 66 with histidine (H) instead at that same position (Tsien, 1994). This single residue replacement created a novel FP referred to as BFP or "Y66H".

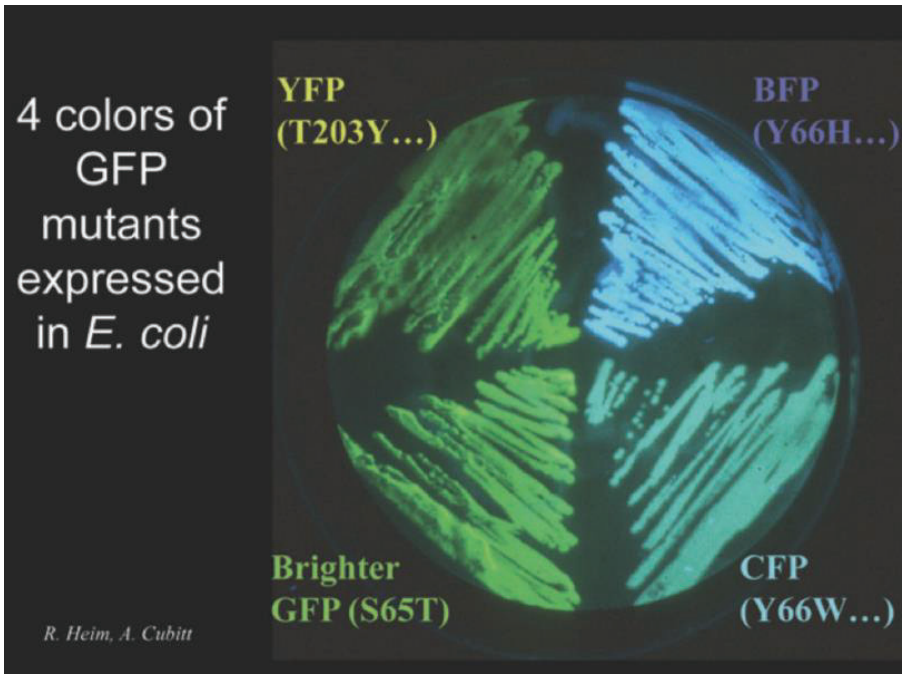


Figure 3. Streak plates of *E. coli* expressing various mutants of GFP. Each mutant was derived from wtGFP and the particular amino acid change to the chromophore is indicated in parentheses (Heim *et al.*, 1995).

The pioneering work by Shimomura, Chalfie, and Tsien: discovering and developing the *Aequorea victoria* GFP gene culminated in them sharing the 2008 Nobel Prize in Chemistry.

GFP STRUCTURE, CHROMOPHORE, AND FOLDING

wtGFP has a molecular weight of 27kDa, a height of 40Å and width of 25Å (Remmington, 2011). As one of the most recognizable proteins in nature, GFP is composed of two main components: 1) an outer, cylindrical β-barrel (sometimes referred to as the "β-can") structure consisting of 11 anti-parallel beta sheets (Craggs, 2009); and 2) a central, internal alpha helix that spans the height of the protein, housing its chromophore which resides almost exactly at the proteins core (Tsien, 1998). The cylindrical β-barrel plays an important role in protecting GFPs chromophore - its source of light emission (Day, 2009) - from oxygen quenching and interactions with hydronium ions and is not tolerant to residue deletions (Zimmer, 2002).

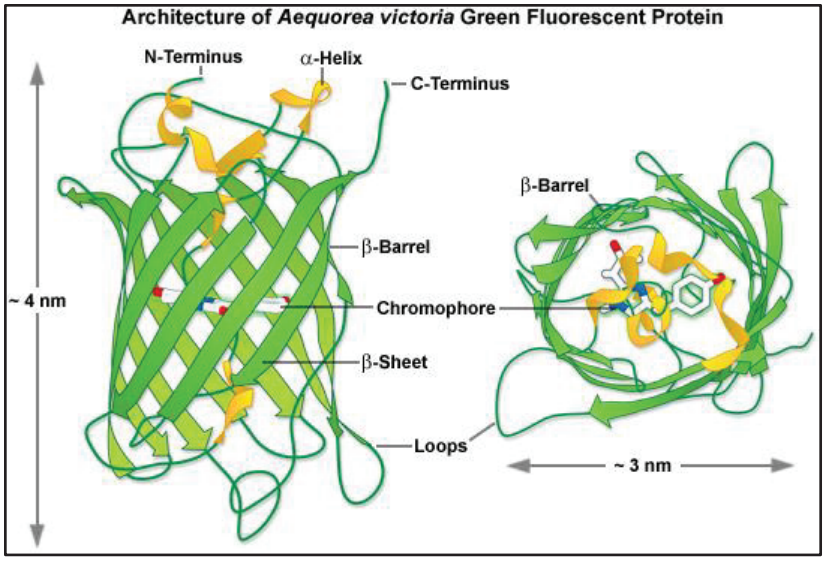


Figure 4. wtGFP side view and aerial view of internal chromophore. Anti-parallel β-sheets are colored green; alpha helical ribbons are in yellow (Source: Day, 2009).

The chromophore is a tripeptide sequence of the residues: S (serine), Y (tyrosine) and G (glycine), at positions S65, Y66 and G67, with Y66 and G67 being a conserved dipeptide seen in

nearly all fluorescent protein chromophores (Craggs, 2009). Able to fold in temperatures below 36°C (Zimmer, 2002), wtGFP chromophore maturation process occurs as an autocatalytic cyclization reaction; a spontaneous post translational modification that does not require additional enzymes or cofactors (Barondeau *et al.*, 2001). This rearrangement of the covalent bonding in the chromophores peptide backbone is characterized in three steps (Figure 5). The first step of the autocatalytic cyclization involves the nucleophilic attack of the G67 amide nitrogen on the adjacent carbonyl carbon atom of S65 forming the imidazolin ring. The second step is dehydration of the α-β bond of Y66 resulting in the formation of the imidazolin-5-1-heterocyclic ring. The third step is oxygen dependent and involves the oxidation of the α-β bond of Y66. This step must occur for wtGFP to fluoresce (Day, 2009).

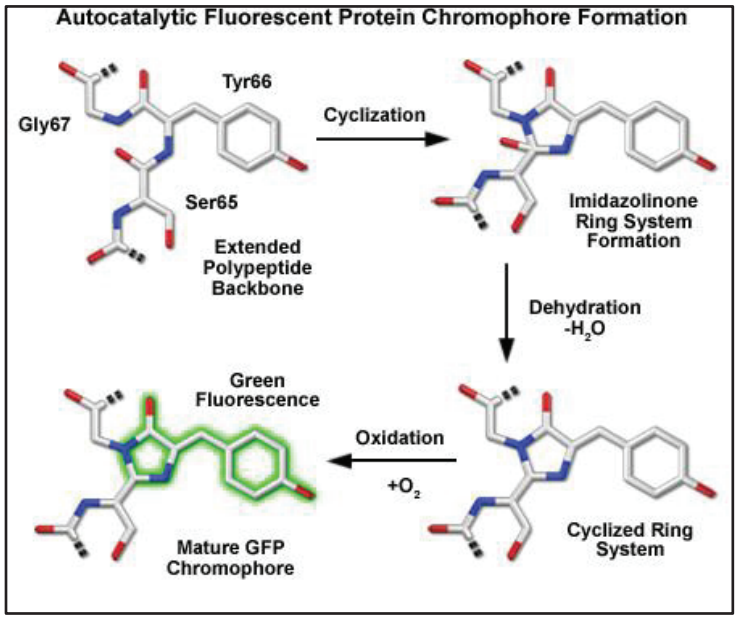


Figure 5. wtGFP autocatalytic chromophore cyclization resulting in fluorescence. This is an oxygen-dependent process and can take anywhere from 90 minutes to several hours *in vivo* for wtGFP. (Source: Day, 2009)

GFP FUNCTION *IN VIVO*

In vitro, GFP requires external excitation with blue or ultraviolet photons (with a peak excitation at 488 nm in order to emit green fluorescence at 509 nm (Zhou, 2005). In the wild however, *Aequorea victoria* uses a coupled, two-protein system to emit light. The first part, the protein aequorin - is excited by relatively short, high energy wavelengths of light; aequorin then remits a longer (blue) wavelength of light that is absorbed by GFP, which in turn is excited and emits green fluorescence. In this coupling of aequorin with GFP, GFP acts as an energy transfer acceptor through radiationless energy transfer, where GFP excitation with blue light causes the protein to emit a longer photon wavelength of green light (Prasher, 1992).

In order for endogenous production of blue light, aequorin must undergo a bioluminescent reaction (Figure 6). Calcium-dependent aequorin consists of two covalently bonded subunits: 1) the protein apo-aequorin (22 kDa) and 2) a small molecule, coelenterazine (400 Da). Upon the binding of three Ca^{2+} ions (one Ca^{2+} ion to each of the three helix-loop-helix domains of apo-aequorin) the covalent bonds between apo-aequorin and coelenterazine break resulting in carbon dioxide, coelentaramide and the emission of blue light (Granatiero, 2014). Interestingly, the blue light emitted by aequorin has a λ_{max} = 465nm (Ohmiya, 1996); the same wavelength needed to excite GFP for the emission of visible green fluorescence. In the case of GvFP, it is unknown whether or not *Gonionemus* synthesizes an aequorin or aequorin-like protein that GvFP acts as a companion to, transferring blue photon energy for green fluorescence.

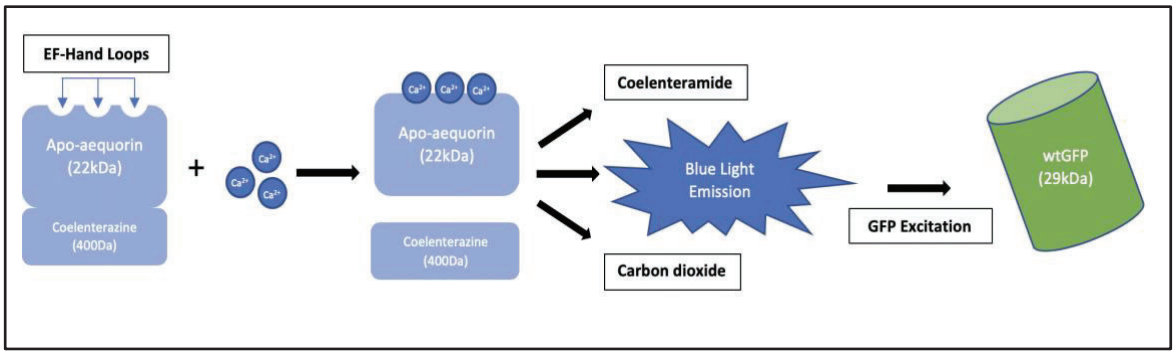


Figure 6. The bioluminescent reaction of calcium sensitive aequorin results in photon transfer enabling blue light excitation and green light emission of wtGFP *in vivo*.

Research exploring the possible applications of fluorescent proteins far overshadows the research looking at *why* these hydrozoans fluoresce. A few theories attempt to explain what benefit fluorescence serves in bioluminescent and chemiluminescent jellyfish. These theories can be placed into two groups; predation and protection. The predatory advantage of GFP in the hydrozoan *Olindias formosus* was observed by Haddock and Dunn (2015). In tanks with a transparent partition separating *Olindias* from juvenile rockfish (*Sebastes* - their prey in the wild), the frequency at which the juvenile rockfish would attempt to attack *Olindias* (and stopped by the partition in the tank) was observed under white, yellow, and blue light. Under white and yellow light, rockfish attacks were at their lowest frequency. Under blue light rockfish attack frequency was at its highest compared to white and yellow light conditions both showing a consistently lower rockfish attack frequency (Haddock, 2015). The attractiveness of green fluorescence to rockfish may be due to the fact they are herbivores and may be mistaking the green tips of *Olindias* tentacles for chlorophyll (Haddock, 2015).

ENGINEERED MUTANTS OF GFP

Random and site-directed mutagenesis of the wtGFP gene (and subsequent mutants) has given rise to the wide variety of engineered fluorescent proteins (FPs) we have today. The earliest single substitution wtGFP derived mutants (BFP: blue fluorescent protein, and CFP: cyan fluorescent protein) all share the same sequence length but differ in sequence each by only a single amino acid. This minute change in sequence creates a change in photon excitation and emission which, in turns, changes the color of visible light these FPs produce.

```
CLUSTAL O(1.2.4) multiple sequence alignment


wtGFP       MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL 60
BFP         MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL 60
CFP         MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL 60
            ************************************************************

wtGFP       VTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV 120
BFP         VTTFSHGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV 120
CFP         VTTFTWGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLV 120
            ****: ************************************************************

wtGFP       NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD 180
BFP         NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD 180
CFP         NRIELKGIDFKEDGNILGHKLEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD 180
            ************************************************************

wtGFP       HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK   238
BFP         HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK   238
CFP         HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMVLLEFVTAAGITHGMDELYK   238
            ************************************************************
```

Figure 7. A Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) amino acid multiple sequence alignment of wtGFP (FPBaseID: 1XF1B) and its derived multi-colored mutants; CFP (cyan fluorescent protein- FPBaseID: ZRLU4), and BFP (blue fluorescent protein- FPBaseID: 671NA). The boxed region denotes residues 65, 66, and 67 that form the chromophore.

Other mutants have been engineered to circumvent certain limitations of wtGFP. For instance, wtGFP fluorescence brightness was increased by five-fold with the creation of EGFP (enhanced GFP). EGFP was engineered by the substitution of the three residues: S65T, F64L, and H231L (Day, 2009). Unlike wtGFP, EGFP also has a notably faster maturation rate, meaning it can emit fluorescence after transfection in cells quicker than wtGFP, an occurrence observed by the quantification of immature protein in cells (wtGFP vs. EGFP) after translational arrest using chloramphenicol (Balleza, 2018). Although mutants may appear attractive over wtGFP, they can come at an expense. EGFP is more pH sensitive than wtGFP and also tends to dimerize. These issues were resolved by the creation of the subsequent engineered FP known as Emerald. Emerald has the same S65T, F64L, and H231L mutations as EGFP but also contains the following substitutions: S72A, N149K, M153T, and I167T. These additional substitutions combined with those of EGFP gives Emerald brighter fluorescence than EGFP and eliminates the tendency of dimerization (Day, 2009).

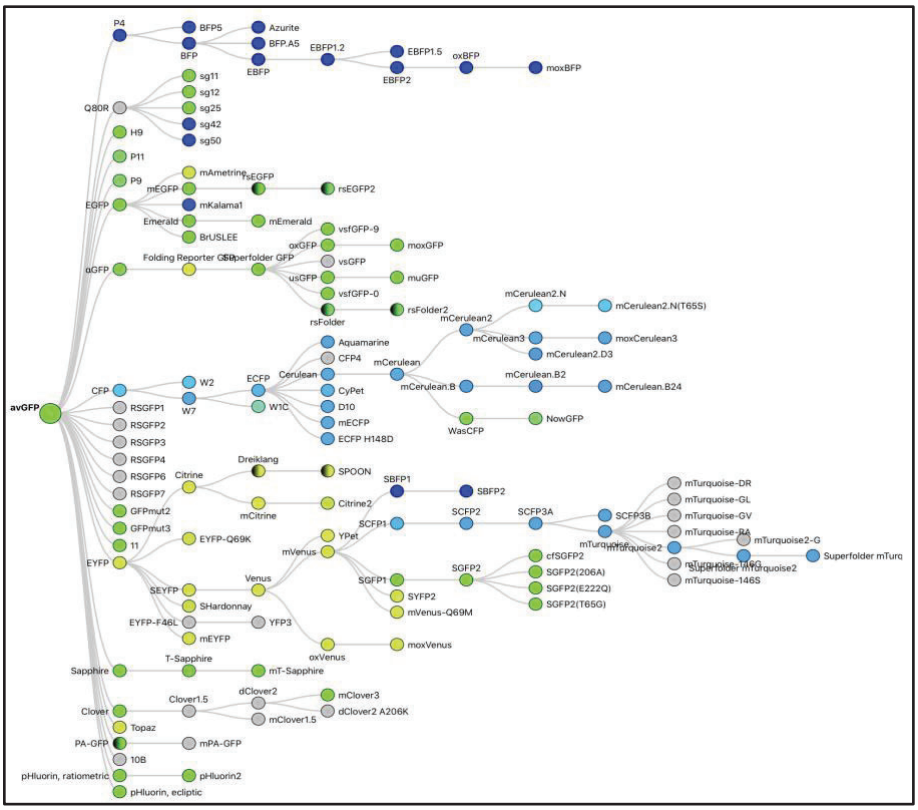


Figure 8. (Source: www.fpbase.org). This tree represents wtGFP derived mutants that form the diverse family of engineered fluorescent proteins with variable excitation and emission wavelengths alongside other modified characteristics.

Aequorea wtGFP derived mutants are limited to a maximum emission of 540 nm which in turn only allows wtGFP mutants a maximum shift to yellowish fluorescence (Chudakov, 2010). Orange (590 nm to 620 nm) and red (620 nm to 750 nm) fluorescence is only seen in Anthozoan FPs like *Discosoma* spp. (coral) which synthesizes a red fluorescent protein (dsRed) that excites at 558 nm and emits at 583 nm (Gross *et al.*, 2000). Like wtGFP, dsRed has been mutated by researchers in order to further exploit the protein for research applications.

APPLICATIONS IN BIOLOGICAL RESEARCH

Protein Localization

Since its initial discovery GFP has been exploited for its ability to report on the location and expression of a protein inside cells or tissues non-invasively, just by looking for a green fluorescent signal. The ability of GFP to tolerate the fusion of other proteins at the amino or carboxyl termini, creating what is known as a fusion protein, without impact on either proteins function, localization or fluorescence is well documented (Chudakov *et al.*, 2009). This makes it possible to determine protein localization *in vivo* as well as the levels of protein expression by quantifying the fluorescent signal. This method can be applied to protein localization in a cell, tissue, organ, or whole organism. As discussed earlier, Chalfie (1994) was the first to demonstrate protein localization in the transparent *C. elegans,* allowing the mapping of the Mec-7-GFP fusion was expressed with resolution to individual neurons (Chalfie *et al.*, 1994). Protein localization can sometimes require a bit of trial and error. Although GFPs are tolerant of N- or C-terminal fusions, some proteins rely on their N- or C- termini to be free depending on where they localize or other molecules they interact with (Chudakov *et al.*, 2009). The addition of a linker peptide between GFP and the tagged proteins C- or N-terminus is useful to avoid a POI that becomes misfolded or is rendered non-functional (Chen *et al.*, 2013). Chromophore maturation can also be problematic when attempting to tag a protein. Since protein half-life can range from hours to years (Toyama, 2013) before proteolysis occurs, using a GFP that matures slowly can be problematic. For wtGFP, which takes between 90 minutes and 4 hours for its chromophore to mature (Zimmer, 2002), its long chromophore maturation time won't allow for fluorescence until after its fused protein is

destroyed (Philipps, 2001). Mutant FPs have been designed with faster folding times to ensure that fused proteins can be visualized before they degrade.

Protein-Protein Interactions

GFP has also proven useful for interrogating protein-protein interactions (Figure 9). FRET (Forster Resonance Energy Transfer) involves the use of two fluorescent proteins to determine whether or not two proteins of interest (POI) interact with one another (Chudakov *et al.*, 2010) based on their proximity. FRET is able to detect proximity of fused chromophores at a maximum distance of 10 nm (Bajar *et al.*, 2016) with increased proximity resulting in an increase in fluorescence signal, indicating the presence of POI-POI interaction, (Chudakov *et al.*, 2011) while the absence of chromophore-acceptor emission indicates the absence of a POI-POI interaction. Both POIs are each fused to a different FPs; one of which acts as an energy donor, and the other acts as an energy acceptor. These two FPs are referred to as FRET-FP pairs. FRET FP pairs are designed so that the donor FP has an overlapping emission spectrum with the excitation spectrum of the acceptor FP (Zimmer, 2002). Conformational changes in protein structure can also be monitored by fusing a donor FP to the amino-terminus and fusing an acceptor FP to the carboxyl terminus (Zimmer, 2002) or a given protein. In the presence of a molecule that induces a structural change, fluorescence emission from the acceptor FP due to energy transfer is indicative of FP proximity (Hennigan *et al.*, 2009) meaning that a structural change has occurred. FRET can also be utilized to determine whether or not a POI interacts with a specific DNA sequence (Blouin *et al.*, 2009) using a similar set up seen in FRET POI-POI interactions.

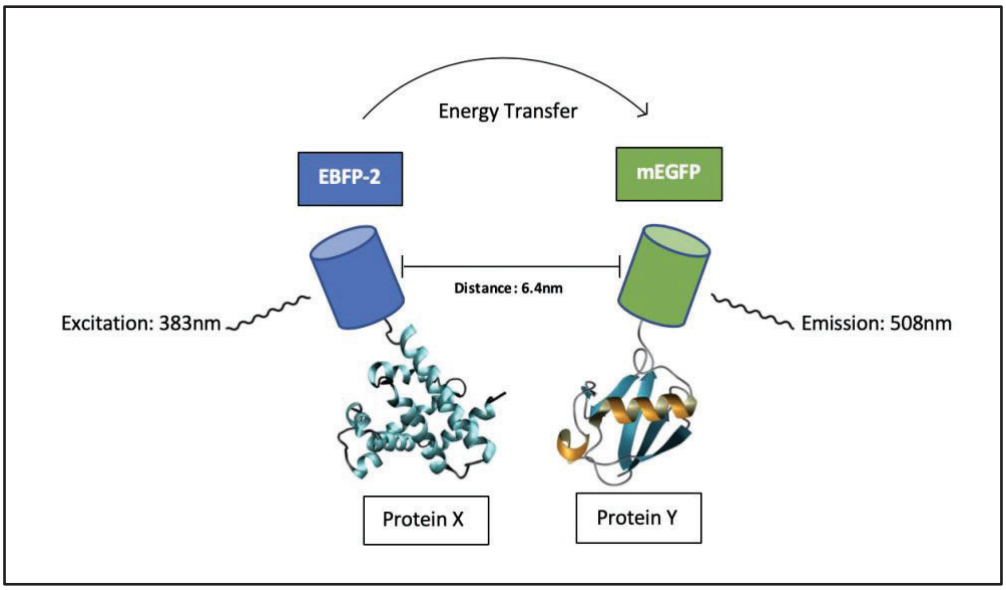


Figure 9. FRET FP Pair EGFP2-mEGFP forming chimeras with POI X and POI Y share a proximity under 10 nm allowing for Forster Energy Transfer between the donor chromophore (EBFP-2) and the acceptor chromophore (mEGFP) causing the emission of green light at 508 nm from mEGFP.

THE BIOLOGY OF *GONIONEMUS VERTENS*

The jellyfish *Gonionemus vertens* is a member of the phylum Cnidaria, class Hydrozoa, order Leptothecata, and the family Olindiidae (Adachi *et al*., 2017). Growing to be 20-25 mm in bell diameter at its adult medusa stage, *Gonionemus* is typically found clinging to sea grass in shallow water (giving it the nickname the "clinging jellyfish") using its 60-80 tentacles that extend from its bell border (Williamson, 1994). In the wild, *Gonionemus* typically feeds on small crustaceans and zooplankton like copepods (Marchessaux *et al.*, 2017).

Although *Gonionemus* is native to the North Pacific Ocean (Govindarajan, 2017), it has been identified as an invasive species appearing on the coastlines of Portugal, France, Norway,

Germany, Denmark, China, Argentina, and the United States (Marchessaux *et al.*, 2017). *Gonionemus* was recently found in the Shrewsbury and Manasquan rivers of New Jersey (Gaynor *et al.*, 2016), making the organism a topic of interest to researchers at Montclair State University. These invaders were genetically identified by isolation and sequencing of DNA barcoding regions (16S rDNA and COI - cytochrome c oxidase subunit I). While the mechanism through which *Gonionemus vertens* invades marine ecosystems in the mid-Atlantic United States and abroad is not completely known, it is suspected that invasions occur due to human activity, specifically ballast water released from cargo ships (Marchessaux *et al.*, 2016).

Figure 10. *Gonionemus vertens* in artificial seawater illuminated with blue light (470 nm) displaying green fluorescence (left) and *Gonionemus vertens* in artificial seawater exposed to white light (right). Fluorescence is usually associated with rhopalia along the margin of the bell, the gonads associated with the radial canals, and the manubrium. The bell diameter in this organism is generally 2 cm.

To the average beach-goer, *Gonionemus* is best avoided due to its potent sting known to cause a variety of neurological symptoms like extreme muscle weakness and cramping, pain, as

well as anxiety, and restlessness manifesting only 15 to 20 minutes post-sting (Pigulevsky, 1969). These symptoms can require hospitalization in some cases. Some *Gonionemus* stings are considerably less toxic, leaving only scarring after a sting. This less toxic population is suspected to have originated from the eastern North Pacific Ocean (Govindarajan, 2019); unlike its toxic counterparts which are thought to have originated from the western North Pacific Ocean (Gaynor *et al.*, 2016). Current research at Montclair State University seeks to determine the protein components that make up the toxic venom of *Gonionemus* which are currently unknown.

RESEARCH OBJECTIVES

Previous research has reported that *Gonionemus vertens* has a natural green fluorescence when illuminated with blue light (Petroff *et al*., 2017) although the specific gene and protein encoding this feature have not yet been identified. Accordingly, this research project has 3 specific objectives: 1). To generate an RNA-seq library from clinging jellyfish RNA; 2). exploit bioinformatic tools to identify *Gonionemus vertens* green fluorescent proteins (GvFP) in these libraries and to characterize the gene and putative protein; and 3). to construct pET-SUMO plasmids containing the full-length gene of this putative GvFP for future expression studies in *E. coli.*

MATERIALS AND METHODS

1. FIELD COLLECTIONS

Mature *Gonionemus vertens* medusae were harvested from Farm Pond, Martha's Vineyard, MA in August 2017. Farm Pond is a 42-acre coastal salt pond near the town of Oak Bluffs on Martha's Vineyard. It is surrounded by 8-acres of salt marsh and is connected to the ocean by an open culvert and, thus, is subject to tidal cycles. Approximately 320 individuals were harvested in eelgrass beds (*ca*. 1 m deep water) and returned live to Montclair State University. All samples were subjected to several changes in artificial seawater (25 ppt) over 24 hours to remove bycatch and to ensure the stomach contents were cleared. Five individuals were selected, labelled MV1 through MV5, and immediately placed in 2 mL of RNALater (www.qiagen.com) and stored at -20°C for subsequent RNA extraction.

2. TOTAL RNA PURIFICATION

The QIAamp RNA Blood Mini Kit from Qiagen was used to purify total RNA from each *Gonionemus vertens* sample (MV1, MV2, MV3, MV4, MV5). The **Purification of Total RNA from Tissue** protocol (www.qiagen.com) provided by the manufacturer was used with the following exceptions. Individual *Gonionemus vertens* specimens were removed from -20°C in RNALater and thawed quickly on ice. The RNALater solution was removed from each jellyfish prior to the addition of the Qiagen RLT buffer. Each sample was homogenized for 20 seconds at high power (35,000 rpm) using a Pro Scientific (PRO200 – Biogen Series) handheld homogenizer with a 7-mm x 95-mm saw-tooth generator probe (https://proscientific.com). Total RNA isolated was quantified by both UV absorbance (NanoDrop ND-1000) as well as by fluorescence on a Qubit 3.0 Fluorometer using the Broad Range (BR) RNA kit. $OD_{260/280}$ ratios were also calculated

for all samples using the NanoDrop ND-1000. All tubes and pipette tips used were sterile and RNAse-free.

3. mRNA ISOLATION

The NEBNext Poly(A) mRNA Magnetic Isolation Module from New England Biolabs (NEBNext Poly(A) mRNA Magnetic Isolation Module) was used to isolate polyadenylated mRNA from previously purified total RNA. In all cases, 1 µg of total RNA was used in 50 µL of sterile nuclease-free water as the starting material. Total DNA-free polyadenylated RNA was isolated using NEBNext Magnetic Oligo $d(T)_{25}$ beads as described by the manufacturer's protocol (NEBNext Poly(A) mRNA Magnetic Isolation Protocol). All extractions, washings, and thermocycler incubations were done in a 96-well PCR tube plate (200 µL). Magnetic beads were isolated by mounting the 96-well plate on a magnetic rack (BEL-ART Magnetic Bead Separation Rack for 96-well plate). All tubes and tips used in the mRNA isolation were nuclease-free and RNA was stored at -80°C until used for library preparation.

4. RNA-SEQ LIBRARY PREPARATION

4.1 Preparation of NGS Library

Library preparation was performed using the NEBNext Ultra II Directional RNA Library Prep Kit (New England Biolabs). Poly A^+ mRNA was utilized to eliminate or minimize the inclusion of rRNA and tRNA in the construction of the RNA-Seq library. Briefly, the protocol included the following steps: 1) poly A^+ mRNA was fractionated (to generate mean lengths of 200 to 300 nt) and annealed with random primers. 2) First strand cDNA was generated with the use of reverse transcriptase and dNTPs. 3) Second strand synthesis using DNA polymerase and dNTP's (including dUTP) followed by cleanup by binding to AMPure XP beads (Beckman-Coulter, Inc.

#A63881). Binding and elution from beads preferentially bind to dsDNA and removes primers, small ssDNA, and unincorporated dNTPs. 4) End repair of dsDNA and the addition of a dA tail to the 3' ends. 5) Unique adaptors are then ligated to the dsDNA fragments to provide a unique identifier for each library prep and to permit multiplexing of samples. 6) USER-enzyme digestion (USER stands for Uracil-Specific Excision Reagent). This enzyme permits the generation of a single nucleotide gap at the location of a uracil residue (dUTP was incorporated into the second strand cDNA). This makes these RNA-Seq libraries directional. Procedure is followed by another round of magnetic bead clean-up as detailed previously. 7) PCR enrichment of adaptor-ligated DNA. Manufacturer's protocol was followed with the exception that 9 amplification cycles were used based on our initial input of 1000 ng RNA into each library prep. This was followed by another round of magnetic bead clean-up as detailed previously. 8) Final libraries were assessed for size distribution and yield on Agilent Bioanalyzer (DNA 1000 chip) following manufacturer's protocol.

4.2 NGS Sequencing

DNA was sequenced on an Illumina MiSeq platform using a V3 Reagent cartridge and 2 x 300 paired ends. Approximately 34,500,000 reads were generated from this run from the five pooled transcriptome libraries.

4.3 Contig Assembly

Raw sequence data (FASTQ files) were processed by eliminating sequences with low quality scores, removal of adaptor sequences, sorted into individual libraries based on index reads, and then assembled using Geneious (version 10.0.6) Assembler to generate a file of 39,732 contigs for the MV5 library.

5. BIOINFORMATIC ANALYSIS

Sequence data was generated as ABI chromatogram files. The ABI file format is a binary file produced by ABI sequencer software. Macintosh compatible software applications 4Peaks (http://4peaks.en.softonic.com/mac) and Geneious (http://www.geneious.com/) were used to trim, assemble and view Sanger sequencing trace files, correct base calls and create consensus sequences. All sequencing data generated locally was assembled both *de novo* and by comparing with a reference sequence generated by RNA transcriptome analysis. Sequences were searched against GenBank (www.ncbi.nlm.nih.gov) using the BLASTn and BLASTx algorithms. BLAST2Seq algorithm was also used to produce alignments between overlapping sequences and to help resolve inconsistencies between forward and reverse sequencing reads. BLASTn searches, unless otherwise specified, were done using standard default values. A match with an e (expect) value of $< 10^{-4}$ was considered a match.

6. PCR

6.1 Master Mixes

I-5™ 2X High-Fidelity Master Mix DNA Polymerase manufactured by Molecular Cloning Laboratories (320 Harbor Way, South San Francisco, CA 94080; https://www.mclab.com) (10 µL) was pipetted into each labeled, sterile 200 µL polypropylene thin-wall PCR tubes. The I-5™ 2X High-Fidelity Master Mix contained all necessary dNTPs, buffer and salts, with a final Mg^{2+} concentration of 2 mM. To every reaction tube containing 10 µL I-5™ 2X High-Fidelity Master Mix, 1 µL each of forward and reverse primer (10 µM stocks), 1 µL of the DNA sample and 7 µL of sterile deionized water was pipetted in to obtain a final reaction volume of 20 µL. See Appendix

X for a complete list of primers used. In each run, a negative control (NTC, or no template control) received 8 µL of sterile deionized water, but no DNA.

In cases where we needed to generate PCR amplicons with a terminal 3' A residue, we utilized a non-proofreading thermostable DNA polymerase. Specifically, when amplicons were required for cloning by TA tailing, we used the NEB Master Mix.

6.2 Cycling parameters

PCR tubes containing 20 µL of reaction mixture were inserted into the Applied Biosystems ProFlex™ 3 x 32-well PCR System thermocycler. Routinely, when we utilized the I-5™ 2X High-Fidelity Master Mix, initial denaturation was carried out one time at 98ºC for 2 minutes. Then, amplification cycles comprised denaturation, annealing and extension phases were run for 30 cycles at 98 ºC for 10 seconds, 55 ºC for 15 seconds, and 72ºC for 30 seconds, respectively. After the 30 amplification cycles, a final extension was run at 72ºC for 5 minutes. The samples were held at 4ºC until they were removed from the thermocycler.

7. Agarose Gel Electrophoresis

7.1 Gel Preparation

Agarose gel electrophoresis was used as a standard method to assess both purity and size of amplicons generated by PCR. To prepare a 1.0% (w/v) agarose gel 0.40 g of agarose was mixed with 40 mL 1X TAE buffer (40 mM Tris - Acetate, 1 mM EDTA) in a 250 mL Erlenmeyer flask. To avoid any spillage and over boiling, the agar was boiled in a microwave at low power (3) for 3 minutes and 15 seconds. It was ensured that the agar had fully dissolved to a clear solution. The solution was allowed to cool briefly. While the solution was still a liquid, 4 µL of Invitrogen™ SYBR™ Safe DNA Gel Stain (10,000X in DMSO, Invitrogen) was added to the flask and gently

swirled avoiding any bubble formation. A casting tray with 10- or 12-well comb was readied in the gel box. The agar was gently poured into the casting tray and allowed to gel for 15 minutes. When solidified, the comb was firmly removed and oriented with the wells on anode. Then, the running buffer 1X TAE was poured into the gel box until gel was submerged about 5 millimeters.

7.2 Preparing Samples for Gel Electrophoresis

Two µL of 6X loading dye (0.25% (w/v) Bromophenol Blue, 0.25% (w/v) Xylene Cyanol, 30% (v/v) Glycerol) was pipetted into each labeled 1.5 mL Eppendorf tube. Of the 20 µL PCR reactions, only 10 µL of PCR amplicon was pipetted into each tube and mixed with the loading dye (remaining sample was stored at –20ºC for subsequent DNA sequence analysis or cloning). The mixture was centrifuged briefly. Ten µL of HiLo DNA ladder (HiLo DNA Ladder, Minnesota Molecular; http://www.mnmolecular.com) was pipetted into the end wells flanking the samples. Each sample (12 µL) was carefully pipetted into individual wells. The gel box was covered and plugged into the power supply and run at 100 volts for 45 minutes. Gel running time was adjusted based on the length of the PCR amplicon.

7.3 Digital Gel Imaging

The gels were visualized by illumination with blue light (470 nm) and imaged as described by Gaynor *et al.* (2016). The bands observed on the gel were compared with the HiLo DNA ladder (Minnesota Molecular, Inc.) to determine DNA length.

8. Automated Sanger Dideoxy Sequencing

Amplicons deduced to be both clean (i.e., a single band) and of sufficient quantity (based on intensity of the band) were processed for DNA sequence analysis. If there was more than 1 µg of an amplicon in a lane, then that sample was diluted accordingly for DNA sequence analysis

(typically between 10- and 100-fold depending on the intensity of the band). All dilutions were done with sterile deionized water. Samples submitted for sequencing contained 1 µL of amplicon (or diluted amplicon), 1 µL of forward or reverse primer (10 µM stock), and 8 µL of sterile deionized water. Samples were sequenced in both the forward and reverse directions. Sequencing was performed using the BigDye Terminator Cycle Sequencing Kit Version 3.1 (Applied Biosystems Inc., Foster City, CA 94404; BigDye Terminator Cycle Sequencing Kit Version 3.1) following the manufacturer's instructions with the exception that we routinely ran 1/16 reactions. Cleanup was performed using an EdgeBio Performa DTR Gel Filtration Cartridges (Gaithersburg, MD; https://www.edgebio.com). The samples were analyzed using an ABI3130 Genetic Analyzer from Applied Biosystems (Foster City, CA) using a 36-cm column array and NANO POP7 polymer (MCLAB, South San Francisco, CA 94080, NP7-100; http://www.mclab.com). Sequence calls were made using the KB Basecaller.

8.1 Analyzing DNA sequences

Sequence data was generated as ABI chromatogram files. The ABI file format is a binary file produced by ABI sequencer software. Macintosh compatible software applications 4Peaks (http://4peaks.en.softonic.com/mac) and Geneious (http://www.geneious.com/) were used to trim, assemble and view Sanger sequencing trace files, correct base calls and create consensus sequences. All sequencing data generated locally was assembled both *de novo* and by comparing with a reference sequence generated by RNA transcriptome analysis. Sequences were searched against GenBank using the BLASTn algorithm. BLAST2Seq algorithm was also used to produce alignments between overlapping sequences and to help resolve inconsistencies between forward

and reverse sequencing reads. BLASTn searches, unless otherwise specified, were done using standard default values. A match with an e (expect) value of $< 10^{-4}$ was considered a match.

8.2 Gene Alignment and Assembly

Sequencing data was analyzed both manually and with the help of Geneious sequence analysis software for *de novo* gene assembly. For manual assembly, sequencing data were extracted from the chromatogram files and copied into a word processor. A short segment of the nucleotide was selected from a sequence and searched for relative abundance. Sequences with similarities were aligned to create a scaffold and other sequences were built around it by joining overlapping fragments. A consensus sequence generated by the manual assembly of the sequence data were aligned to the mRNA seq data for validation. The built-in Geneious assembler, Tadpole, SPAdes, Velvet, MIRA and CAP3 assemblers available in the Geneious sequence analysis software were also used to validate the assembled sequences.

9. CONSTRUCTION OF pET-SUMO-GvFP EXPRESSION VECTOR

9.1 Ligation Reaction

The Champion™ pET-SUMO Expression System by Life Technologies (Catalog # K30001) was used for cloning and expression of our gene. A 1:1 molar ratio of vector to insert was recommended by the manufacturer for optimum ligation efficiency. One µL of PCR sample was determined to be optimal for ligation. In a 200 µL tube, 10 µL ligation reaction was carried out with 1µL of fresh PCR product, 1 µL of the 10X Ligation buffer, 2 µL of the 25 ng/µL pET SUMO vector (linearized), 5 µL of deionized water and 1 µL of the 4.0 Weiss units T4 DNA Ligase. The reaction was incubated overnight at 15ºC in a thermocycler for the full length (690 bp) insert.

9.2 Transforming One Shot® Mach1™-T1R Competent Cells

Before transformation was carried out, the S.O.C medium (Invitrogen, Catalog No. 15544-034) was equilibrated to room temperature. LB plates containing 50 µg/mL Kanamycin (LB-KAN) were warmed in an incubator at 37ºC. The One Shot® Mach1™-T1R competent cells were removed from -80ºC freezer and thawed on ice. Each ligation product (2 µL) was pipetted into individual vials containing One Shot® Mach1™-T1R Competent *E. coli* (Invitrogen, Cat. No. C8620-03). The cells and ligation mixture were mixed by gently flicking the walls of the tube and incubated on ice for 15 minutes. Cells were heat shocked in a water bath at 42ºC for 30 seconds and immediately transferred to ice for 2 minutes. Room temperature S.O.C medium (250 µL) was added to the tube containing cells. The tubes were tightly capped and horizontally shaken in a 37ºC shaking incubator for 1 hour at 200 rpm. Two separate volumes, 100 µL and 200 µL, of the transformants were aseptically spread on a pre-warmed LB-KAN plates and incubated overnight at 37ºC.

9.3 Screening for Inserts

Screening for the correct insert was carried out by colony PCR and sequencing of the isolated plasmid DNA (as described in section X of the Materials and Methods). Only those clones that contained an insert in the 5' to 3' direction and correct reading frames were subcultured onto fresh LB-KAN plates and LB-KAN broth for plasmid mini-prep. The NEB Monarch® Plasmid DNA Miniprep Kit (NEB #T1010) was used to isolate and purify plasmid DNA.

9.4 Transforming BL21(DE3) One Shot® Cells for Expression

9.4.1 Ligation and Transformation

BL21(DE3) One Shot® cells (Invitrogen, Cat. No. C6000-03) were removed from -80ºC freezer and quickly thawed on ice. The volume of the plasmid DNA was adjusted so the final concentration was 10 ng/µL. Plasmid DNA (1 µL) was added to each vial of BL21(DE3) One Shot® cells and stirred gently with the pipette tip. The mixture was incubated on ice for 30 minutes. The tubes containing the ligation mixture were then introduced to a 42ºC water bath to heat-shock the cells for 30 seconds. Tubes were then immediately transferred to ice. To the mixture, 250 µL of room temperature S.O.C. medium was added and incubated at 37°C for 1 hour with shaking at 200 rpm.

The entire transformation reaction was added to 10 mL of LB broth containing 50 µg/mL Kanamycin. The culture was grown overnight at 37°C with shaking at 200 rpm.

9.5 IPTG Induction and Sample Processing

An aliquot (500 µL) of the overnight culture was inoculated into 10 mL of LB broth containing 50 µg/mL kanamycin (in a 125-mL Erlenmeyer flask) and allowed to grow for two hours at 37°C with shaking at 200 rpm. The optical density (OD600) reading was taken at the end of the two hours. The sub-culture was then split into two 5 mL cultures. To one of the 5 mL cultures, IPTG was added to a final concentration of 1 mM to induce protein expression, while the other culture was left uninduced. From each culture, 500 µL aliquot was taken immediately and centrifuged at maximum speed (13,200 rpm) in a microcentrifuge for 1 minute. The supernatant was aspirated, and the cell pellets were immediately frozen at -20°C. These were marked as the zero-time point samples. The remaining cultures were incubated continuously at 37°C with shaking at 200 rpm. Time points for each culture was taken every hour for 4 hours. For each time

point, 500 µL of both the induced and uninduced cultures were taken and processed as described above.

9.6 Sample Analysis

Samples were analyzed by direct screening of aliquots for green fluorescence when illuminated by blue light (475 nm). Each 500 µL aliquot was centrifuged for 2 minutes @16,000 x g (Eppendorf 5615C Microfuge) to pellet cells. Pellets were resuspended in ice-cold 10 mM Tris (pH 8.0) and washed one more time by a brief centrifugation. Then, each 500 µL washed aliquot was sonicated on ice with a microprobe tip for 30 seconds at full power (7-mm x 95-mm saw-tooth generator probe) and centrifuged for 2 minutes @16,000 x g (Eppendorf 5415C Microfuge) to pellet unbroken cells and cellular fragments. The remaining supernatant was measured in triplicate for green fluorescence in a Qubit 3.0 Fluorometer. A control consisting of 10 mM Tris (pH 8.0) was used to normalize the baseline of fluorescence.

10. MEASUREMENT OF GvFP FLUORESCENCE

Briefly, fluorescence was measured in a Hitachi F-7000 spectrofluorometer at 25˚C. Emission and excitation spectra were also determined on crude extracts of GvFP from *G. vertens* or *E. coli* strains expressing a recombinant GvFP. Alternatively, the Qubit 3.0 Fluorometer was used for determining changes in relative green fluorescence in *E. coli* expressing GvFP cloned into the pET-SUMO expression vector when induced by 1 mM IPTG. Crude GvFP was produced homogenizing two live medusa of *Gonionemus vertens* in 10 mM Tris (pH 8.0) in a glass hand homogenizer (Potter-Elvehjem homogenizer) using 10 strokes. The homogenate was clarified by centrifuging for 2 minutes @16,000 x g (Eppendorf 5615C Microfuge) to pellet unbroken cells

and cellular debris. Clarified supernatant was kept one ice and diluted up to 2 mL with ice-cold 10 mM Tris (pH 8.0) to facilitate fluorescence scans.

RESULTS

1. Fluorescence Emission Spectrum of Native GvFP from *Gonionemus vertens*

In an effort to understand the green fluorescence seen in native *G. vertens*, a number of intact, live medusae were extracted for total soluble protein as described in Materials & Methods. The native GvFP emission spectra was analyzed in 10 mM Tris (pH 8.0) by excitation with 470 nm (blue) light and scanned from 500 to 600 nm in 1 nm steps (Figure 11). The peak emission was seen at 513.8 nm, which is the green portion of the visible spectrum. Excitation with 90$_o$ polarized light (@470 nm) shows a dramatic reduction in fluorescence with no change in λ_{max}.

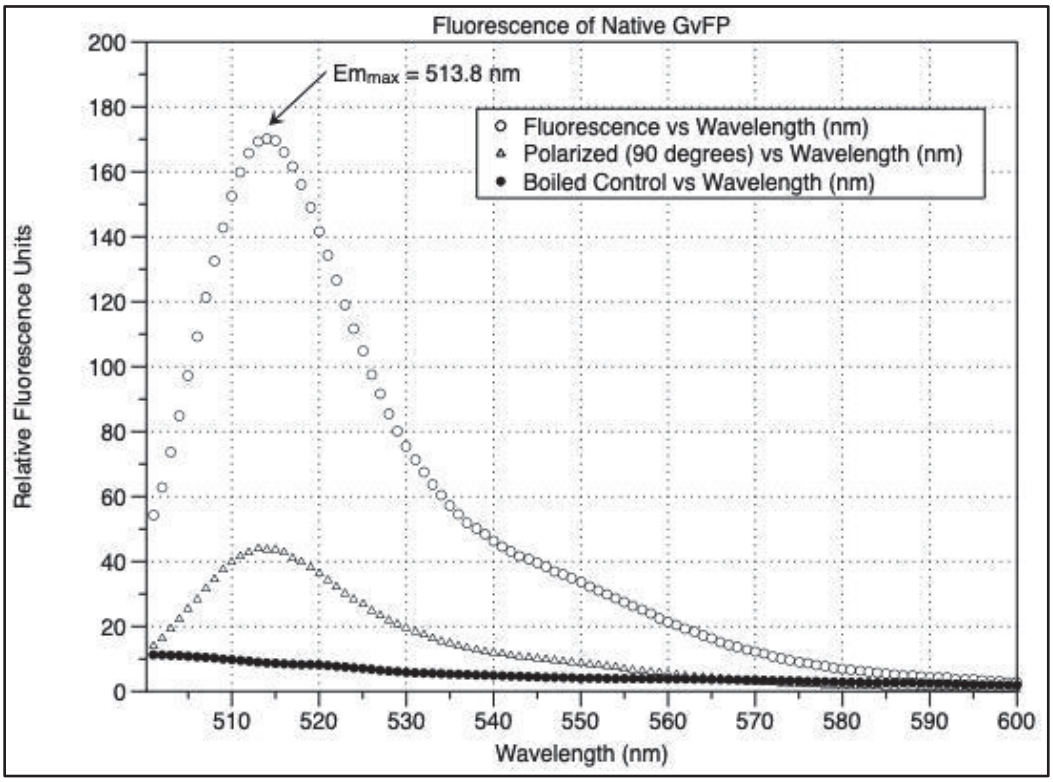


Figure 11. Native GvFP homogenate emission fluorescence spectrum excited at 470 nm (in 10 mM Tris, pH 8.0, 1-cm path length cuvette). Polarized sample was excited at 90$_o$; boiled sample was heated @100$_o$C for 10 minutes.

This dramatic reduction implies that the fluorescence is likely due to a large macromolecule (*e.g.*, protein) and not a small, rapidly diffusing compound (wi). In addition, the disappearance of fluorescence in the boiled sample is consistent with this phenomenon being associated with a protein that becomes denatured (unfolds) at high temperatures. Wild-type GFP (wtGFP) is known to lose fluorescence after denaturation (Remmington, 2011). Extraction and the resulting spectra was unchanged by the inclusion of 1 mM EDTA in the Tris extraction buffer (data not shown).

2. Amplification of the GvFP Gene from *Gonionemus vertens* by PCR Using Conserved and Degenerate Primers

Traditionally, many genes have been cloned using a strategy of designing primers from homologous genes that have been cloned in related organisms. The assumption is that many genes share enough homology to permit using synthetic primers as molecular probes or as PCR primers to isolate the gene in question. In this study, a total of 32 primers were designed and synthesized in an effort to clone the *Gonionemus vertens* GFP gene. Table A lists 5 primer sets (forward and reverse) initially designed to amplify various regions of the original *A. victoria* GFP gene (cDNA). Table B lists degenerate primers designed from conserved regions of the *Aequorea* GFP protein. Single-letter abbreviations of both DNA and protein are based on the IUB code (see Appendix D). Table C lists primers designed from the GFP of *Olindias formosa,* which is another hydrozoan and considered a very close relative to *Gonionemus vertens*. Lastly, Table D lists primers designed to amplify the wtGFP full-length cDNA, as well as internal primers that target the conserved chromophore region.

Despite near saturation coverage of the *Aequorea* and *Olindias* GFP genes, none of these primers succeeded in amplifying a clean amplicon from *G. vertens* gDNA. This also included using these primers in combinatorial arrangements (*i.e.*, not mated with its original pair) to compensate for the possibility of one member of a primer pair not working. For example, each forward primer of a set was tried individually with each reverse primer of a set. Despite all of these attempts, as well as using lowered annealing temperatures during PCR reactions, we failed to generate any clean amplicons. All the primers designed from the wtGFP of *A. victoria* did amplify the corresponding region in the pGLO plasmid that harbors the full-length cDNA of wtGFP (data not shown). These results suggested that the GvFP of *Gonionemus* does not share compatible priming sites and may be considerably different than those of *Aequorea* and *Olindias*.

Table A. GFP Primers - designed using PrimerQuest (IDT-DNA) tool off of the original GFP mRNA in GenBank (*Aequorea victoria)*

Primer Name and Direction	5'-3' Sequence
1F	GGA GTT GTC CCA ATT CTT GTT G
1R	GTC TTG TAG TTC CCG TCA TCT T
2F	CCA TTA CCT GTC CAC ACA ATC T
2R	ACA CTC CAA GTA GCC TGT TTA AT
3F	GAT GAC GGG AAC TAC AAG ACA C
3R	GCA GAT TGT GTG GAC AGG TAA
4F	CCA TGC CCG AAG GTT ATG T
4R	ACA GGT AAT GGT TGT CTG GTA AA
5F	GGC CAA CAC TTG TCA CTA CT
5R	TCT GCT AGT TGA ACG CTT CC

Table B. Design of degenerate primers to isolate GFP homologs from *Gonionemus vertens*. Degeneracies follow the IUB Code (see Appendix B). Single letter amino acid code is listed below each set of degenerate primers targeting a specific region of the GFP protein (*Aequorea victoria*).

5'-3' Primer Sequence	GAY GGC TGC GTN AAY GGD CA
Single Letter Amino Acid IUPAC Code	D G C V N G H (Positions: 19-25)
5'-3' Primer Sequence	GTT ACA GGT GAR GGM GAR GG
	GTT ACA GGT GAR GGK GAR GG
	GTT ACA GGT GAR GGM AAY GG
	GTT ACA GGT GAR GGK AAY GG
Single Letter Amino Acid IUPAC Code	V T G E G E/N G (Positions: 29-35)
5'-3' Primer Sequence	TTC CAY GGT RTG AAY TTY CC
Single Letter Amino Acid IUPAC Code	F H G M/V N F P (Positions: 125–131)
5'-3' Primer Sequence	CCT GCC RAY GGT CCN GTM ATG
	CCT GCC RAY GGT CCN GTK ATG
Single Letter Amino Acid IUPAC Code	P A D/N G P V M (Positions: 131–137)

Table C. Primers used on GV whole genome DNA designed using the *Olindias formosus* GFP gene.

Primer name	Primer Sequence (5'-3')
Assay Set 1 (Olindias GFP) F	ACG AGC ATT GTT CCA GTA TCC
Assay Set 1 (Olindias GFP) R	ACG TGG CTC AGA AGT CTT ATT C
Assay Set 2 (Olindias GFP) F	CCA CAT GTT CGC GCA TTA TC
Assay Set 2 (Olindias GFP) R	CGA GTA TCA TTG GGA TCC TTC TT
Assay Set 3 (Olindias GFP) F	CCA ATG TCC TGG GTA GTG ATT
Assay Set 3 (Olindias GFP) R	TTA TAC TGC CGA TGC TGA AGG
Assay Set 4 (Olindias GFP) F	TTC AAG GTT GCT GGA GAA GG
Assay Set 4 (Olindias GFP) R	TCT GTG TGT GCA GGA AAT GAT A
Assay Set 5 (Olindias GFP) F	TTG CTT CGC CTC TGC AAT A
Assay Set 5 (Olindias GFP) R	CTC GGT CTG AAC AAT GTG ATC T

Table D. Primers used on GV whole genome DNA designed using the *Aequorea victoria* GFP gene.

Primer name	Primer sequence (5'-3')
wtGFP F	ATG GCT TCC GGA CGA GCA TTG TT
wtGFP R	GCT GGC AAC CCA TGG CAT GAG CCT CAG CAT CGG CAG TAT AA
wtGFP Chromophore F	GGA GAT CTG CCA ATG TCC TGG
wtGFP Chromophore R	AAG GAT CCC AAT GAT ACT CGA GAT CAC

3. Isolation of the GvFP Gene from the Transcriptome of *G. vertens*

The inability to directly amplify the GvFP gene from genomic DNA of *G. vertens* necessitated an alternative approach to isolate this gene. Since many researchers have found success searching for expressed genes in RNA-Seq libraries, we decided to analyze the transcriptome of this important hydrozoan to find this gene.

3.1 RNA Isolation

As detailed in Materials and Methods, five individual *G. vertens* were collected from a salt pond on Martha's Vineyard, MA in August 2017 and were selected for RNA extraction representative of the morphological diversity displayed in this population (Figure 12).

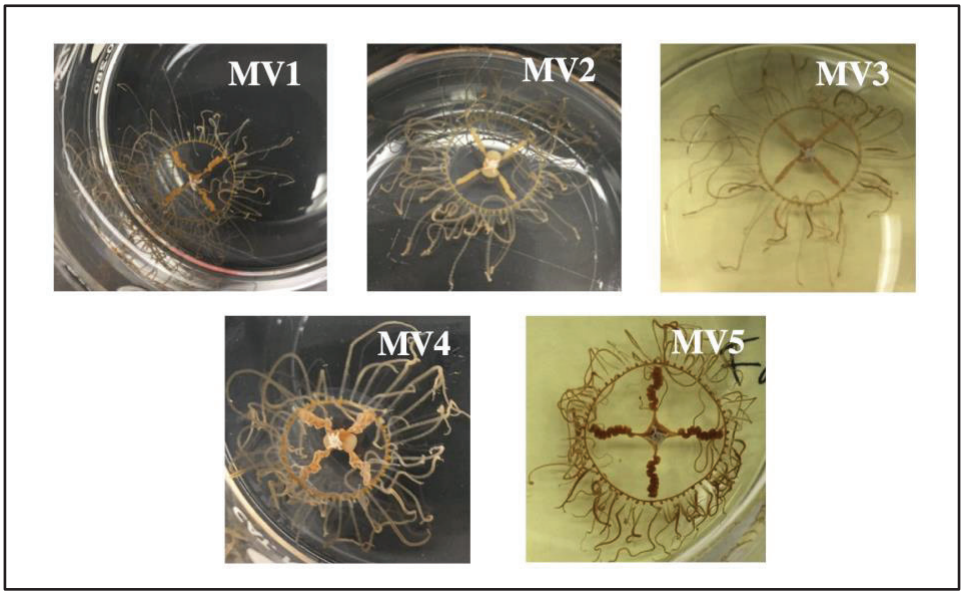


Figure 12. Five *G. vertens* selected for RNA extraction. Each individual was held in artificial seawater for 48 hours prior to extraction to ensure that stomach contents had been expelled. Several changes of seawater also ensured the removal of bycatch. Each was stored in

RNALater and extracted for total RNA as detailed in Materials & Methods. (Source: Alorah D. Bliese, personal communication, May, 2019)

Prior to RNA-Seq library construction, total extracted RNA was quantified using the Nanodrop-1000 (Thermo Scientific: Nanodrop 1000 Spectrophotometer) and the Qubit 3.0 Fluorometer (Thermo Scientific Invitrogen: Qubit 3.0 Fluorometer). See Table E.

RNA Sample	ND-1000 reading (ng/µL)	OD $_{260/280}$	Qubit 3.0 Fluorometer (ng/µL)	Total RNA Yield
MV1	323.8	2.15	220	11.0 µg
MV2	557.3	2.14	556	27.8 µg
MV3	625.9	2.13	584	29.2 µg
MV4	850.5	2.13	870	43.5 µg
MV5	307.1	2.12	292	14.6 µg

Table E. Total yields based on Qubit 3.0 readings. OD $_{260/280}$ readings are from the Nanodrop (ND-1000); readings between 2.0 - 2.2 are considered "pure" and free from protein.

3.2 RNA-Seq Library Synthesis

RNA-Seq libraries were generated from total RNA by first selecting poly A+ mRNA as described in Materials & Methods. Prior to pooling and loading of libraries on the Illumina MiSeq Nextgen sequencer, each library was quantified on an Agilent 2100 High Sensitivity Bioanalyzer. Figure 13 shows a comparison of all five RNA-Seq libraries synthesized on an Agilent 2100 Bioanalyzer. Pooling of libraries normalized the quantity of libraries loaded for sequencing and were loaded at a concentration of 10 pM as suggested by the manufacturer.

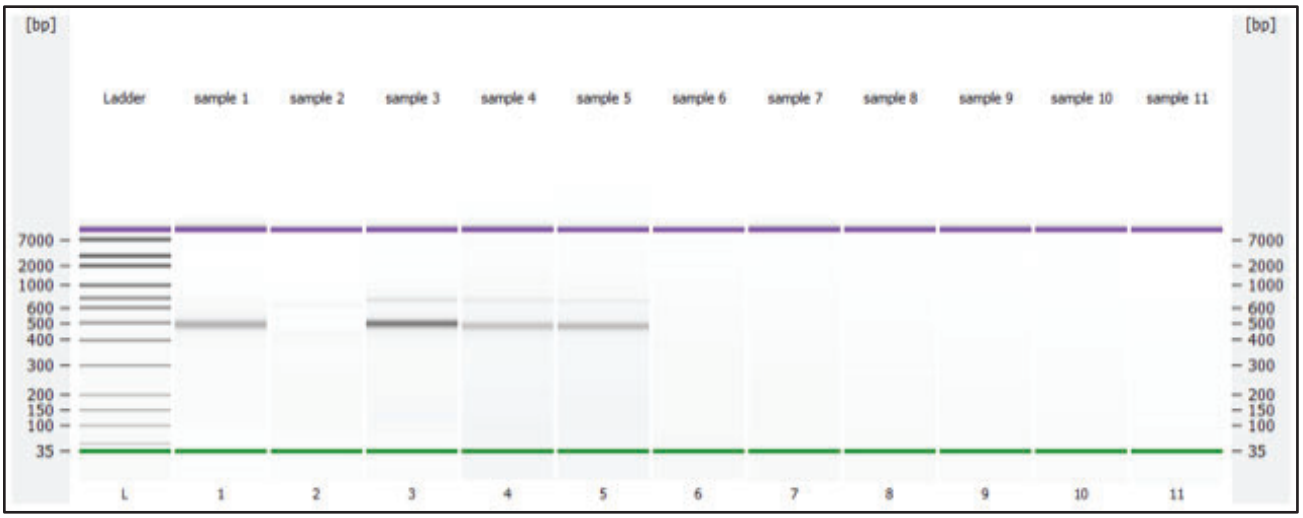


Figure 13. Agilent 2100 High Sensitivity Bioanalyzer analysis of cDNA library fragment sizes. Sample 1- MV1, sample 2- MV2, sample 3- MV3, sample 4- MV4, sample 5- MV5. Purple and green bands represent the migration of internal size markers of 35 bp (Green) and 10,380 bp (Purple) co-electrophoresed with each library on the chip.

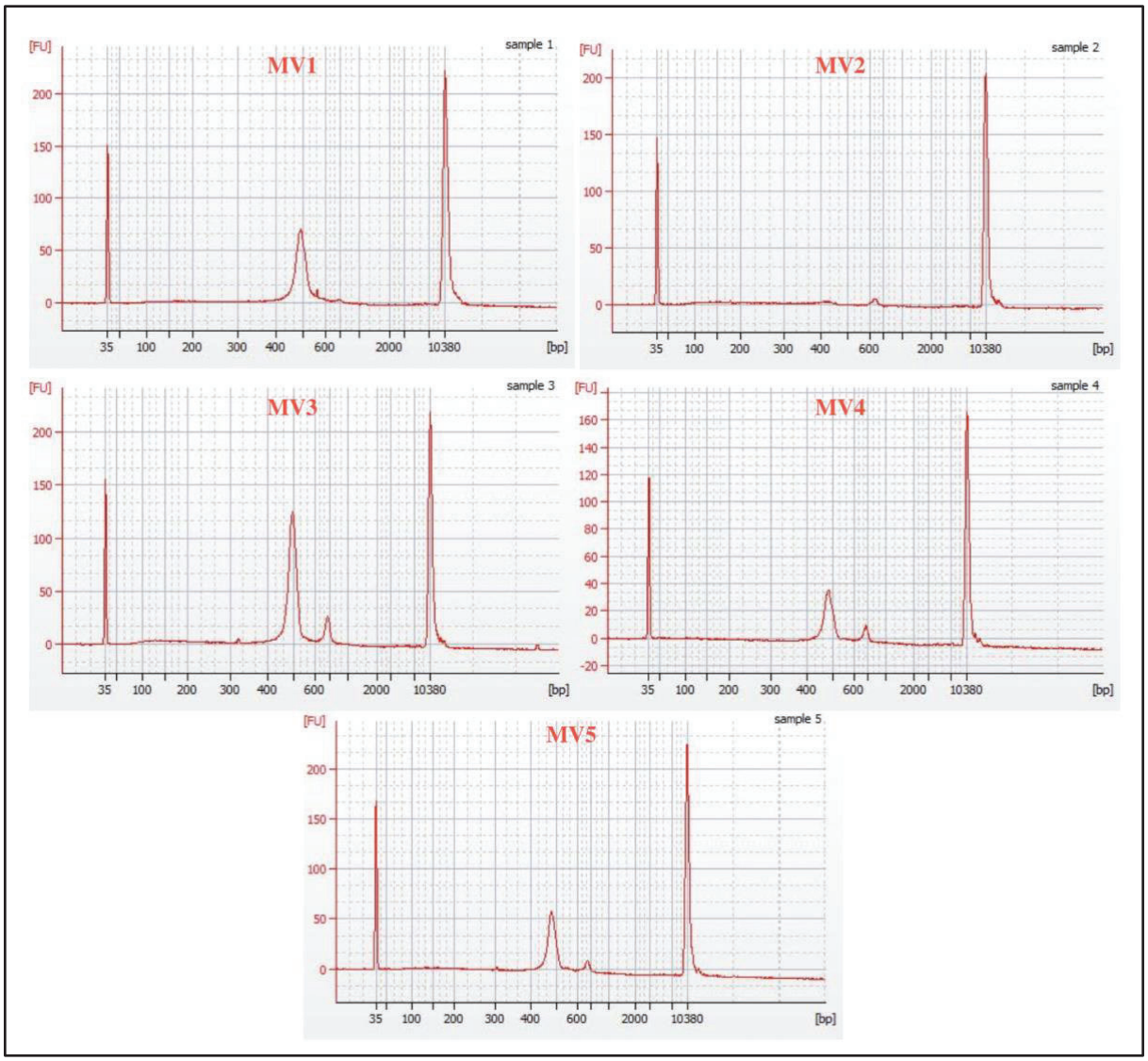


Figure 14. cDNA library fragment size distribution of each library. Plots are Size of dsDNA in bp (x-axis) *vs*. Relative Fluorescence Units (y-axis) generated by an Agilent 2100 Bioanalyzer using the High Sensitivity dsDNA Chip. Large peaks at 35 bp and 10,380 bp represent internal size markers that were co-electrophoresed with each library on the chip. Each library was between 400 and 600 bp long as expected. Only the MV2 library was low; when libraries were

pooled for sequencing, all were normalized to 10 pM from their Qubit 3.0 readings to compensate for varying levels of input DNA.

3.3 NGS Sequencing

RNA-Seq libraries (each containing a unique index) were pooled and run on an Illumina MiSeq using a V3 Reagent Cartridge. Summary statistics for the flowcell are provided in Figure 15. Over 34 million clusters were generated on the flowcell from these 5 pooled libraries at a density of >1600K clusters/mm2. The percentage of clusters passing filter (PF) was 83.66%. The percentage of reads that were greater >= Q30 was 64.70%, with a mean quality score of 29.26. Overall, the performance of the run was quite acceptable, generating more than 17.4 GB of data. From a cost perspective, a quick calculation shows that this experiment generated more than 5 million bases of sequence per US dollar.

Flowcell Summary

Clusters (Raw)	Clusters(PF)	Yield (MBases)
34,583,078	28,930,572	17,416

Lane Summary

Lane	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
1	28,930,572	100.00	98.99	1.01	17,416	83.66	64.70	29.26

Figure 15. Flowcell Statistics

Each library, encoding a unique 6 bp barcode index, can be separated out from the total library of 28,930,572 reads (that passed filter). Of this total, we chose the MARTVIN5 (MV5) library for our initial analysis since it represented 26.19% of all reads generated, more than the

other libraries in this NGS run (see Table F. Lane Summary data below) and had a >=Q30 of 64.33%.

Lane Summary

Lane	Project	Sample	Barcode sequence	PF Clusters	% of the lane	% Perfect barcode	% One mismatch barcode	Yield (Mbases)	% PF Clusters	% >= Q30 bases	Mean Quality Score
1	default	MARTVIN1	CGATGT	6,831,999	23.62	99.09	0.91	4,113	100.00	63.61	28.98
1	default	MARTVIN2	TGACCA	3,915,119	13.53	98.92	1.08	2,357	100.00	70.18	30.70
1	default	MARTVIN3	ACAGTG	5,192,371	17.95	98.97	1.03	3,126	100.00	65.90	29.55
1	default	MARTVIN4	GCCAAT	5,144,483	17.78	98.70	1.30	3,097	100.00	61.53	28.46
1	default	MARTVIN5	CTTGTA	7,575,933	26.19	99.13	0.87	4,561	100.00	64.33	29.15
1	default	Undetermined	*unknown*	270,667	0.94	100.00	NaN	163	4.57	60.71	28.40

Table F. Lane Summary Statistics for NGS Transcriptome Run

Although all five libraries were pooled to contain the same molar input to the flowcell (10 pM each), there was some variation seen in the percentages of each library represented in the whole. We expected to see *ca*. 20% of each library in the final set of sequence reads, but this varied from a low of 13.52% (MV2) to a high of 26.19% (MV5). Despite this, all were equally good in quality and reliability of the data. Figure 16 graphically illustrates the size distribution of NGS reads for the MV5 library.

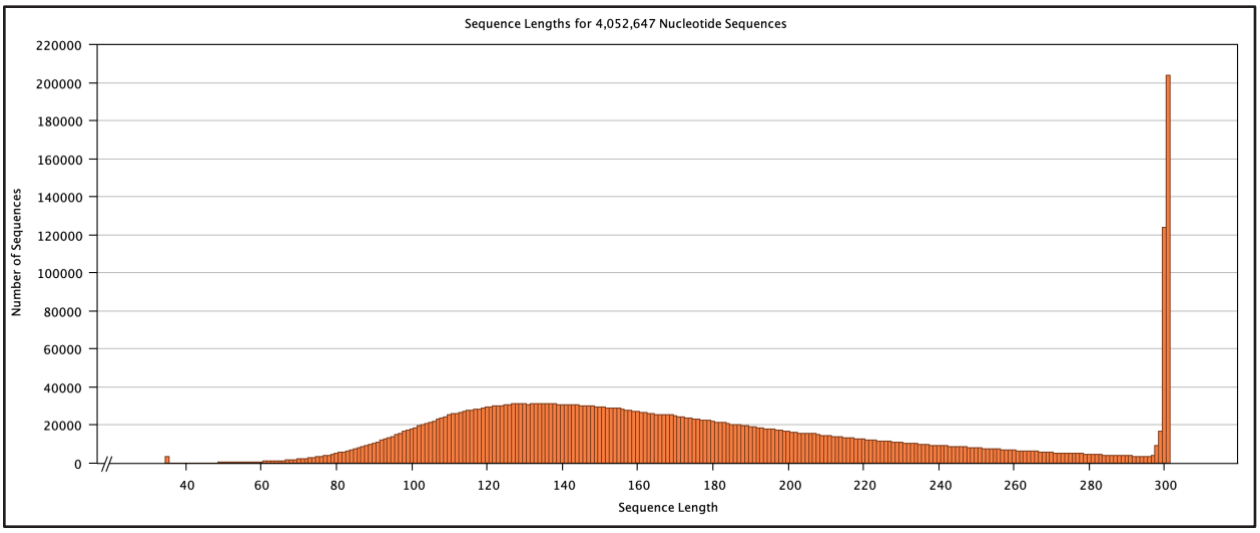


Figure 16. Sequence lengths for the 4,052,647 reads from the NGS run.

4. *De novo* Assembly of RNA-Seq Library MV5

Since the MV5 library showed the largest number of reads by NGS, it was the first library we attempted to build using the Geneious 10.1 Assembler as described in Materials & Methods. The statistics for this *de novo* assembly is detailed in Table G. The statistics for the Geneious assembly quality scores are found in Table H.

Table G. Geneious Assembly Statistics

Assembler	Geneious (Default)
Contigs in Geneious Assembly	39,732
Total Bases	16,883,364 nt
Time for Assembly	42.3 hr
Mean Length of Contigs	424.9 nt
Std Dev of Contig Length	347.5 nt
Minimum Contig Length	57 nt
Maximum Contig Length	8485 nt

Table H. Geneious Assembly Quality Scores

Confidence Mean	281.5
Expected Errors	15,764
At least Q20	98.9%
At least Q30	97.8%
At least Q40	70.4%

The length graph of the RNA-Seq assembly for the MV5 library using Geneious assembler can be seen in Figure 17. The maximum contig was 8,485 nt, the minimum contig was 57 nt. The mean contig length was 424.9 nt with a standard deviation (SD) of 347.5.

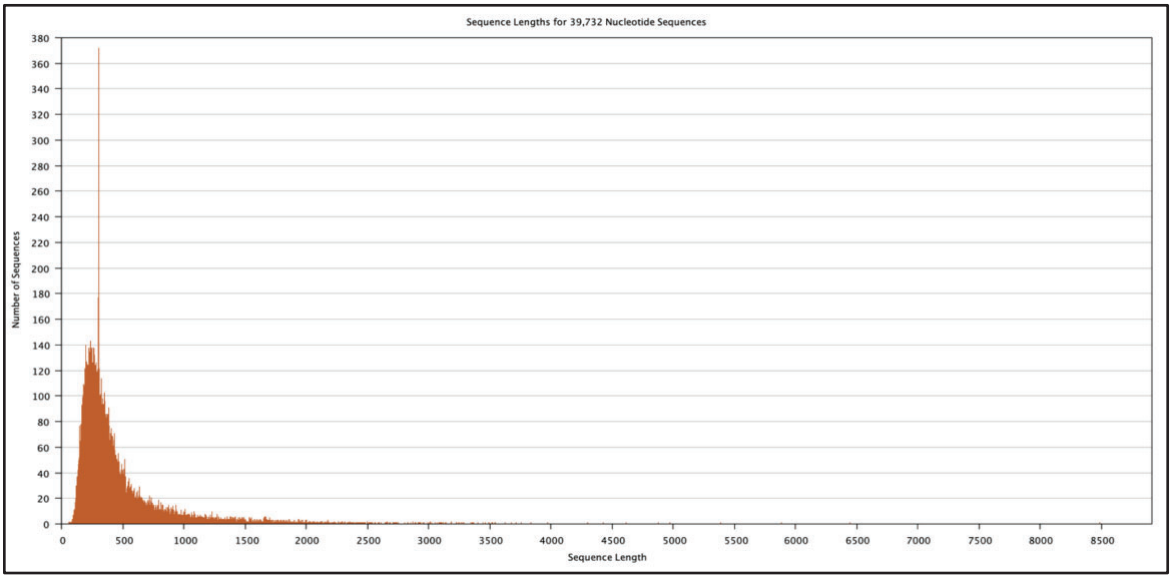


Figure 17. Length Graph of RNA-Seq Assembly of MV5 using Geneious assembler. Maximum contig was 8,485 nt; Minimum contig was 57 nt. Mean contig length = 424.9 nt; SD = 347.5.

Table I details the base frequency found in this dataset. The percentages of all four common nucleotides, as well as the percentages of all degenerate bases (using the IUB Code in Appendix D) are expressed. The GC content of this transcriptome is 43.0%.

Read (nt)	Frequency	%	% non-ambiguous
A	4,827,904	28.6%	28.7%
C	3,950,173	23.4%	23.5%
G	3,290,717	19.5%	19.6%
T	4,762,493	28.2%	28.3%
Y	14,630	0.1%	0%
W	7,001	0.0%	0%
V	46	0.0%	0%
S	3,856	0.0%	0%
R	13,653	0.1%	0%
N	926	0.0%	0%
M	6,530	0.0%	0%
K	5,250	0.0%	0%
H	74	0.0%	0%
D	59	0.0%	0%
B	52	0.0%	0%
GC	7,244,746	43.0%	43.0%
TOTAL	16,833,364	100%	100%

Table I. Frequency of nucleotides and degenerate bases.

5. Discovery of Putative GvFP in Assembled Contigs

The Geneious assembly generated 39,732 unique contigs from the 4,052,647 reads from our NGS run of MV5 (Table G). Each contig theoretically represents an mRNA molecule that existed in the organism at the time it was collected. The strategy employed to locate putative GFP genes in this library was to translate each contig, in all 6 possible reading frames (3 forward frames and 3 reverse frames), and to search for proteins that contain a conserved motif associated with the chromophore found in nearly all GFP's. One of the best conserved regions of most GFP identified is the tripeptide SYG (serine-tyrosine-glycine). This chromophore is usually found at positions 64 through 66 from the amino terminal end of the protein chain. In particular, the latter two amino acids - YG - are most highly conserved and it is their side chains that autocatalyze to form the unique chromophore of GFP's (Remmington, 2011; Zagranichny *et al*., 2004). So, Geneious was used to generate all possible proteins from this set of contigs, a total of 238,392 protein sequences. These were then searched for the dipeptide YG (tyrosine-glycine) and positive hits were collected in a separate file. Of these YG containing peptides, each was searched using the BLASTx algorithm against all known Cnidarian proteins in GenBank. Using this strategy, three peptides were found that had an 86% match to the green fluorescent protein (ScSuFP) from the hydrozoan ***Sc****olionema* ***su****vaense* discovered by Horiuchi *et al.* in 2018. These 3 contigs were named Contig 674, Contig 6046, and Contig 6770, relating to their position in the Geneious assembly of contigs (Figure 18). Two of these three hits were full-length (contig 674 and contig 6770) based on the presence of an in-frame start (ATG) and stop codon (TAG) predicting a protein of the expected size. These two contigs were used in all future analyses. To differentiate the GFP from *Gonionemus vertens* from others, we propose to name this GvFP (for ***G****onionemus* ***v****ertens*

Fluorescent **P**rotein). This novel GvFP has a MYG tripeptide chromophore that is unique among hydrozoans. This MYG chromophore has only been reported in one other GFP from the sea anemone *Anomonea sulcata* (Zagranichny *et al.*, 2004).

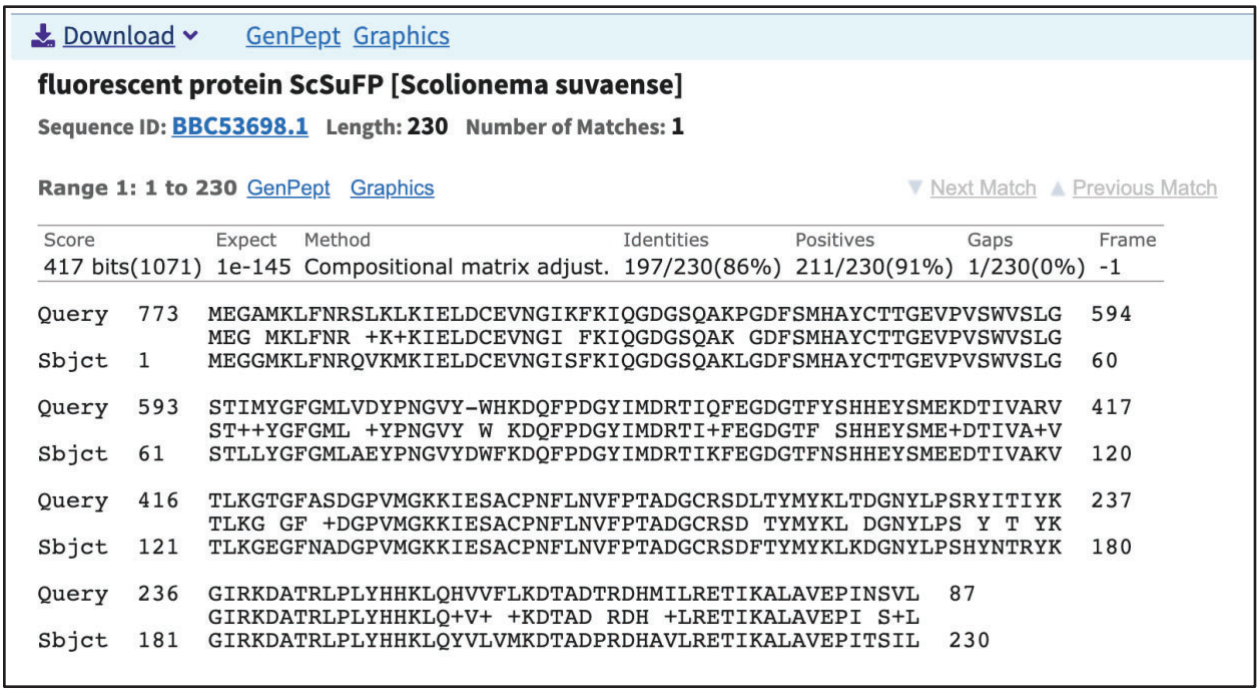
>674 MV5 assembly
CGATTCTTCATATAAACAATCATCAAGATAATTACATCCTTGTCACAGATTCAGGGCAAAATAATTAAAGCCCTTCTTGATGTATTTTCCATCATCTTTAATACCTCTAATGATTAATAGTTT
GTTTGTTTTGGGGGTTTCAGCCACCGTAGGTTTTGTGCAAGAGTATCAGCGCCCAACCATATCCCTGGTCTTCTTCTTTCTCCTGTCTTGACTGATCAATCAATACAGAACAAAGAC
GAGGCCCAGGGACTAGGCAAGCACTAGTACTCTTGAACAAAACTYAGGTTTACTGAAAACCCCCCAAAACAAACTAATTATTAACCATGAACAGTCAAGTATCCTACAGTCAATTAAA
CCTCTAATGAATTTCTCTTCATTTGAAACTGAAATTTCCTATCAGAAGGAAGTCTACCAGGAACAAATATACTCGTAGATCATCATCGGCGACATGTTTACAGGGCAGAGTTGATGGGT
TCAACAGCTAAAGCTTTGATAGTTTCCCTCAAGATCATATGATCGCGAGTGTCAGCAGTATCTTTCAAGAACACAACATGCTGTAGTTTGTGGTGGTACAGAGGCAGACGGGTGGG
GTCCTTGCGGATTCCTTTGTAGGTAGTGATGTAGCGAGAGGGGAAATAACTGCCATCAGTAAGCTTGTACATGTAAGTTAAGTCGGATCGGCAACCACCGGCAATGGGGTATACATT
CAAGAAATTAGGACAAGCTGATTCGATTCTCTTTCCCATGACTGGTCCATCGGCGGCAAATCCTGTACCTTTTAGGGTAACCCTCGCTACAATGGTGTCTTCTTCCATCGAGTATTCA
TGGTGAGAATTAAAGGTTCCATCCCCTTCGAACTGAATTGTACGGTCCATGATGTAACCATCGGGGAACTGGTCTTTATGCCAATCATAAACACCGTTGGGGTAATCCACTAACATGC
CAAATCCATACATAATCGTTGAACCAAGTGACACCCATGATACAGGAACTTCTCCAGTCGTACAGTAGGCATGCACGGTGAAGTCACCAGGTTTGGCTTGAGATCCATCGCCTTGTA
TCTTGAATTCGATGTCGTTTACTTCACAGTCCAACTCAATCTTCAATTTTAACGATCGATTGAAGAGTTTCATGCCGCCTTCCATTGCAAAGCTAGCAATCTTGCGAACTTTGAACTAA
AGAAATGAAGTCTTGCAGTGGTGATGTCTGGATGCCAGTTCAGTTC

>6046 MV5 assembly
CAGTATCTTTCAAGAACACAACATGCTGTAGTTTGTGGTGGTACAGAGGCAGACGGGTGGCATCCTTGCGGATTCCTTTGTAGATAGTGATGTAGTGAGAGGGCAAATAGTTGCCA
TCAGTAAGCTTGTACATGTAAGTGAGGTCGGATCGGCAACCATCGGCAGTGGGGAACACATTCAAGAAATTTGGACAGGCTGATTCGATTTTCTTTCCCATGACTGGTCCGTCAGA
AGCAAATCCTGTACCTTTCAGGGTAACCTTCGAAATAATGGTGTCTTCTTCCATGGAGTATTCATGGTGAGAATAGAAGGTTCCATCCCCTTCGAACTGAATTGTACGRTCCATGATG
TAACCATCGGGGAACTGGTCTTTATGCCAATCATAAACACCGTTGGGGTAATCCACTAACATGCCAAATCCATACATAATAGTTGAACCAAGAGACACCCATGATACAGGAACTTCTC
CAGTCGTACAGTAGGCATGCATGGAGAAGTCACCAGGTTTGGCTTGAGAGCCATCGCCTTGTATCTTGAATTTGATGCCGTTCACTTCACAGTCCAACTCAATCTTCAATTTCAACG
ATCGATTGAAGAGTTTCATGGCGCCTTCCATTCCTAAGCAGTCTTGTGCACTTTGAACGGGAATCGAAAGAAAATGAAGTCTTGCAGAGGTGATGTCCGGATGCCAG

>6770 MV5 assembly
CACTTTATTTGAAACTGATATTTCCTATCAAAAGGAAGCGTACAAGAACCAAATACACTCGCMGATTATCATCAGCGACATGGTTACAACACAGAGTTGATGGGCTCAACGGCTAAAG
CTTTGATAGTTTCCCTCAAGATCATATGATCGCGAGTGTCAGCAGTATCTTTCAAGAACACAACATGCTGTAGTTTGTGGTGGTATAGAGGCAGACGGGTGGCATCCTTGCGGATTC
CTTTGTAGATAGTGATGTATCGAGAGGGCAAATAGTTGCCATCAGTAAGCTTGTACATGTAAGTGAGGTCGGATCGGCAACCATCGGCAGTGGGGAATACATTCAAGAAATTTGGAC
ATGCTGATTCGATTTTCTTTCCCATGACTGGTCCGTCAGAAGCAAATCCTGTACCTTTCAGGGTAACTCTCGCTACAATAGTGTCTTTTTCCATGGAGTATTCATGGTGAGAATAGAA
GGTTCCATCCCCTTCGAACTGAATTGTACGGTCCATGATGTAACCATCGGGGAACTGGTCTTTATGCCAATAAACACCGTTGGGGTAATCCACTAACATGCCAAATCCATACATAATA
GTTGAACCAAGAGACACCCATGATACAGGAACTTCTCCAGTCGTACAGTAGGCATGCATGGAGAAGTCACCAGGTTTGGCTTGAGAGCCATCGCCTTGTATCTTGAATTTGATGCC
GTTCACTTCACAGTCCAACTCAATCTTCAATTTCAACGATCGATTGAAGAGTTTCATGGCGCCTTCCATTCCTAAGCAGTCTTGTGCACTTTGAACGGGAATCTTATATTTTCT

Figure 18. Raw sequences of contigs 6770, 674, and 6046.

A BLASTx of these contigs on GenBank produces a remarkable match to a fluorescent protein (ScSuFP) from the hydrozoan *Scolionema suvaense*. Figure 19 shows the BLASTx match of our Contig 6770 GvFP protein. A comparison shows an identity of 86% (197/230) and a positive score of 91% (211/230), with one gap in our query sequence. The e-value was 1e-145. Similarly, impressive values are generated when Contig 6770 is compared. There is no doubt that these proteins (and the genes encoding these proteins) are highly homologous and closely related.



Download GenPept Graphics

fluorescent protein ScSuFP [Scolionema suvaense]

Sequence ID: BBC53698.1 Length: 230 Number of Matches: 1

Range 1: 1 to 230 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
417 bits(1071)	1e-145	Compositional matrix adjust.	197/230(86%)	211/230(91%)	1/230(0%)	-1

```
Query  773  MEGAMKLFNRSLKLKIELDCEVNGIKFKIQGDGSQAKPGDFSMHAYCTTGEVPVSWVSLG  594
            MEG MKLFNR +K+KIELDCEVNGI FKIQGDGSQAK GDFSMHAYCTTGEVPVSWVSLG
Sbjct  1    MEGGMKLFNRQVKMKIELDCEVNGISFKIQGDGSQAKLGDFSMHAYCTTGEVPVSWVSLG  60

Query  593  STIMYGFGMLVDYPNGVY-WHKDQFPDGYIMDRTIQFEGDGTFYSHHEYSMEKDTIVARV  417
            ST++YGFGML +YPNGVY W KDQFPDGYIMDRTI+FEGDGTF SHHEYSME+DTIVA+V
Sbjct  61   STLLYGFGMLAEYPNGVYDWFKDQFPDGYIMDRTIKFEGDGTFNSHHEYSMEEDTIVAKV  120

Query  416  TLKGTGFASDGPVMGKKIESACPNFLNVFPTADGCRSDLTYMYKLTDGNYLPSRYITIYK  237
            TLKG GF +DGPVMGKKIESACPNFLNVFPTADGCRSD TYMYKL DGNYLPS Y T YK
Sbjct  121  TLKGEGFNADGPVMGKKIESACPNFLNVFPTADGCRSDFTYMYKLKDGNYLPSHYNTRYK  180

Query  236  GIRKDATRLPLYHHKLQHVVFLKDTADTRDHMILRETIKALAVEPINSVL  87
            GIRKDATRLPLYHHKLQ+V+ +KDTAD RDH +LRETIKALAVEPI S+L
Sbjct  181  GIRKDATRLPLYHHKLQYVLVMKDTADPRDHAVLRETIKALAVEPITSIL  230
```

Figure 19. BlastX (https://blast.ncbi.nlm.nih.gov/Blast.cgi?LINK_LOC=blasthome) search result of contiguous sequence 6770 with an E-value of 1e-145, and an 86% identity to ScSuFP (*Scolionema suvaense*).

6. Characterization of GvFP Protein from *Gonionemus vertens*

The MW (molecular weight) and the pI (isoelectric point) of our GvFP protein was predicted using the ExPASy Compute pI/Mw Tool (https://web.expasy.org/compute_pi/). Compute pI/Mw predicts the molecular weight of GvFP (for contig 6770) to be nearly 26kDa (25,945.53 Da) making it slightly less than the molecular weight of wtGFP from *Aequorea victoria* which is approximately 27kDa (Hink *et al.*, 2000). The isoelectric point refers to the pH level at which the protein has a neutral charge. The pI of our GvFP is 5.74. Slightly acidic pI's are typical for cytosolic proteins.

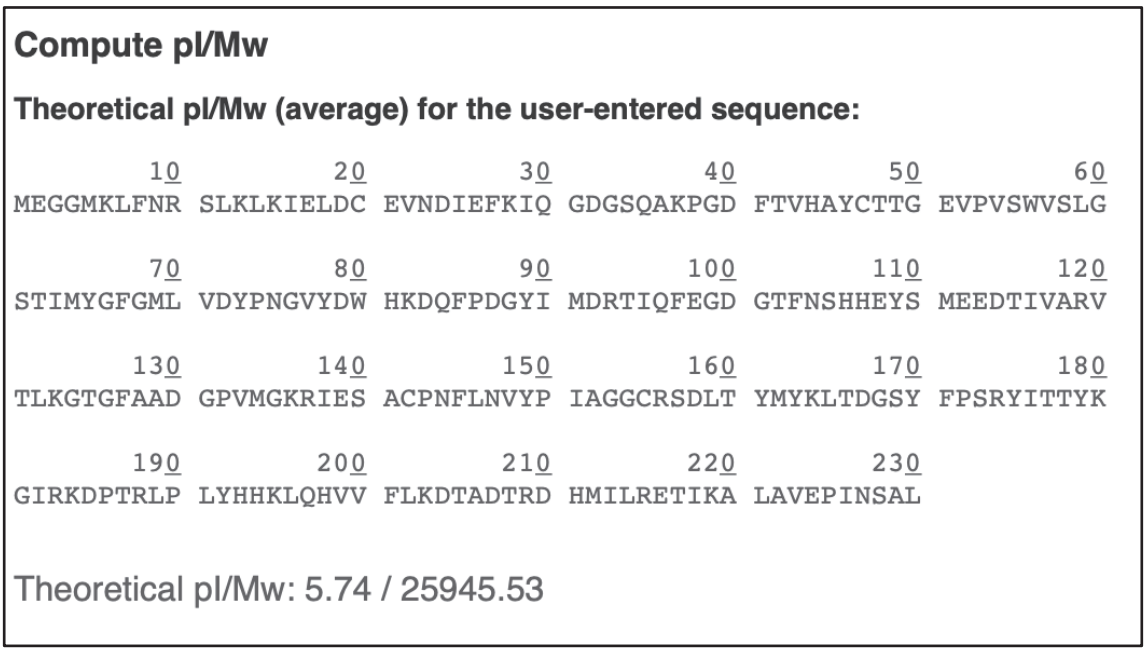


Figure 20. Predicted pI and Mw of GvFP. (Source: https://web.expasy.org/compute_pi/)

SignalP-5.0 (http://www.cbs.dtu.dk/services/SignalP/) was used to analyze the N-terminus of the GvFP peptide for hydrophobic residues and predict if this peptide contains a signal peptide involved in intracellular localization. At their N-terminus, signal peptides have a span of 16-30 amino acids which are usually hydrophilic, and positively charged, with a central hydrophobic amino acid region, and a C-terminus with a signal peptidase cleavage site. Analysis by SignalP-5.0 demonstrated a 0% probability that this putative amino terminus contained a signal peptide (Figure 21).

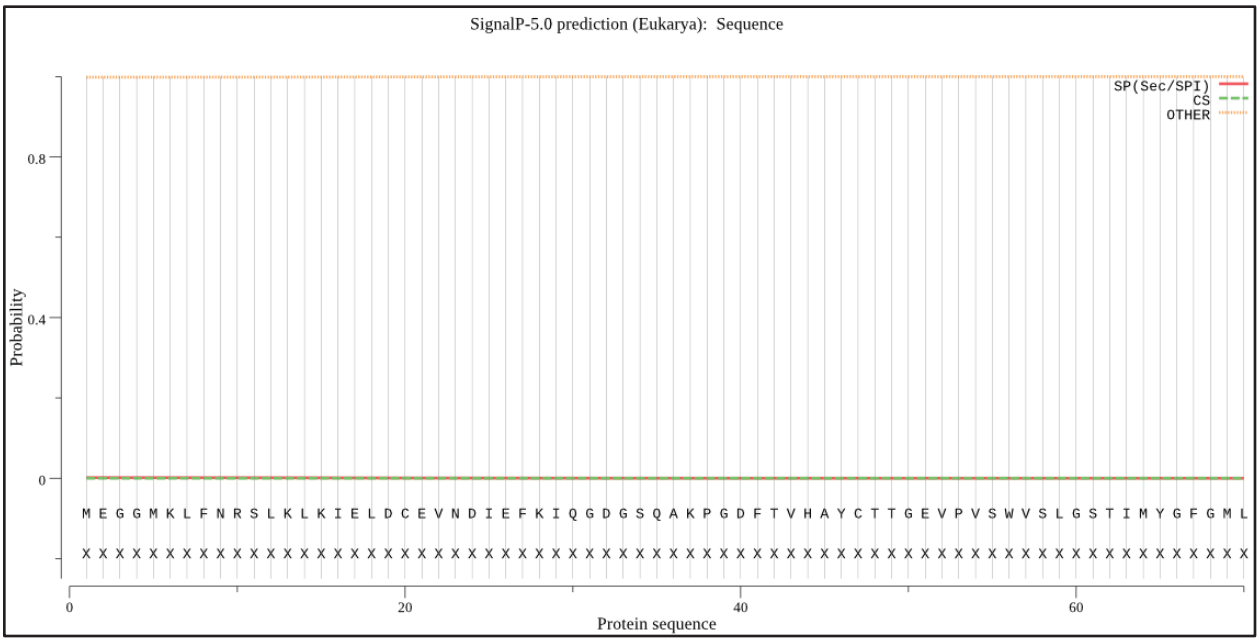


Figure 21. SignalP-5.0 predicts that the GvFP sequence has a 0% probability of harboring a signal peptide sequence.

Likewise, the companion algorithm, DeepLoc-1.0, was used to predict probable intracellular localization of this protein based upon primary sequence data in our GvFP protein (http://www.cbs.dtu.dk/services/DeepLoc/). This program predicted a 95.7% probability that this protein would be cytosolic in nature, and not associated with membranes or within any subcellular compartment (Figure 22).

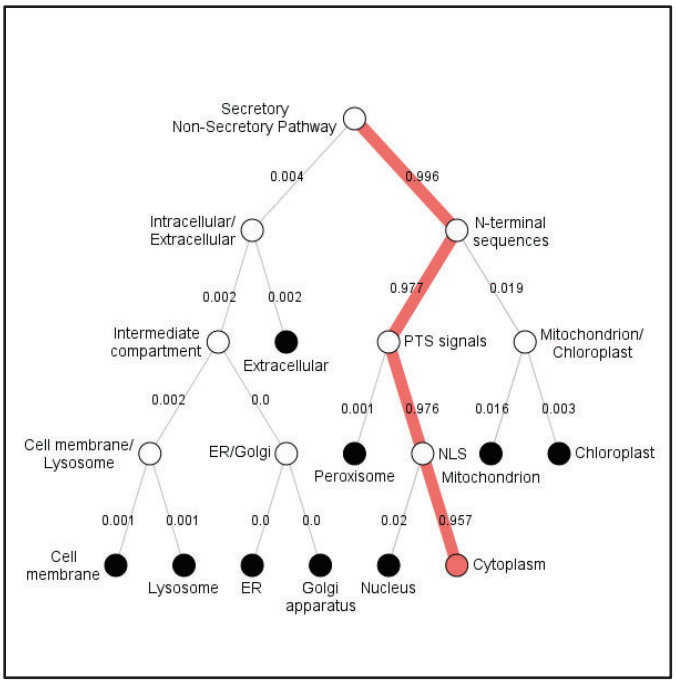


Figure 22. DeepLoc-1.0 analysis of the GvFP primary amino acid sequence predicts that this protein will be localized to the cytoplasm of the cell (http://www.cbs.dtu.dk/services/DeepLoc/).

6.1 Predicted 3-D Structure Using PHYRE2

After translating the contig 6770 nucleotide sequence into an amino acid sequence using EMBOSS Transeq (https://www.ebi.ac.uk/Tools/st/) we wanted to know how our GvFP protein compared to the standard folded wtGFP structure. The contig 6770 amino acid sequence was imported into the PHYRE2: Protein Fold Recognition Server (http://www.sbg.bio.ic.ac.uk/phyre2) and the predicted PDB structure file produced by PHYRE2 was uploaded to the EzMOL1.3 (http://www.sbg.bio.ic.ac.uk/~ezmol/) molecule display wizard in order to obtain high quality images of the predicted protein structure. The model of our GvFP (based on the 6770 contig) can be seen in Figure 23.

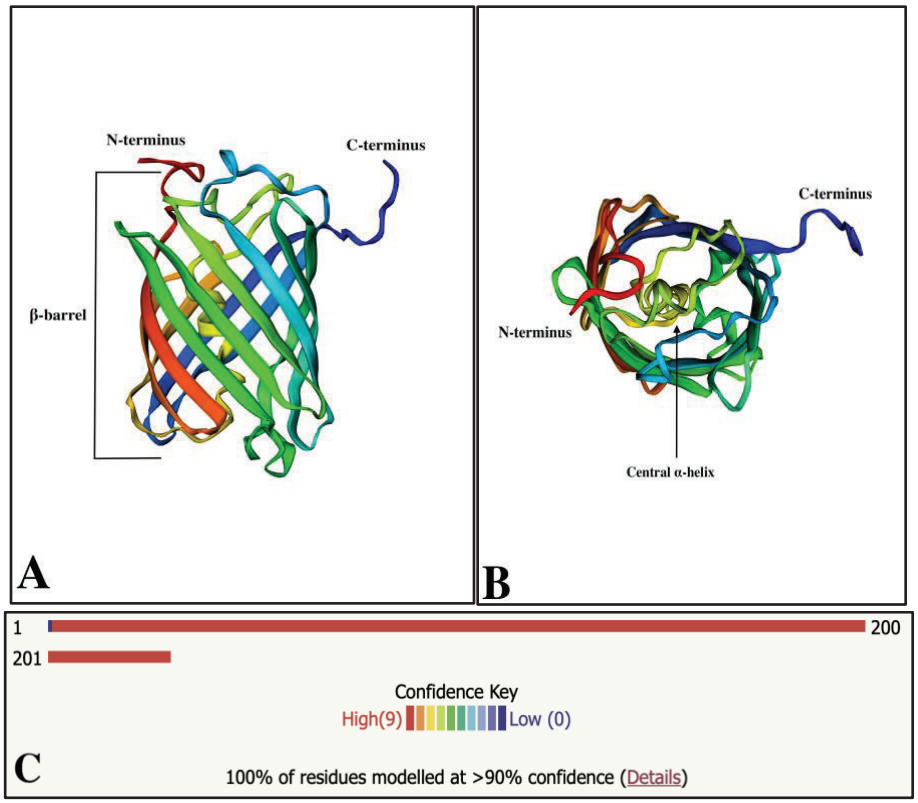


Figure 23. A: Side view of predicted protein structure. B: Top view of predicted protein structure with β-can surrounding internal alpha helix. C: The confidence key indicates that PHYRE2 modeled the majority of this sequence with high confidence (above 90% confidence).

The structure of wtGFP is iconic and one of the best recognized of all proteins. It has a β-barrel or β-can motif constructed by wrapping the barrel with 11 alternating antiparallel β-sheets stabilized by hydrogen bonding to one another. This barrel is open at the top and bottom of the structure and has an alpha-helical ribbon coiled on the inside. Three adjacent amino acids, typically SYG (at positions 64, 65, and 66), auto-catalyze when the protein folds *in vivo*, to generate the green fluorescent chromophore. This chromophore is entirely protected within the barrel.

Although our protein is only 43% homologous to the wtGFP from *A. victoria*, PHYRE2 predicts a three-dimensional structure that is virtually indistinguishable from wtGFP. GvFP has a predicted β-barrel conformation and a 2_o structure of 55% β-sheet, 7% alpha-helix, and 13% disordered structure. It has a molecular weight of 25.95 kDa and a pI of 5.74, in agreement with predictions made by the ExPASy Compute pI/Mw Tool. The dimensions of the protein are 6.8 nm tall, 5.4 nm wide, and 3.8 nm deep. By looking down on the top view of the 3D model you can see that this protein is not a perfect circle but instead, a somewhat flattened can structure.

6.2 Phylogeny of GvFP

In an attempt to understand how our newly discovered GvFP is related to other known Hydrozoan GFPs, three gene regions known to exist in all hydrozoan GFPs (YG at positions 66 and 67; R at position 98; and E at position 222) are analyzed in the Clustal Omega alignment seen in Figure 24 (Remmington, 2011). R98 and E222 are both members of the β-sheets that form the β-can surrounding the chromophore and known to be important in the chromophores autocatalytic cyclization. As seen in this alignment, the YG - found in all GFPs - is conserved. But overall, there does not seem to be large stretches of homology between these six hydrozoan GFP's. But this is misleading because if you look carefully at this alignment, especially between our newly discovered *G. vertens* GvFP and that from *Scolionema,* one would see a near perfect conservation of both type and position of amino acids in these two hydrozoans.

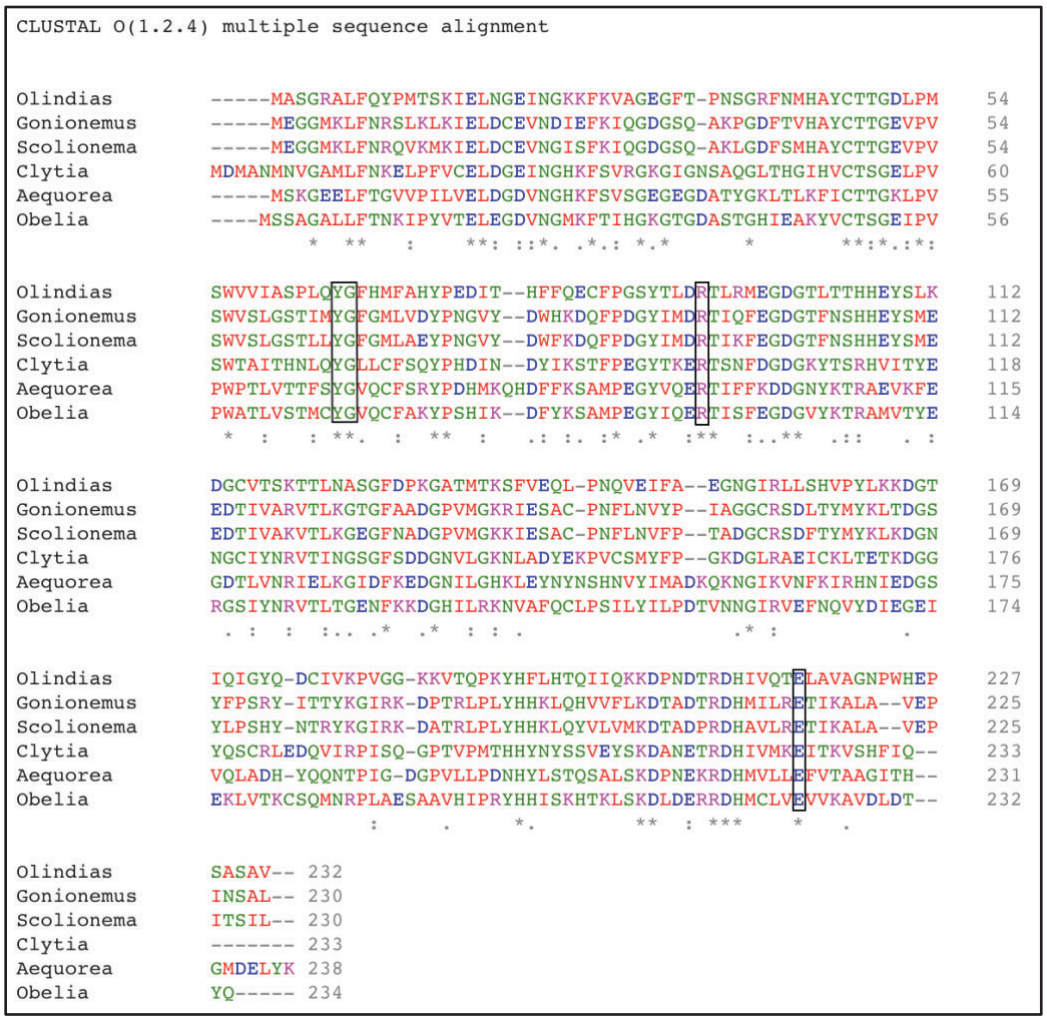


Figure 24. Clustal Omega amino acid sequence alignment of the GFP genes from the hydrozoans: *Olindias formosus* (Uniprot Entry ID: A0A2Z5X7U3), *Gonionemus vertens* (our novel GvFP), *Scolionema suvaense* (Uniprot Entry ID: A0A2Z5XV34), *Clytia hemisphaerica* (Uniprot Entry ID: J9PIH6), *Aequorea victoria* (Uniprot Entry ID: P42212), and *Obelia spp.* MH-2011 (Uniprot Entry ID: G1JSF2).

Using our predicted primary sequence of the GvFP protein and comparing to its closest homologues in GenBank, we can generate a phylogenetic tree (using the NJ algorithm) showing the relationship of our protein to others. Figure 25 shows the relationship of GvFP to the GFP of other organisms. As expected, GvFP is closest to hydrozoans, but in some cases not all hydrozoans. Note that *A. victoria* is a considerable distance away from our GvFP.

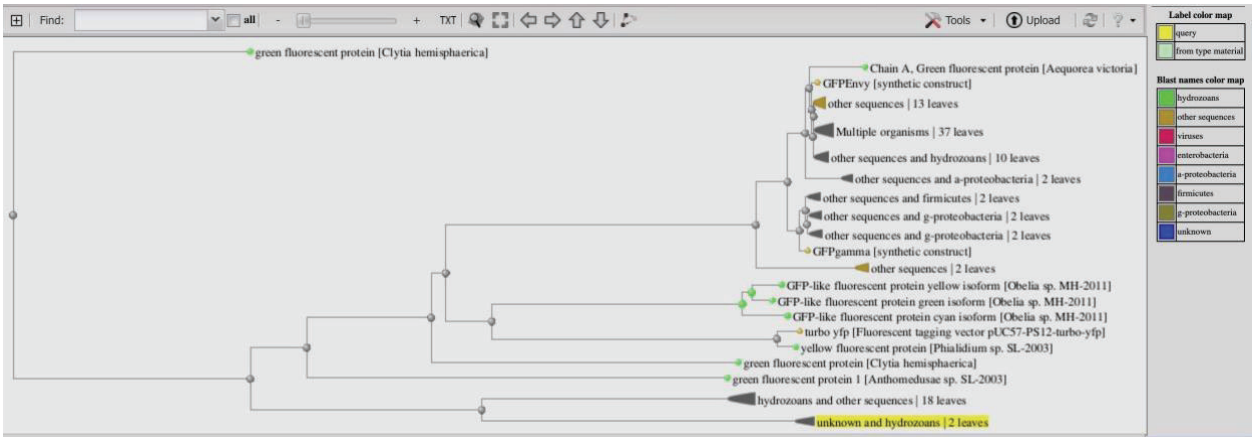


Figure 25. Phylogenetic tree graphically showing the relationship of GvFP to its nearest neighbors using the NJ (Neighbor Joining) algorithm. GvFP can be seen at the bottom of the tree in yellow.

7. Amplification of GvFP (Contig 6770) from *Gonionemus vertens* Genomic DNA

In an attempt to generate genomic clones of these GvFP constructs we generated primers that bound to both the start and end of the gene as well as to numerous internal positions within the gene (Table J). Amplification of gDNA with these primer sets are predicted to generate a known amplicon size based on the original gene sequence. Actual PCR amplicons generated are listed next to predicted amplicon sizes. As can be seen, in all cases the actual amplicon generated

was considerably larger than that predicted from the contig 6770. The most logical explanation of these results is that the genomic regions of these gene(s) must contain introns, that lengthen the gene. The difference between these two values informs us of the likely length of the introns in these genes. These were not encountered in the RNA-Seq derived sequences, since our starting material (poly A+ mRNA) did not contain introns.

Primer Name	Primer Sequence (5'-3')	Predicted Amplicon Size (bp)	Approximate Gel Band Size (bp)
GvFP Set 1 Forward	CCT GTA TCA TGG GTG TCT CTT G	434 bp	~500 bp - ~600 bp
GvFP Set 1 Reverse	TGC TGT AGT TTG TGG TAT AG	434 bp	~500 bp - ~600 bp
GvFP Set 2 Forward	CGA CTG GAG AAG TTC CTG TAT C	384 bp	~700 bp
GvFP Set 2 Reverse	TGA TGT ATC GAG AGG GCA AAT AG	384 bp	~700 bp
GvFP Set 3 Forward	GGA CCA GTC ATG GGA AAG AAA	288 bp	~1,200 bp
GvFP Set 3 Reverse	GAT GGG CTC AAC GGC TAA A	288 bp	~1,200 bp
GvFP Set 4 Forward	GTA GCG AGA GTT ACC CTG AAA G	279 bp	~1,200 bp
GvFP Set 4 Reverse	GCG AGT GTC AGC AGT AGT ATC TT	279 bp	~1,200 bp
GvFP Set 5 Forward	GCG CCA TGA AAC TCT TCA ATC	213 bp	~1,500 bp - ~2,000 bp
GvFP Set 5 Reverse	GGT AAT CCA CTA ACA TGC CAA ATC	213 bp	~1,500 bp - ~2,000 bp
GvFP Set 6 Forward	GGG TGT CTC TTG GTT CAA CTA TTA	227 bp	~4,000 bp
GvFP Set 6 Reverse	TGG TCC GTC AGA AGC AAA TC	227 bp	~4,000 bp
GvFP Set 7 Forward	CCT ACT GTA CGA CTG GAG AAG T	417 bp	No band
GvFP Set 7 Reverse	CCT TGC GGA TTC CTT TGT AGA T	417 bp	No band

GvFP Set 8 Forward	TGT GAA GTG AAC GGC ATC AA	307 bp	~1,200 bp
GvFP Set 8 Reverse	TCA GGG TAA CTC TCG CTA CAA	307 bp	~1,200 bp

Table J. Eight sets of forward and reverse primers, their sequences, and the predicted amplicon size (bp), and approximate amplicon size (bp).

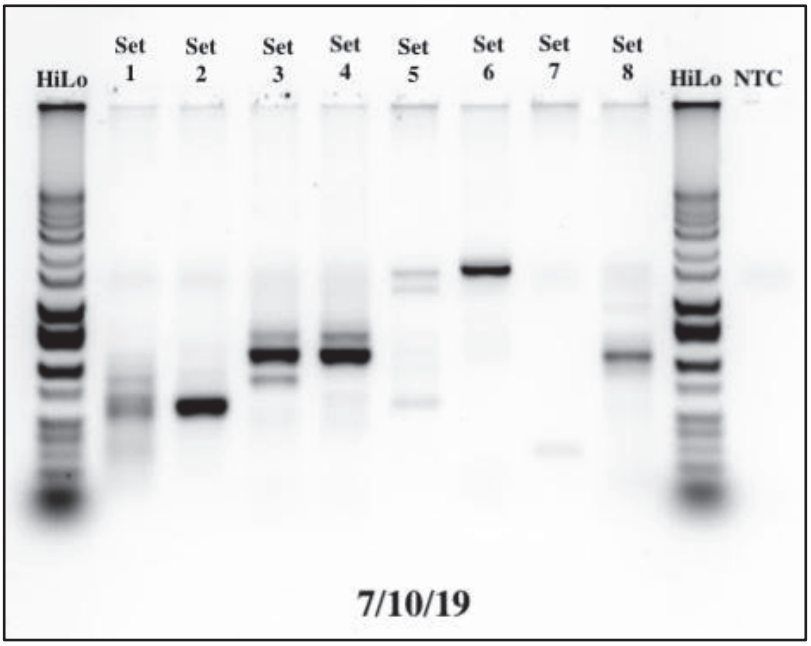


Figure 26. Separation of amplicons generated with the Contig 6770 primer sets on *Gonionemus vertens* DNA (Table J) on a 1% agarose gel. gDNA used in this experiment were from samples collected in Martha's Vineyard, MA. HiLo represents DNA size markers and NTC is our No Template Control.

8. Synthesis of Full-Length Recombinant GvFP (Contig 6770) in pET-SUMO

In order to express recombinant GvFP in *E. coli*, we needed a full-length copy of the contig 6770 that did not contain and introns. Since it is clear that the genomic clones of the GvFP gene

contained one or more introns, we turned to a synthetic route. We had the contig 6770 synthesized by IDT-DNA (Coralville, IA). We used this synthetic construct to generate large copy numbers of this gene by PCR using a pair of primers that bound to the 5' and 3' end of the contig. PCR, using a non-proof-reading thermostable Taq DNA polymerase was required here since it was imperative to have a 3' adenine attached to the amplicon. This amplicon was checked for size on an agarose gel for correct size (data not shown) and a diluted sample was sequenced by standard automated Sanger dideoxy sequencing using both the forward and reverse primers (see Appendix C for electropherogram of forward sequencing reaction). Both size analysis on agarose gels and sequence analysis verified that we had the correct, full-length version of contig 6770. With the addition of 3' adenines this amplicon could be directly ligated into a plasmid expression vector.

We chose the pET-SUMO Champion expression system since it accommodated T/A cloning of PCR-generated amplicons with 3' overhangs. Our contig 6770 was ligated into the pET-SUMO vector (Figure 27) and transformed into both Mach1-T1 and BL21(DE3) competent cells and plated on LB-Kanamycin plates as described in Materials & Methods. The Mach1-T1 *E. coli* strain is a non-expression strain only used for propagation of the plasmid and allows one to check plasmid constructs for correct orientation. The *E. coli* BL21(DE3) cells are engineered for efficient expression of recombinant fusion proteins. Since time was short at the end of this project, we directly screened two BL21(DE3) recombinants for green fluorescence. Single colonies were picked and grown up as overnight liquid cultures (37C with 200 rpm). A small aliquot of this overnight culture was inoculated into 10 mL of sterile SOB with Kanamycin. At time zero 1 mM IPTG was introduced and cultures were monitored hourly for 4 hours. A 1-mL aliquot was removed and measured for green fluorescence using a Qubit 3.0 Fluorometer. Controls were cells

transformed with an unrelated gene (CAT - chloramphenicol acetyltransferase gene) which should not express any green fluorescence. Figure 28 demonstrates that we have green fluorescence expressed in one BL21(DE3) clone harboring a recombinant pET-SUMO plasmid. This induction was performed in triplicate.

Lastly, using the cells harvested from the previous experiment at the 4-hour mark (Figure 28), cells were pelleted, washed, and lysed by sonication as described in Materials & Methods. The clarified supernatant was subjected to both emission and excitation scans on a spectrofluorometer. Figures 29 and 30 show the excitation and emission spectra, respectively, of our induced recombinant *E. coli* expressing recombinant GvFP in the clarified supernatant. In both cases, the λ_{Max} for both emission and excitation were near perfect matches to the native GvFP isolated from intact medusae (compare to Figure 11). As seen previously with native GvFP extracted from medusae, excitation of the recombinant GvFP with 90_o polarized light (@470 nm) shows a dramatic reduction in fluorescence with no change in λ_{max}. This reduction suggests that the fluorescence is again likely due to a large macromolecule (*e.g.*, protein) and not a small, rapidly diffusing compound (Lakowicz, 2006).

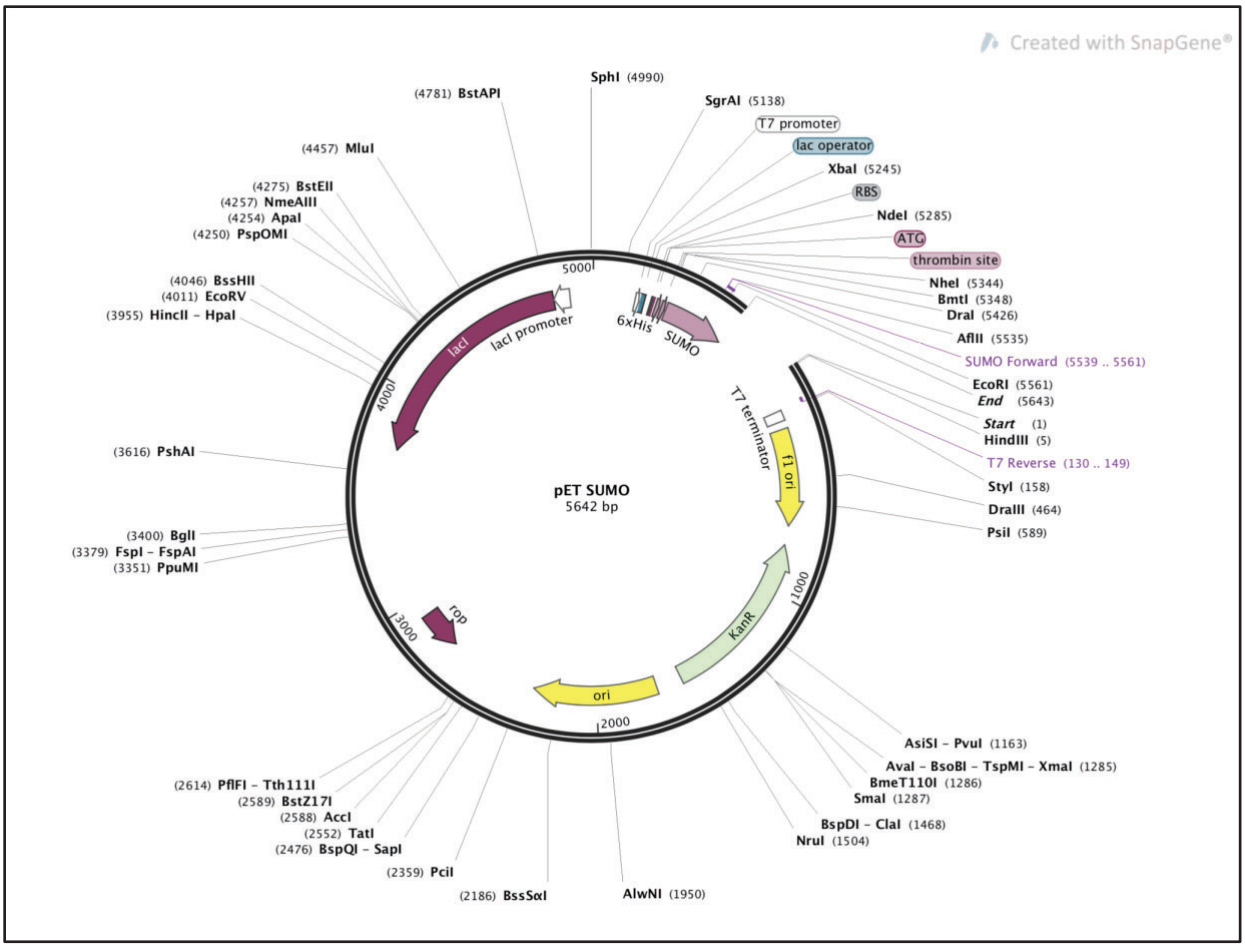


Figure 27. pET-SUMO vector used for ligation of the PCR-amplified contig 6770 gene.

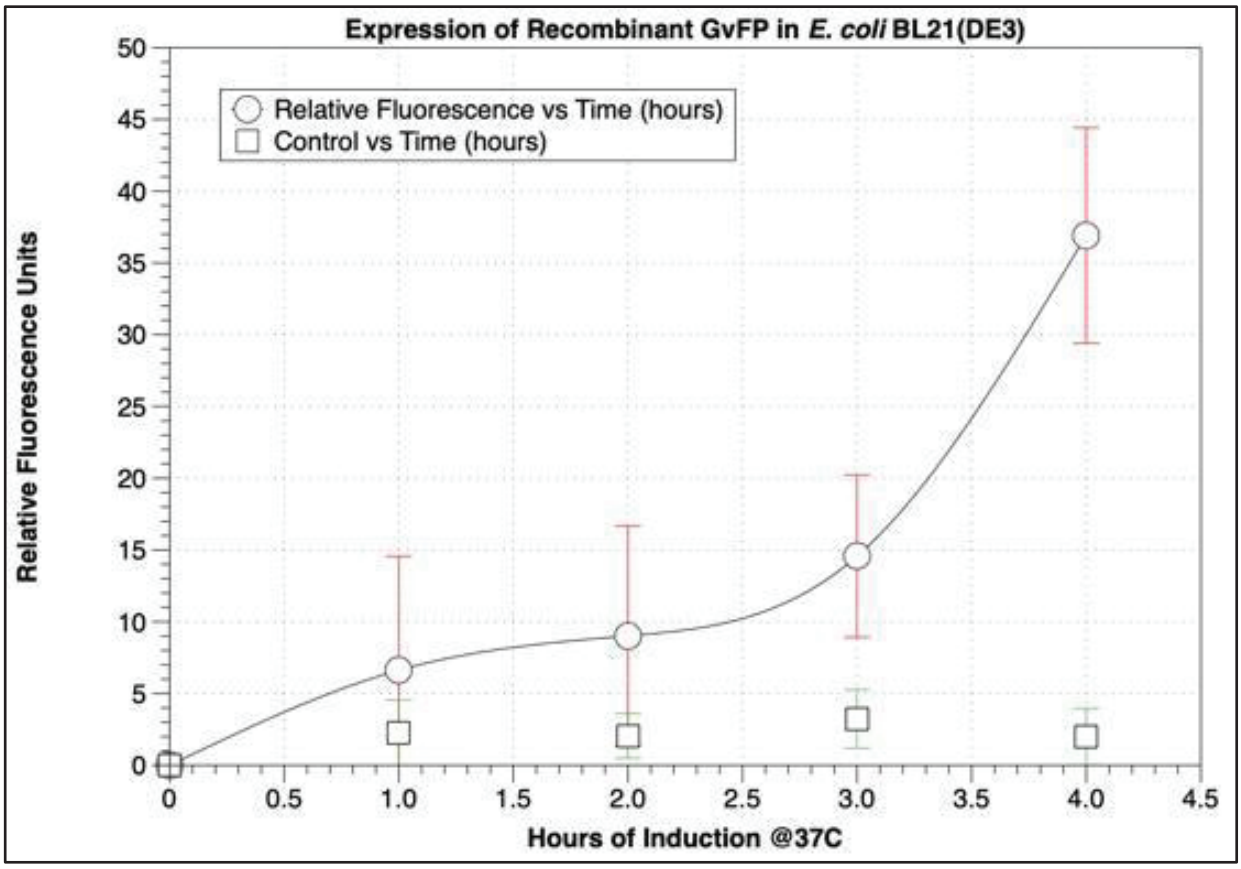


Figure 28. Expression of recombinant GvFP in *E. coli* BL21(DE3) cells measured by fluorescence in a Qubit 3.0 fluorometer after induction with IPTG at time 0. Control cells (transformed with a pET-SUMO control plasmid harboring and expressing a non-fluorescent protein) were induced in an identical fashion. All samples are measured in triplicate.

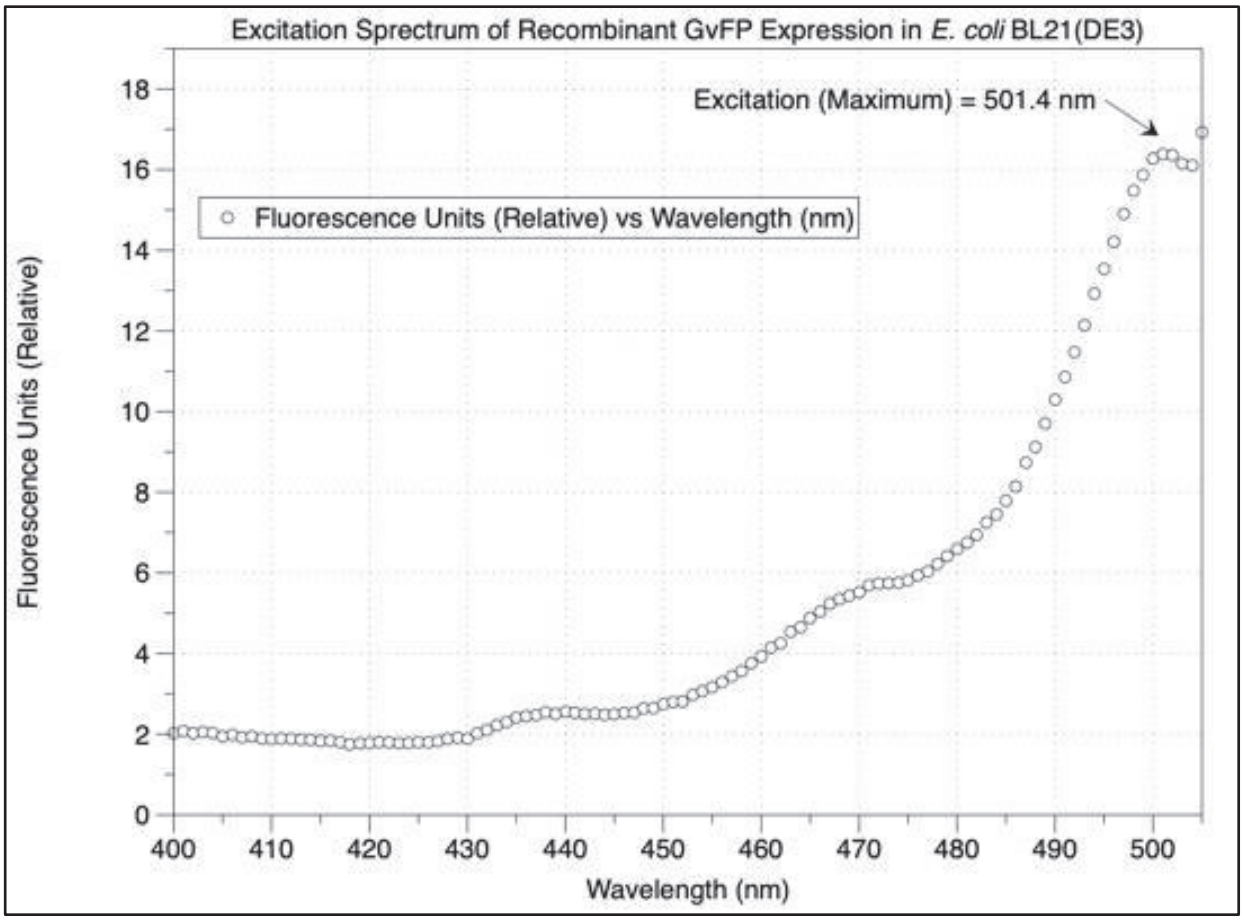


Figure 29. Recombinant GvFP excitation maxima measured at 501.4 nm using the Hitachi-7000 Fluorescence Spectrophotometer, illuminated with a range of wavelengths from 340 nm to 510 nm.

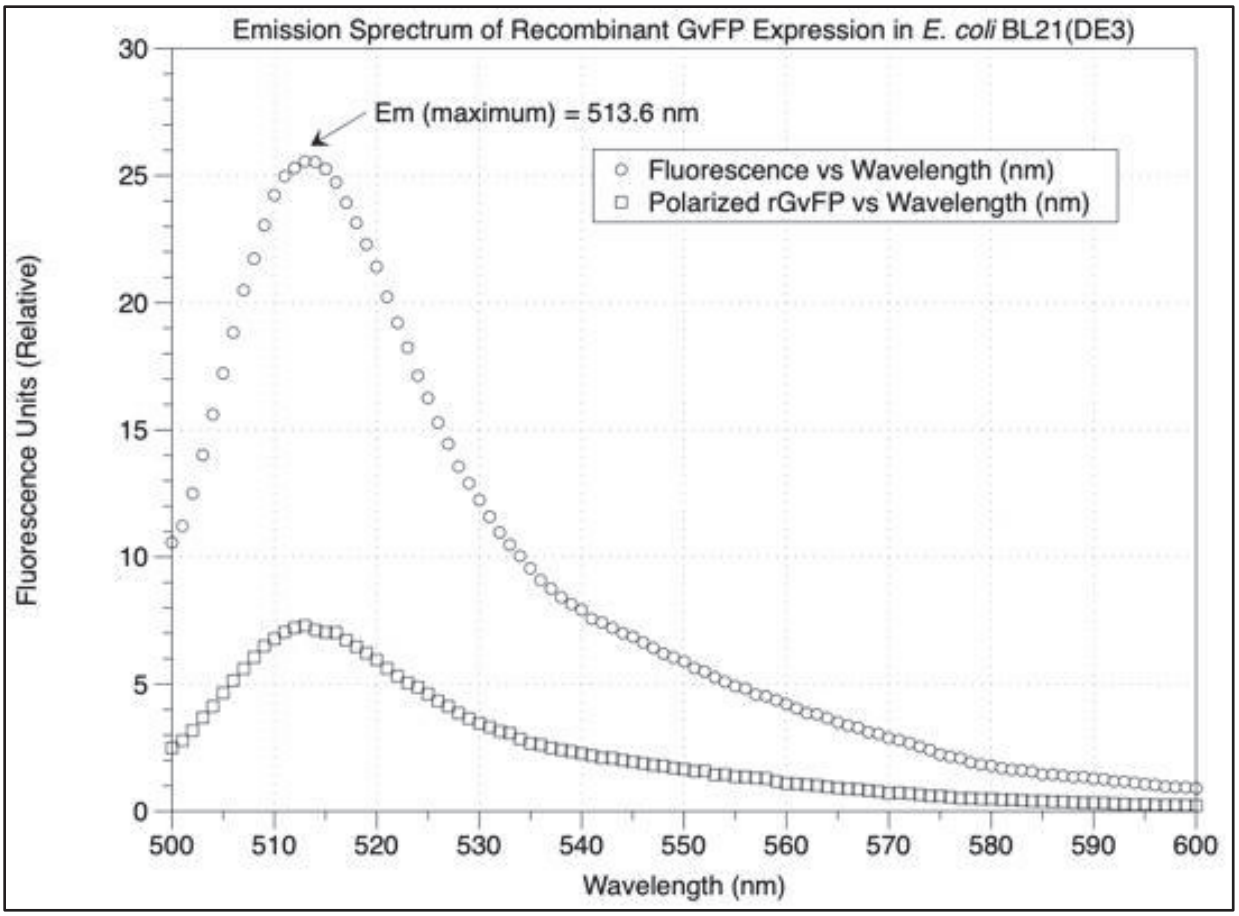


Figure 30. Recombinant GvFP homogenate emission fluorescence spectrum excited at 470 nm (in 10 mM Tris, pH 8.0, 1-cm path length cuvette). Polarized sample excited at 90$_{o}$. Maxima measured at 513.6 nm using the Hitachi-7000 Fluorescence Spectrophotometer.

DISCUSSION

Aequorea victoria wtGFP is known to excite at λ_{ExMax} = 488 nm and emit at λ_{EmMax} = 509 nm (Zhou, 2005) and the fluorescence spectra observed of intact, live *Gonionemus vertens* GFP is λ_{ExMax}= 501.4 nm and a λ_{EmMax}= 512.9 nm (Petroff *et al.*, 2017). The fluorescence spectra of native homogenate *Gonionemus* GFP closely mirrors that of the live *Gonionemus* spectra with λ_{ExMax}=501.4 nm and λ_{EmMax}=513.6 nm. In both cases (*in vivo* spectroscopy and *in vitro* spectroscopy of *Gonionemus* GFP) the Stokes Shift (the difference between λ_{ExMax} and λ_{EmMax}) is remarkably short, observed as 11.5 nm (*in vivo*) and 12.2 nm (*in vitro*) when compared to the Stokes shift of wtGFP of 21 nm. Considering ScSuFP has the highest percent identity to that of our newly discovered GvFP sequence identified here, it is not surprising to see similar λ_{ExMax} and λ_{EmMax} observed at λ_{ExMax} = 498 nm and λ_{EmMax} = 511 nm (Horiuchi *et al.*, 2018) making its Stokes shift 13 nm; extremely similar to that of *Gonionemus* GFP. Thus, GvFP has one of the smallest Stokes shifts seen in nature.

In attempts to isolate and amplify a GFP gene from *Gonionemus*, we started by using a traditional approach: amplify using primers designed to anneal to conserved regions of other hydrozoan GFP genes on *Gonionemus* DNA. In this case, we started by using primers designed for *Aequorea victoria* and *Olindias formosus* GFP. Even after mixing primer sets (amplifying with mismatched primers and same direction primers) in order to maximize the chances of amplification, all attempts failed. These failures led us to think that the GFP gene of *Gonionemus* is unique and may not share enough homology to that of *Aequorea* or *Olindias*.

To elucidate a GFP sequence from *Gonionemus*, we made the first ever *Gonionemus vertens* transcriptome using RNA-Seq on *Gonionemus* specimens collected at Farm Pond in Martha's Vineyard, MA. The transcriptome utilized total RNA followed by mRNA selection of whole organisms in order to sequence the mRNA transcripts of every jellyfish cell type and tissue. The coverage of RNA-Sequencing was further broadened by complexing five *Gonionemus* samples (all of which were collected in Farm Pond) in order to increase RNA-Seq output data, effectively allowing us to create five distinct transcriptomic libraries. The five samples were chosen based on anatomical features like color, size, and possible sex. These morphological differences allow us to assess transcriptomic variation between organisms from the same population that vary in their anatomy. The complexing of different cDNA libraries of different organisms allows us to carry out future studies on any gene family.

The creation and sequencing of five cDNA libraries allowed us to avoid further amplification by means of traditional methods which had previously resulted in experimental failure. Upon *de novo* assembly, we were able to obtain a high quantity of mRNA sequences, some of which we theorized would be the mRNA transcript of the suspected protein causing green fluorescence in *Gonionemus*. By searching for the conserved YG dipeptide of the tripeptide GFP chromophore we were able to identify possible transcripts that coded for a fluorescent protein of some type. Upon identifying putative fluorescent protein transcripts, we were able to understand why the previously designed primers for *Aequorea* and *Olindias* GPF failed to result in amplification using *Gonionemus* DNA.

After moving on and designing primers based on the contig 6770 sequence (generated by the Geneious *De Novo* assembly) that had the highest percent identity match to the *Scolionema*

suvaense fluorescent protein (ScSuFP), amplification was clearly visible upon gel electrophoresis. The 8 primer sets (sequences are listed in Table J, this thesis) were used in PCR reactions with genomic *Gonionemus* DNA from Martha's Vineyard, Massachusetts, and Potters Pond, Rhode Island. These primers were designed to anneal at different regions of the GvFP gene and therefore expected to produce different size fragments. Clear bands were produced only from the *Gonionemus* DNA from Martha's Vineyard, while bands produced from the Potters Pond samples were blurry and infrequent. This leads us to believe that the GFP gene in *Gonionemus* populations may differ to the extent that primers designed based on a contig discovered from a particular population may not amplify in organisms from a different population. This suggests that these GFP-encoding genes might be subject to rapid evolutionary change due to selective pressure in populations as has been suggested for venom-encoding genes in cnidarians.

Interestingly, of the amplicons that were produced in the Martha's Vineyard *Gonionemus* samples, each fragment size observed was considerably larger than predicted fragment sizes. This occurrence may be due to the fact that in whole genome DNA, the GvFP gene is present, but contains one or more non-coding introns. Sequencing using the same primer sets should allow us to determine whether or not introns are actually present and, if they are, identify both their location and size within the gene.

Due to failures in expected amplicon size after using primers designed from contig 6770 of *Gonionemus* DNA due to the possible presence of introns, we synthesized the full-length Contig 6770 GvFP (G-Block synthesis from IDT-DNA). For expression studies it was imperative that we use an intron-less version of the GvFP gene for expression in *E. coli*, since prokaryotes cannot process and remove introns. Subsequently, we ligated this synthetic contig into the pET-SUMO

expression plasmid and transformed BL21(DE3) and Mach1-T1 *E. coli* cells. Induction of the lac operon with the lactose-mimic IPTG initiates transcription in this vector, and we did see induction of green fluorescence in *E. coli* BL21(DE3) cells over the course of 4 hours. Interestingly, the induction showed a lag of 2 to 3 hours, which may be required for the correct auto-catalyzing of the chromophore of the GvFP. The presence of this recombinant protein in *E. coli* still needs to be verified by PAGE analysis and, preferably, with Western blotting using antibodies to GFP. Analysis of the fluorescence spectrum of cells expressing the recombinant GvFP at the 4 hour time point show that an excitation and emission spectrum that is nearly identical to the wild-type GvFP in *Gonionemus vertens*.

Future Work with GvFP

Based on the findings in this project, much work remains to be done:

1. **Increase Expression Levels of Recombinant GvFP.** Improving the expression level of the recombinant GvFP in *E. coli* would greatly facilitate future efforts to purify and study the purified protein. The pET-SUMO vector expresses proteins as a fusion protein, containing the 6X His motif that affords the ability to purify the recombinant fusion protein by affinity chromatography to Ni-columns. In addition, the amino terminal fusion peptide (SUMO) has an engineered protease site that allows cleavage with the SUMO Protease to completely remove this amino terminal leader from the GvFP protein. But greater quantities of the recombinant protein are needed to do this biochemistry. Larger amounts of purified GvFP would allow much cleaner spectral scans and permit characterization of other properties of this important protein. Although many GFPs are known to maintain

their fluorescent properties (maintaining their same excitation and emission wavelengths) when fused to another protein, isolation of larger amounts of the recombinant fusion protein may allow us to better understand this chimera.

2. **Search for Other FP's in Transcriptome**. It is likely that other fluorescent proteins remain to be discovered in the transcriptome libraries. Initial screening through the data set has suggested the presence of other homologs to GvFP, some with different chromophores than the MYG of GvFP.
3. **Look at Population Differences in MV Transcriptome Libraries.** The fact that we created a transcriptome from five individuals affords us the future possibility of examining population differences in the expression of various FP's in these samples.
4. **Discovery of Novel Venom and Non-Venom Proteins in *Gonionemus*.** Since the transcriptome libraries were created from total RNA of the entire organism, all genes expressed at the time of capture and extraction are likely to be represented in these libraries. That means that these should be screenable for any venom or non-venom protein or gene depending on the particular search strategy employed. For example, if one was interested in studying the histone genes of *Gonionemus vertens*, this would be easy. In fact, a quick search through the data set indicated that all five known histone proteins (H1, H2A, H2B, H3, and H4) are represented. This can be a very valuable resource for future researchers.
5. **Examine Putative Intron-Exon Structure of *Gonionemus* GvFP Genes.** It was clear that gDNA from *Gonionemus vertens* contained introns in the genomic GvFP genes. Although it was beyond the proposed scope of this thesis, future work characterizing the number, position, and sizes of these introns should be done.

6. **Examine Biological Functions of GvFP in the Biology of *Gonionemus*.** Lastly, understanding the role and function of GvFP and other FP's in the biology of *G. vertens* is important. The organism clearly expends considerable energy in expressing these proteins and we assume there must be a rationale for their presence. It is clear that most interest surrounding fluorescent proteins focuses on exploiting them to serve the needs of researchers (by means of mutagenesis and discovery of novel FP's), but the role that native GFPs play in the lives of hydrozoans is vastly understudied. Understanding what potential benefit comes from expressing GFP in living hydrozoans would allow us to better understand *Gonionemus* biology.

REFERENCES

Bajar BT, Wang ES, Zhang S, Lin MZ, Chu J. 2016. A Guide to Fluorescent Protein FRET Pairs. Sensors (Basel). 16(9).

Balleza E, Kim JM, Cluzel P. 2018. Systematic characterization of maturation time of fluorescent proteins in living cells. Nat Methods. 15(1): 47-51.

Barondeau DP, Putnam CD, Kassmann CJ, Tainer JA, Getzoff ED. 2003. Mechanism and energetics of green fluorescent protein chromophore synthesis revealed by trapped intermediate structures. Proc Natl Acad Sci USA. 100(21): 12111-16.

Blouin S, Craggs TD, Lafontaine DA, Penedo JC. 2009. Functional studies of DNA-protein interactions using FRET techniques. Methods Mol Biol. 543:475-502.

Chalfie M, Tu Y, Euskirchen G, Ward WW, Prasher DC. 1994. Green fluorescent protein as a marker for gene expression. Science. 263(5148): 802-5.

Chen X, Zaro JL, Shen WC. 2013. Fusion protein linkers: property, design and functionality. Adv Drug Deliv Rev. 65(10):1357-69.

Chudakov DM, Matz MV, Lukyanov S, Lukyanov KA. 2010. Fluorescent proteins and their applications in imaging living cells and tissues. Physiol Rev. 90(3):1103-63.

Craggs TD. 2009. Green fluorescent protein: structure, folding and chromophore maturation. Chemical Society Reviews. 38(10): 2865-75.

Davis T. Profile of Martin Chalfie. 2008. Proc Natl Acad Sci USA. 105(5):1393-5.

Day RN, Davidson MW. 2009. The fluorescent protein palette: tools for cellular imaging. Chemical Society Reviews. 38(10): 2887-921.

Gaynor JJ, Bologna PAX, Restaino DJ, Barry CL. 2016. First occurrence of the invasive hydrozoan *Gonionemus vertens* A. Agassiz (Cnidaria: Hydrozoa) in New Jersey, USA. BioInvasions Records. 5(4): 233-237.

Govindarajan AF, Källström B, Selander E, Östman C, Dahlgren TG. 2019. The highly toxic and cryptogenic clinging jellyfish sp. (Hydrozoa, Limnomedusae) on the Swedish west coast. PeerJ. 7: e6883.

Granatiero V, Patron M, Tosatto A, Merli G, Rizzuto R. 2014. Using targeted variants of aequorin to measure Ca2+ levels in intracellular organelles. Cold Spring Harbor Protocols 2014(1): 86-93.

Gross LA, Baird GS, Hoffman RC, Baldridge KK, Tsien RY. 2000. The structure of the chromophore within DsRed, a red fluorescent protein from coral. Proc Natl Acad Sci USA. 97(22):11990-11995.

Haddock SH, Dunn CW. 2015. Fluorescent proteins function as a prey attractant: experimental evidence from the hydromedusa *Olindias formosus* and other marine organisms. Biol Open. 4(9):1094-104.

Heim R, Cubitt AB, Tsien RY. 1995. Improved green fluorescence. Nature. 373(6516): 663-664.

Heim R, Prasher DC, Tsien RY. 1994. Wavelength mutations and posttranslational autoxidation of green fluorescent protein. Proc Natl Acad Sci USA. 91(26): 12501-12504.

Hennigan RF, Chaiken MF, Foster LA, Ip W. 2009. A FRET-based approach for studying conformational changes of a cytoskeleton-related tumor suppressor molecule. Methods Mol Biol. 586:143-56.

Horiuchi Y, Laskaratou D, Sliwa M, Ruckebusch C, Hatori K, Mizuno H, Hotta JI. 2018. Frame-Insensitive Expression Cloning of Fluorescent Protein from *Scolionema suvaense*. Int J Mol Sci. 19(2)

Lakowicz JR. 2006. Principles of Fluorescence Spectroscopy (Third Edition). Springer. Pg. 12-13.

Liu L, Li Y, Li S. 2012. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012-251364.

Marchessaux G, Gadreaud J, Martin-Garin B, Thiéry A, Ourgaud M, Belloni B, Thibault D. 2017. First report of the invasive jellyfish *Gonionemus vertens* A. Agassiz, 1862 in the Berre Lagoon, southeast France. BioInvasions Records. 6(4):339-344.

Ohmiya Y, Hirano T. 1996. Shining the light: the mechanism of the bioluminescence reaction of calcium-binding photoproteins. Chem Biol. 3(5):337-47.

Petroff E, Snitsarev V, Lopez R, Bologna PAX, Gaynor JJ. 2017. Green fluorescence from non-aequorea hydrozoa *Gonionemus vertens* (Clinging Jellyfish). FASEB J. 31(1):631.

Pigulevsky SV, Michaleff PV. 1969. Poisoning by the medusa *Gonionemus vertens* in the Sea of Japan. Toxicon, 7(2); 145–149.

Phillips GJ. 2001. Green fluorescent protein--a bright idea for the study of bacterial protein localization. FEMS Microbiol Lett. 204(1): 9-18.

Prasher DC, Eckenrode VK, Ward WW, Prendergast FG, Cormier MJ. 1992. Primary structure of the *Aequorea victoria* green-fluorescent protein. Gene. 111(2): 229-233.

Remington SJ. 2011. Green fluorescent protein: a perspective. Protein Science. 20(9): 1509-1519.

Shimomura O. 2009. The discovery of the green fluorescent protein (GFP) (Nobel Lecture). Angewandte Chemie International Edition. 48(31): 5590-5602.

Shimomura O. 1979. Structure of the Chromophore of *Aequorea* Green Fluorescent Protein. FEBS Letters. 104: 2.

Shimomura O, Johnson FH, Saiga Y. 1962. Extraction, purification and properties of aequorin, a bioluminescent protein from the luminous hydromedusan, *Aequorea*. J Cell Comp Physiol. 59:223-239.

Toyama BH, Savas JN, Park SK, Harris MS, Ingolia NT, Yates JR, Hetzer MW. 2013 Identification of long-lived proteins reveals exceptional stability of essential cellular structures. Cell. 154(5): 971-982.

Tsien RY. 1994. The green fluorescent protein. Annual Review of Biochemistry. 67: 509-544.

Tsien RY. 2010. Nobel lecture: constructing and exploiting the fluorescent protein paintbox. Integr Biol (Camb). 2(2-3):77-93.

Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.

Wiedenmann J, Elke C, Spindler K-D, Funke W. 2000. Cracks in the β-can: Fluorescent proteins from *Anemonia sulcata* (Anthozoa, Actinaria). Proc. Nat. Acad. Sci. 97 (26) 14091-14096.

Williamson J. 1996. Venomous and Poisonous Marine Animals: A Medical and Biological Handbook. New South Wales. University of New South Wales Press.

Yang F, Moss LG, Phillips GN. 1996. The molecular structure of green fluorescent protein. Nat Biotechnol. 14(10):1246-51.

Zagranichny VE, Rudenko NV, Gorokhovatsky AY, Zakharov MV, Balashova TA, Arseniev AS. 2004. Traditional GFP-type cyclization and unexpected fragmentation site in a purple chromoprotein from *Anemonia sulcata*, asFP595. Biochemistry. 43(42): 13598-13603.

Zimmer M. 2002. Green fluorescent protein (GFP): applications, structure, and related photophysical behavior. Chemical Reviews. 102(3): 759-781.

Appendix A: Amino Acid Coding Frequency of *G. vertens* Transcriptome

Amino Acid	Frequency	%
A	1,632,046	4.8%
C	1,314,787	3.8%
D	1,040,603	3.0%
E	1,247,379	3.6%
F	1,797,286	5.2%
G	1,438,665	4.2%
H	1,216,057	3.5%
I	1,899,228	5.5%
K	1,697,530	5.0%
L	3,325,678	9.7%
M	704,059	2.1%
N	1,316,719	3.8%
P	1,462,226	4.3%
Q	1,391,963	4.1%
R	2,378,594	6.9%
S	3,364,903	9.8%
T	2,013,246	5.9%
V	1,981,572	5.8%
W	547,215	1.6%
Y	1,009,972	2.9%
*	1,501,573	4.4%
X	287,235	
TOTAL	34,568,536	100%

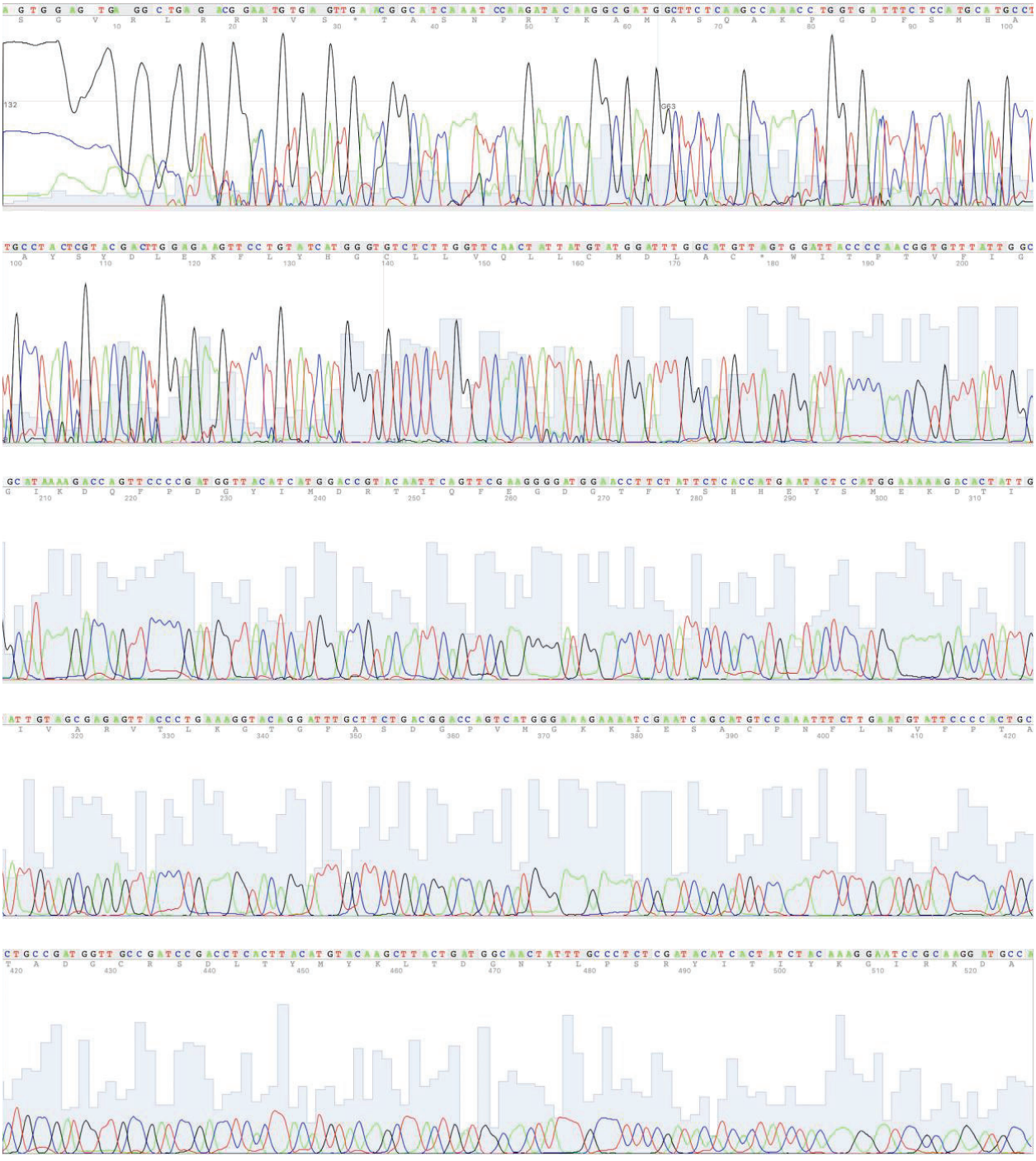
Appendix B: Codon Use Frequency in *G. vertens* Transcriptome

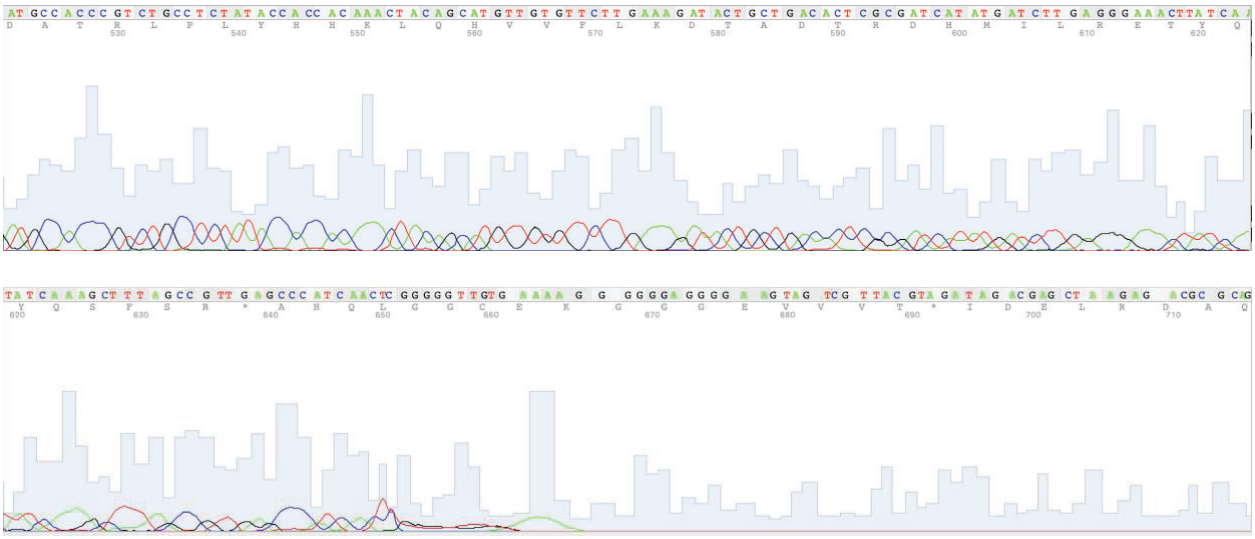
Codon	Amino Acid	% of Amino Acid	Frequency
GCA	A	37.0%	602,026
GCC	A	18.7%	304,295
GCG	A	14.9%	241,741
GCT	A	29.4%	478,373
TGC	C	45.9%	602,026
TGT	C	54.1%	710,514
GAC	D	40.4%	419,379
GAT	D	59.6%	619,616
GAA	E	63.6%	792,674
GAG	E	36.4%	453,263
TTC	F	44.2%	792,674
TTT	F	55.8%	1,001,784
GGA	G	34.6%	496,496
GGC	G	21.2%	304,296
GGG	G	17.6%	252,837
GGT	G	26.5%	380,379
CAC	H	42.0%	510,555
CAT	H	58.0%	704,059
ATA	I	30.4%	576,777
ATC	I	32.7%	619,616
ATT	I	36.9%	699,614
AAA	K	59.1%	1,001,784
AAG	K	40.9%	693,830
CTA	L	9.7%	320,956

CTC	L	13.7%	453,263
CTG	L	18.6%	615,938
CTT	L	20.9%	693,830
TTA	L	13.8%	458,581
TTG	L	23.3%	774,442
ATG	M	100%	704,059
AAC	N	46.8%	615,196
AAT	N	53.2%	699,614
CCA	P	37.8%	547,215
CCC	P	17.4%	252,837
CCG	P	17.7%	255,995
CCT	P	27.1%	393,067
CAA	Q	55.7%	774,442
CAG	Q	44.3%	615,938
AGA	R	28.8%	683,885
AGG	R	16.6%	393,067
CGA	R	18.1%	429,510
CGC	R	10.2%	241,741
CGG	R	10.8%	255,995
CGT	R	15.5%	368,369
AGC	S	14.3%	478,373
AGT	S	16.3%	548,344
TCA	S	21.4%	719,704
TCC	S	14.8%	496,496
TCG	S	12.8%	429,510
TCT	S	20.4%	683,885

ACA	T	35.4%	710,514
ACC	T	18.9%	380,379
ACG	T	18.3%	368,369
ACT	T	27.3%	548,344
GTA	V	21.8%	431,707
GTC	V	21.2%	419,379
GTG	V	25.8%	510,555
GTT	V	31.1%	615,196
TGG	W	100%	547,215
TAC	Y	42.8%	431,707
TAT	Y	57.2%	576,777
TAA	*	30.6%	458,581
TAG	*	21.4%	320,956
TGA	*	48.0%	719,704

Appendix C: Electropherogram of Contig 6770 Sequenced with Primer Set 8 (Forward primer).





Electropherograms of the synthetic contig 6770 DNA, using the forward primer (of set 8) viewed with 4Peaks (https://nucleobytes.com/4peaks/index.html) and sequenced by chain termination sequencing on the ABI 3130.

Appendix D: IUB Nucleotide Codes

IUB Nucleotide Codes

Code	Definition	Mnemonic
A	Adenine	**A**
C	Cytosine	**C**
G	Guanine	**G**
T	Thymine	**T**
R	AG	pu**R**ine
Y	CT	p**Y**rimidine
K	GT	**K**eto
M	AC	a**M**ino
S	GC	**S**trong
W	AT	**W**eak
B	CGT	Not A
D	AGT	Not C
H	ACT	Not G
V	ACG	Not T
N	AGCT	a**N**y