THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF CHEMICAL ENGINEERING

RICE EMBRYOGENIC PROTEIN PRODUCTION AND DISTRIBUTION

KATIE ANNMARIE LEGENSKI
SPRING 2016

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree in Chemical Engineering
with honors in Chemical Engineering

Reviewed and approved* by the following:

Wayne Roger Curtis
Professor of Chemical Engineering
Thesis Supervisor

Darrell Velegol
Professor of Chemical Engineering
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

The purpose of this project is to identify a cost- and resource-effective production platform for the useful embryogenic protein ENODLI from rice (*Oryza sativa*). OsENODLI has been shown to induce somatic embryogenesis in plant calli and somatic tissue when applied in tissue culture. Somatic embryos are relevant to plant molecular biologists because of the utility of this tissue when genetically transforming plants. Somatic embryogenesis also provides a method of propagating clones of superior plant cultivars. Current techniques for inducing somatic embryogenesis are inefficient, expensive or time-consuming, relying on conditioned media or highly specific mixtures of plant hormones that vary with plant type and even cultivar. Genetically engineering plants is proving to be vital to food security and faces greater challenges due to rampant disease and climate change. Over 1/5 of calories globally consumed are supplied by rice. A staple in predominantly agrarian economies like those in South East Asia, rice is being threatened by encroaching shorelines due to saline sensitivity. OsENODLI could help ensure rice security by accelerating rice salt tolerance for a large portion of the global population. With the proper production platform in hand, OsENODLI will be made available to rice researchers around the world to aid in their transgenic and propagation efforts for this critically relevant crop.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Dr. Curtis for his continued support for this work in his lab. I would also like to thank the National Science Foundation and the Department of Chemical Engineering Penn State Bioendowment summer fellowship program which largely gave me the opportunity to learn much of the theory necessary over summer 2015 for conducting this work.

I would also like to thank Bastian Minkenburg in Yang lab at Penn State for teaching me seed sterilization and other basic plant tissue culture techniques, as well as Tina Lai for working with me for other procedures I needed to carry out this work. A special thanks as well to Sergio Florez who helped me with initial primer design over summer 2015 and to Nymul Khan who taught me the basics of working in CurtisLab. Finally, a thanks to Nate Hamaker for finishing out the BBrady algae bacterial symbiont project.

Chapter 1

Preface

The purpose of this portion of my thesis is to inform the reader of how my research experiences evolved into the thesis project. This chronicle of my time in CurtisLab also serves as context for those who would like to continue my project after I graduate or perhaps any reader interested more generally in the research topics or the nature of undergraduate research. I have worked on two major projects in CurtisLab: *Bradyrhizobium* bioinformatics and PLA molecular cloning, which are discussed more below.

Because this thesis is written for the benefit of those who want to learn, rather than to be pedantically specific, I've included "Recall" paragraphs that help explain the biology in more depth than what a typical chemical engineering student would know (a glossary of terms is also included). Most importantly, DNA sequences are listed to remove all doubt of what I actually worked with or accomplished during my time in CurtisLab. I hope this thesis can serve as a learning tool and chronicle of my mistakes and successes for any student who would like to continue work on the project.

BBrady

The first project I worked on in CurtisLab was on *Bradyrhizobium* bioinformatics. This project was a continuation of another CurtisLab member's work (Justin Yoo, now at University of California, Santa Barbara pursuing a PhD in chemical engineering). In the course of his work,

Justin managed to isolate and sequence a novel bacterium with 16s rRNA that categorized the bacteria as a member of the *Bradyrhizobium* family.

Recall: 16s rRNA

16s rRNA refers to a DNA sequence that is highly conserved from organism to organism. You can think of it as being the bacterium's fingerprint. The 16s is used to identify bacteria or determine which organism a bacterium is most similar to, which is what was done in this case.

Justin isolated this bacterium when he was working with *Botryococcus braunii*, a colony-forming eukaryotic microalga that prolifically produces hydrocarbons that can be used in biofuels (Yoo, 2013). His work encompassed determining the algae's growth kinetics so we could better understand how to utilize the algae for energy security. In order for Justin to determine kinetic parameters, he needed an axenic culture of the algae. This means that the algae could have no bacteria living with it as it naturally does, but instead the algae must be a monoculture. Obtaining an axenic culture is typically a routine and trivial process, but axenically isolating this algae was particularly difficult. The bacterium that was the most difficult to remove from the algae was fully sequenced at the Penn State Huck Institute of Life Sciences Genomics Core Facility on an Illumina MySeq next generation sequencer.

My job on this project was to work with Dr. Istvan Albert (iua1@psu.edu, **(814) 865-2281**), Associate Professor of Bioinformatics at Penn State, to assemble the bacterium's genome. This organism was sequenced in 250 bp pieces that had to be assembled into an 8 Mbp genome (genome size was determined via k-mer analysis). Although bacterial genomes are relatively

easy to assemble, I had very little knowledge of, let alone experience using, assembly software (assembling genomes is a highly intensive process done by complex computer programs), thus I am very grateful to Dr. Albert for his aid throughout the process.

The first step to genome assembly was assembling contigs. A contig is a contiguous sequence of DNA assembled from the short 250 bp sequences (called raw reads). This was done in the SPAdes bacterial assembly program (Bankevich et al., 2012). The report for the contig assembly is shown in Figure 1. The information that I found pertinent as I continued my work on this project was the GC content (listed as "GC (%)" in Figure 1). The GC content for this bacterium was notably high at 65%. This detail is important when amplifying a sequence via PCR.

	scaffolds
# contigs (>= 0 bp)	147
# contigs (>= 1000 bp)	56
Total length (>= 0 bp)	8053200
Total length (>= 1000 bp)	8011077
# contigs	76
Largest contig	997947
Total length	8024132
GC (%)	64.58
N50	541543
N75	206566
L50	6
L75	12
# N's per 100 kbp	0.00
# predicted genes (unique)	7618
# predicted genes (>= 0 bp)	7627
# predicted genes (>= 300 bp)	6773
# predicted genes (>= 1500 bp)	984
# predicted genes (>= 3000 bp)	94

Figure 1: Contig Assembly report from SPAdes. This report contains details on how well the contigs were assembled from the raw reads. The largest contig is quite large (almost 1Mb) implying the program did not have much difficulty assembling the contigs.

After assembling the contigs, the next step in genome assembly is to assemble the contigs into the whole genome. This is done by scaffolding the contigs to an already assembled genome (called a reference genome) that is fairly close to what you expect your bacteria to be. However, we could not finish the assembly in this relatively simple way because the bacterium was sufficiently novel that no suitable reference genome was available to scaffold the contigs onto.

Two possibilities existed for taking this project forward: optical mapping or SMRT sequencing. Optical mapping creates a genome-wide restriction digest map that can be used as the reference genome. To optically map a genome, a highly-specific restriction enzyme is used to digest genomic DNA. The DNA is then assembled by size of DNA between cut sites rather than by sequence. The size of each piece of DNA is determined by fluorescence, hence the name, optical mapping. This can be very helpful as long as the enzyme is chosen carefully.

One of the most difficult parts of finishing a genome (assembling your contigs into the final product) is sequencing the non-coding regions that are typically tandem repeats. If the enzyme does not cut in those regions, you still do not know what lies between contigs, though you may have a better idea of the order and orientation of your contigs (Boers, Burggrave, van Westreenen, Goessens, & Hays, 2014).

The second option was SMRT (single molecule, real time) sequencing. This technique can sequence DNA in up to 20,000 bp (recall the MySeq from Penn State only gives 250 bp reads) by isolating a single strand of DNA with a single DNA polymerase in a well. This technology is expensive but not too expensive when sequencing a small bacterial genome. SMRT sequencing has no trouble sequencing tandem repeats within non-coding regions of genomes and is quite optimal for finishing small bacterial genomes (Roberts, Carneiro, & Schatz, 2013).

Because I did not understand the differences between the two technologies nor did I know that the SMRT sequencing was within our budget for the project, I sent the genome for MapIt optical mapping at OpGen (Figure 2). The genome was mapped with KpnI, but still had several gaps in non-coding regions even after analysis with MapSolver. This optical map answered questions pertaining to order and placement of genes, but we did not have the clean finished genome we desired.

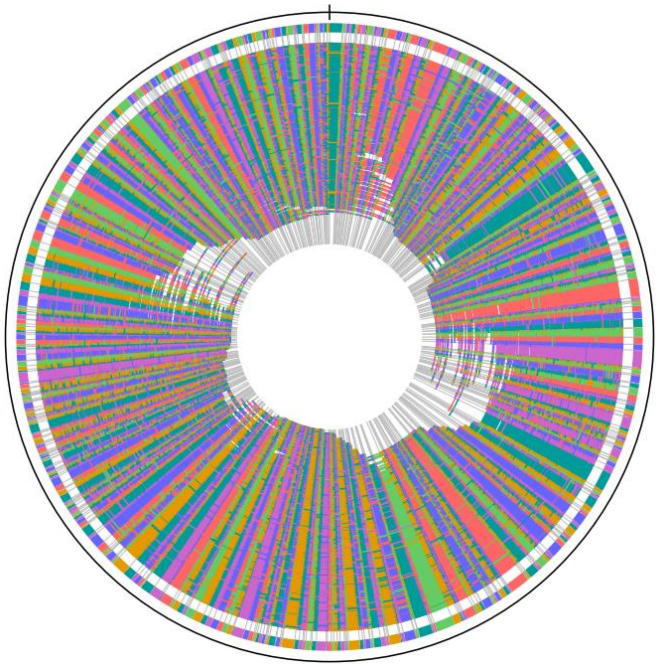

Figure 2: BBrady Optical Map. Each concentric circle represents 1X coverage of the whole genome. The different colors represent the regions between cut sites. The consensus restriction map is the outermost circle.

At this point in the project, Nate Hamaker began to close the gaps in the genome by sequencing PCR products with primers designed to amplify regions between the gaps. Please contact him if you would like to know more about how he finished the genome.

A piece of information that may be of use to future lab members is a set of primers I designed to be specific to our BBrady. The primers are called “Rand_fwd” and “Rand_rev” (fwd:

CGGCAGCGGCGTTAGGTATTC; rev: AAATTGCCCGATCTCTTGCCTCATC) and amplify a 753 bp region spanning part of a histidine kinase. These primers are also extremely specific (currently being used Antonio Xu to identify the presence of the bacterium in algal cultures), amplifying nothing else in the Bradyrhizobiaceae family when tested with a permissive primer blast. These primers can be used, for example to identify the presence of BBRADY by colony PCR based on the size of the amplicon.

OsENODLI

While working with BBrady, I was interested in expanding my research experience to include molecular biology. Specifically, I have an interest in international food security and a specific interest in rice (in part due to my research mentor Nymul Khan, who is from Bangladesh). This lead to my submission of a Fulbright research grant related to rice genetic modification for saline tolerance and my joining Tina Lai in her cloning efforts with other embryogenic proteins in cacao.

This molecular biology work started out as something that I found to be interesting and turned into my thesis work and the basis of this paper.

Chapter 2

Background

Identifying OsPLAI

Phytocyanin-like arabinogalactan (PLA) proteins are members of the arabinogalactan protein (AGP) family, a family of structurally complex proteoglycans (heavily glycosylated proteins – covalently linked arabinose sugar polymers). AGPs are typically membrane proteins utilized in extracellular signaling in diverse roles influencing gametogenesis, root and stem regeneration, and vascular development/reinforcement (Arey, 2012). AGPs were identified as the component of conditioned media (reused media from a previous culture often rich in trace but necessary elements) responsible for inducing somatic embryogenesis (SE), the rate-limiting step in producing transgenic plants and propagating slow-growing plants. Further work isolated a single molecule responsible for inducing somatic embryogenesis, opening the door for more timely higher plant transformation: PLA (Poon, Heath, & Clarke, 2012).

Nomenclature

All PLAs in this paper will be referred to as "Gs"PLA where G is the plant genus and s is the plant species. For example, the PLA identified in cotton (*Gossypium hirsutum*) will be referred to as GhPLA. All higher plants relevant to PLA will be introduced with their respective genus and species.

PLA is the family of proteins OsENODLI belongs to. You will find I try to use OsENODLI in reference to the protein that is the focus of this thesis because it is annotated in the *Oryza sativa japonica* cultivar nipponbare genome as such. However, when I started the project, the protein was referred to as OsPLA1 because of its homology to similarly named proteins in the lab. CurtisLab nomenclature refers to all proteins with PLA-like properties to be PLAs. Note that OsENODLI and OsPLA1 refer to the same gene and subsequent protein.

PLA was first identified as a strong SE inducer in cotton. Instead of harvesting the protein from plant tissue, Poon et. al. cloned and overexpressed the protein in both tobacco (*Nicotiana tabacum*) and *E. coli*. This posed an interesting functionality question: will a plant protein expressed in bacteria still be functional? This question is relevant because plant proteins are subjected to many post-translational modifications that influence their structural conformation and hence functionality. Bacterial proteins are simpler and do not require as much post-translational modification. Poon proved that the post-translational modifications to GhPLA were trivial, showing the bacterially-produced GhPLA to be just as effective in inducing SE as the plant-produced protein (Poon et. al, 2012).

Several orthologs to GhPLA that have been identified via NCBI tblastx are listed in Table 1. The sequence I chose as my basis for OsPLA1 was the sequence Mashiguchi isolated to have SE inducing potential (Mashiguchi, Yamaguchi, & Suzuki, 2004). The PLAs identified from cacao (*Theobroma cacao*), TcPLA1 and TcPLA2, are currently being studied by CurtisLab master's student Tina Lai. Tina has served as my mentor for this project and has taught me many lab techniques. My work was meant to expand on hers by cloning a monocot PLA from rice

(*Oryza sativa*). Monocots and dicots have significant enough differences in embryogenesis (and in sequence) to warrant this side project.

Table 1: PLA Orthologs from ClustalOmega. **This matrix lists the identities of several PLAs, i.e. the percent similarity of amino acids. Observe the diagonal of "100" illustrating each protein is identical to itself. Note also that OsPLA is significantly different from all other PLAs isolated thus far by those in and outside of CurtisLab.**

	OsPLA rice	DaPLAI yam	DrPLAI yam	GhPLAI cotton	TcPLAII cacao	TcPLAI cacao
OsPLA rice	100	35.9	38.1	34.3	36.5	37.0
DaPLAI yam	35.9	100	91.9	52.7	50.5	50.5
DrPLAI yam	38.1	91.9	100	61.2	60	60
GhPLAI cotton	34.3	52.7	61.18	100	83.2	80.9
TcPLAII cacao	36.5	50.5	60	83.2	100	99.1
TcPLAI cacao	37.0	50.5	60	80.9	99.1	100

Table 1: PLA Orthologs from ClustalOmega shows the homology of the gene I chose as the PLA ortholog in rice (*Oryza sativa*) (Bawono, Heringa, & Editor, 2014). The matrix shows that all the rice gene shows little homology to other PLAs, even other monocots like yam. I thought this was odd, so I blasted (tblastx with DNA sequence and blastp with translation) the gene against all *Oryza sativa japonica* sequences in the database. Both results indicated that OsENODLI was the closest to GhPLA, though the similarity is still unconvincing (<70% protein sequence similarity). By BLASTing the rice OsENODLI domain against rice, this provides an assessment of whether there are other distant homologs, or if the entire class of PLA proteins in rice reflects this large difference. Since no similar proteins to OsENODLI were significant (small enough E value) in tblastx searches of the *Oryza sativa* database, it may concluded that the embryogenic proteins in rice are unique.

Rather than clone the entire protein, I decided to clone only the active domain 'PLA domain'. Cloning only the active domain reduces cell resources required to produce the protein. The entire gene is codes for 237 amino acids, while the active domain only encodes for 116. This 2-fold reduction in amino acid production puts less stress on the cell to produce the protein.

The domain I identified spanned amino acids 27-138 in the whole protein. This window was chosen because of the InterPro prediction listed in Figure 3. PLA is not a well-studied protein (the NCBI still lists OsENODLI as a predicted protein, not having an experimentally verified structure or function), which is why the protein has no overall predicted family. Because PLA is not well known, I took the domain to be from all three families: cupredoxin, plastocyanin-like, and unintegrated signatures. Arabinogalactans are copper-binding proteins, and plastocyanin is synonymous with phyto/phycocynanin, therefore, I thought it reasonable to take these predicted domains as correct. The sequence circled in red in Figure 3 is the domain that I chose because the other families' domains were included within it. The resulting sequence is 116 amino acids, which is comparable in length to TcPLA1 (115) and TcPLA2 (107 aa). In short, this domain that I predicted may be wrong or correct. Only functionality assays will confirm whether this prediction was correct. The full translated gene as predicted by Mashigushi and the domain I chose are in Appendices.

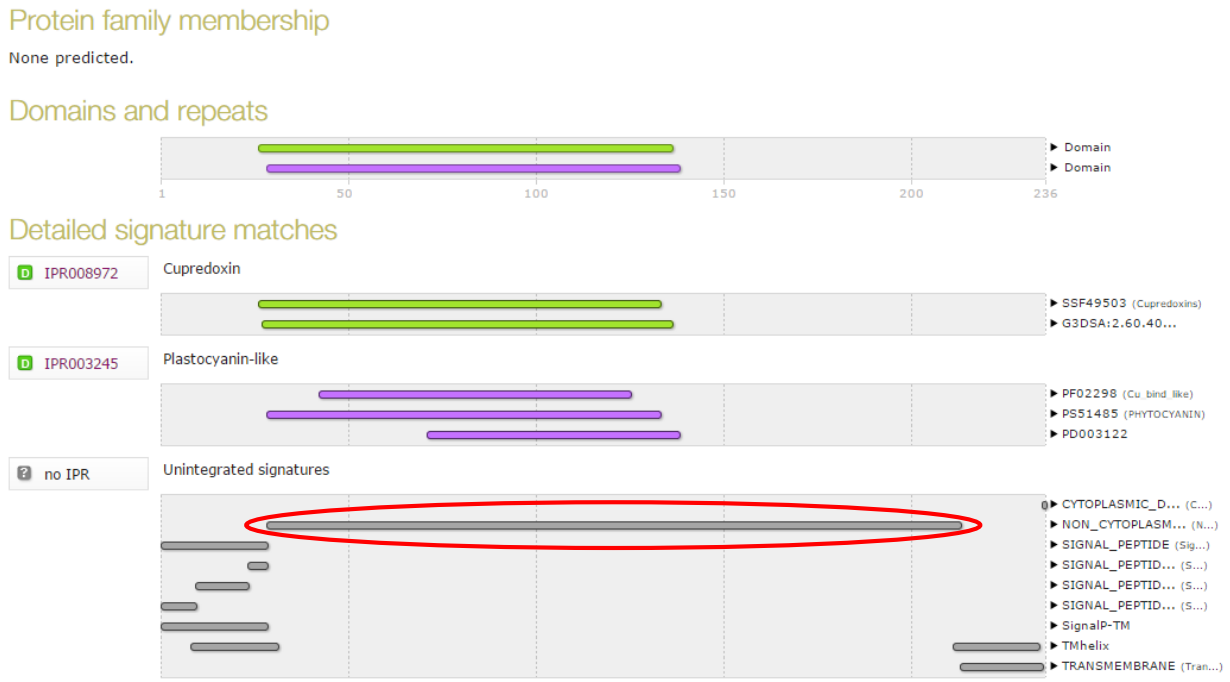

Figure 3: InterPro Domain Predictions. Each bar represents a predicted protein domain. The cupredoxin, plastocyanin-like, and NON_CYTOPLASMIC PEPTIDE predicted domains were chosen to isolate OsPLA's domain.

E. coli Expression Platform

E. coli was chosen as the initial expression platform for OsENODLI because it is well-characterized and relatively simply to clone into vectors designed specifically for heterologous protein expression. That is, all the available tools were readily available for cloning the gene into an *E. coli* vector. This simple initial cloning step was also used to verify the accuracy of the spliced gene (i.e. to determine whether or not the fusion PCR was successful).

E. coli expression was taken to be the base case for protein expression as well as cost analysis for protein production. Because *E. coli* is well-characterized, it is often used for expressing proteins although it is expensive to feed (i.e. the medium in which is grows is costly

when compared to cyanobacteria medium). The goal of this thesis was to produce protein at a low cost, making *E. coli* an illogical choice for expression. However, cloning into *E. coli* is better documented than cloning into cyanobacteria, making *E. coli* a convenient middle man for building constructs.

The pET14b vector was chosen to clone OsENODLI into because of its availability in the lab and its utilization of the T7-promtoer / polymerase system for high-level heterologous protein expression. The basic idea of this expression vector is providing a promoter that interacts with (thereby pretending to be) a viral gene that is uniquely transcribed by the chromosomally encoded T7 polymerase. The multiple cloning site (MCS) chosen for this cloning was the BamHI and NdeI MCS because of the lack of restriction homology to OsENDOLI and the economy of primer use (this same MCS is utilized in cyanobacteria vectors). This MCS does include a poly-histidine tag in addition to a thrombin cleavage site to purify the protein from the poly-histidine tag post purification (Figure 4).

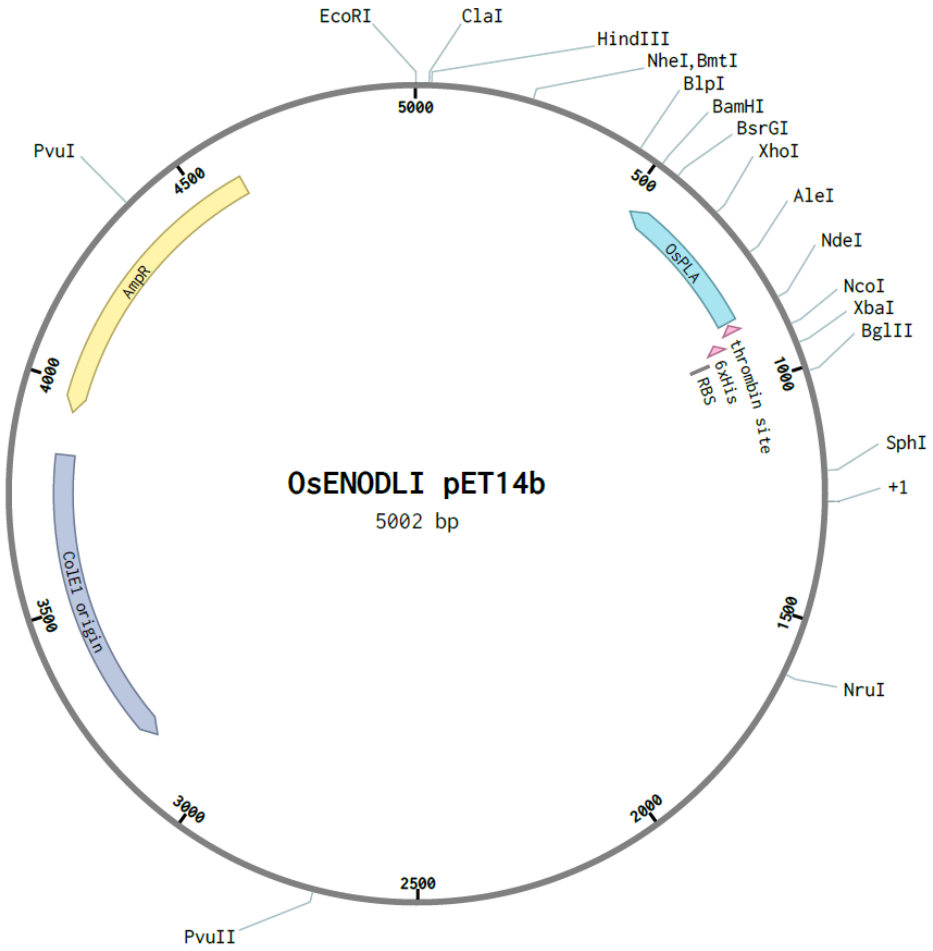

Figure 4: *E. coli* expression vector for OsENODLI. All single cutting enzymes made by NEB are listed. Cloning primer sequences are listed in App. B.

Cyanobacteria Expression Platform

Expressing OsENODLI in cyanobacteria holds promise for reducing production cost of the protein because cyanobacteria is capable of fixing atmospheric CO_2, essentially eliminating the cost of substrate in producing the protein. Secondly, harvested cyanobacteria can be used in biodiesel production through hydrothermal liquefaction or fermentation, further reducing the production cost of OsENODLI.

Chapter 3

Materials and Methods

Restriction-Based Cloning

OsENODLI was cloned into the pET14b vector via restriction cloning. This cloning approach utilizes DNA homology to insert a new DNA sequence. The homology is created by digesting the DNA to be inserted and the vector into which it will be inserted with restriction enzymes. These enzymes must be chosen so that the sticky ends they leave behind are identical for both insert and vector. The digested DNA fragments are then mixed together with DNA ligase. At random, the fragments associate and dissociate into fragile complexes that are held together only by hydrogen bonding. Once this fragile complex is recognized by DNA ligase, the enzyme covalently seals the phosphodiester backbone, creating the desired vector. This reaction product is then transformed into competent cells and selected for on antibiotic medium (NEB, 2015).

The restriction enzymes NdeI (NEB, R0111S) and BamHI HF ® (NEB, R3136S) were used to digest the insert and vector. Pure plasmid (> 2 µg) was digested with 20 U of each enzyme for 1 hour at 37°C. Similarly, PCR product was digested directly after PCR amplification with 20 U of each enzyme for 1 hour at 37°C. These products were purified separately using Omega Biotek's E.Z.N.A ® Cycle Pure Kit (product number D6492-00).

Insert and vector were added to yield a 18:1 (insert:vector) ratio by molecular weight. 400 U of T4 ligase (NEB, M0202S) was used to ligate the insert and vector in a 20 µL volume

for 24 hours. The reaction was started on ice to allow DNA fragment association and allowed to rise to 22°C over the 24 hour period.

I used TOP10 chemically competent cells for transforming my ligation product into, and selected for clones using ampicillin 100 µg/mL LB plates. Protocols for transformation and LB recipe may be found in 0.

Note: the restriction sites chosen include the poly-histidine affinity tag for purification in addition to a thrombin cleavage for pure protein isolation. This was not the original plan, but the cloning primers were designed to be compatible with cloning both into E. coli and cyanobacteria. The NdeI site is compatible with both the E. coli multiple cloning site and the cyanobacateria cloning site at the expense of including the his tag in E. coli. The his tag was meant to be left out because of purification costs. His tag purification is much more expensive than inclusion body purification (which is the protocol that was meant to be employed).

Colony PCR

Colony PCR is a quick method of probing for a gene in a sample of DNA. PCR is used to test for diseases such as Zika virus and other infections. This type of PCR does not require DNA cleanup such as a plasmid isolation or an even more demanding genomic DNA isolation, making it very desirable for selecting clones with the desired gene.

Recall: PCR

PCR is the acronym for polymerase chain reaction. For a PCR reaction, you need primers specific to the sequence of interest, dNTPs for making new DNA, template DNA from which

you are amplifying the sequence of interest. the DNA polymerase and any cofactors or buffers it might require. You only make a product if the sequence of interest is present in the template.

To screen for the desired gene, primers are designed to amplify a region specific to the gene of interest. For the primers that flank the MCS (multiple cloning site), clones may be selected based on the band size on the analytical gel.

Recall: Agarose Gels

DNA may be visualized on an agarose gel. The agarose forms a matrix that can separate DNA via friction based on length of the sequence. The separation occurs due to DNA's attraction to the positively charged cathode of the electrophoresis box. DNA is intrinsically negatively charged due to its phosphodiester backbone.

However, this does not confirm that the desired gene inserted, but rather only something similar in size did. Since CurtisLab also works with several other PLA genes similar in size to OsENODLI, a more specific approach to gene insertion confirmation is required. Thus, a primer specific to OsENODLI (i.e. a primer inside the gene) was designed and paired with a universal T7 plasmid primer. Obtaining an amplicon with this primer set not only ensures that OsENODLI inserted itself into the bacterial DNA, but also confirms that the insertion was not chromosomal. This sequence for this primer is listed in Appendix B under the name "OsPLA int rev."

The following paragraphs overview the procedure itself. First the cells that contain the plasmid with the gene of interest are lysed to expose whole genome DNA which then serves as

template for the PCR reaction. This initial step corresponds to "$t_{inital\ denature}$ (min)" and "$T_{denature}$ (°C)" in Figure 5: PCR Cycle.

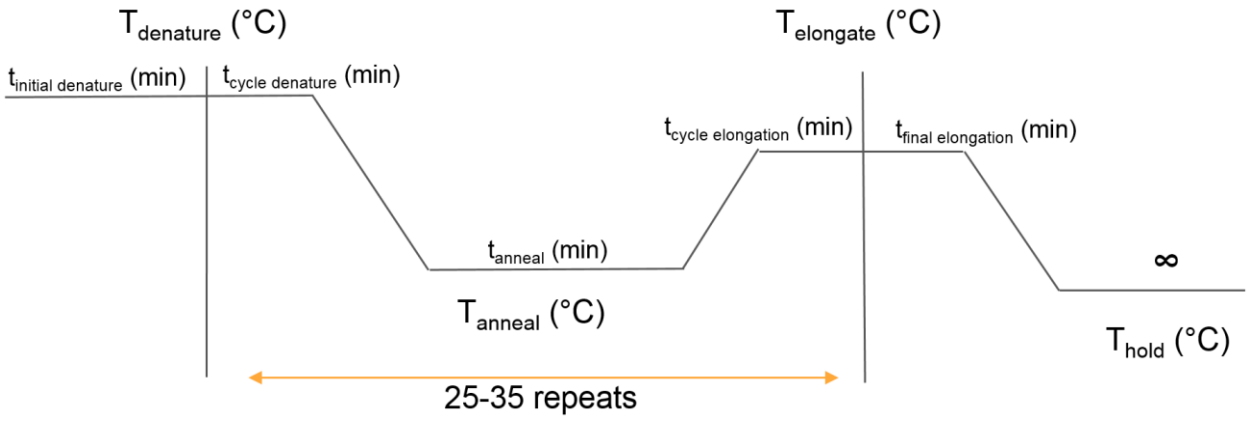

Figure 5: PCR Cycle

Second, the repeat portion of the cycle takes place in which the template repeatedly melts, anneals to the new primers, and elongates to obtain the target DNA sequence. It is in this portion of the reaction that the DNA is replicated. The number of repeats must be chosen carefully to preserve the DNA sequence. One must take into account the error rate of the polymerase when choosing the number of repeats. For especially long sequences of DNA, care must be taken to reduce the number of repeats in the PCR to reduce over amplifying DNA with errors. Having the correct DNA is far more valuable than having a large amount of DNA.

The T_{anneal} is arguably the most critical parameter to PCR success. It is the temperature at which have the total DNA mixture is melted into single strands, and half is associated in double strands via hydrogen bonding. It is calculated by thermodynamic parameters of the DNA basepairs by any one of many available programs. The annealing temperature should be the annealing temperature of the primers to the template. For my work, the Tm as predicted by Benchling is a usually a good annealing temperature for a taq-based PCR. However, some primers still need to be optimized via gradient PCR.

The final step in PCR is final elongation to ensure that all products have elongated fully. The final hold should be included in a PCR cycle to preserve DNA if it cannot be transferred to a freezer immediately after the reaction is finished.

The primers I used to verify my OsENODLI pET14b clone were the T7 terminator universal primer and OsPLA int fwd primer (both are listed in 0). I used Promega GoTaq ® (M3001, 2 U/µL), GoTaq Reaction Buffer ® (M7911, 1X), 10 mM dNTPs, 10 µM of each primer, and a stab of *E. coli* I was testing for each colony PCR reaction. The reaction cycling times and temperatures are in Figure 6.

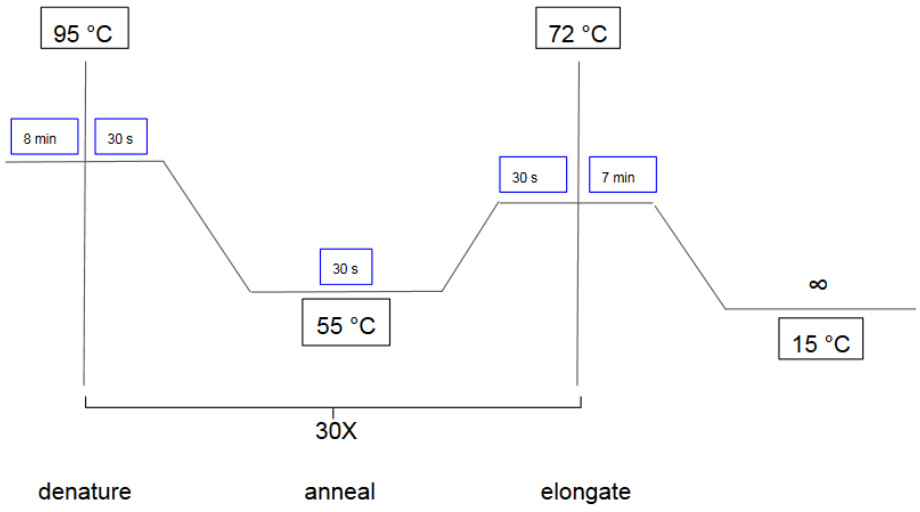

Figure 6: Colony PCR Cycle Times and Temperatures. This cycle was used to probe clones for OsENODLI. A 0.8 µL volume was run on a 1% agarose gel at 40 V/cm for 25 minutes to achieve a properly concentrated result band.

Fusion PCR

Fusion PCR is a method of overlap PCR applied to the specific area of gene splicing (Figure 7). In essence, fusion PCR can be used to glue any two sequences of DNA together (Tropea, Cherry, & Waugh, 2009). Two sets of primers were designed to accomplish the splicing. The first set of primers bracket the whole gene (Cloning primer in Figure 7). The

second set of primers (Fusion primer) has three conceptual regions: homology to the piece that is going to be glued to it, the piece that the other fusion primer will have homology to, and the piece that binds to the rest of the gene.

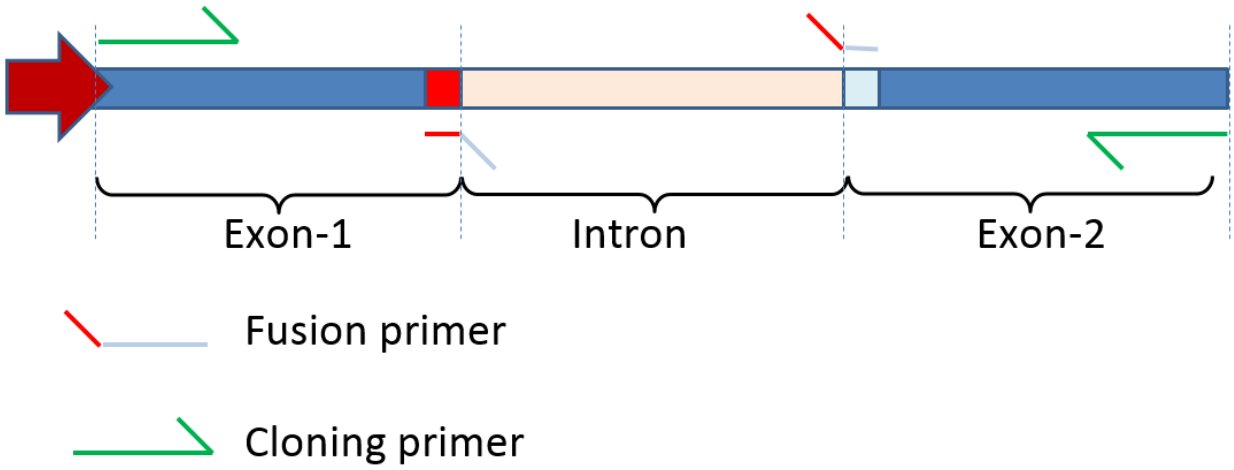

Figure 7: Fusion PCR Schematic

Three PCRs must be run in order to get the final product with no intron. The first PCR is to amplify Exon-1 with appropriate the cloning-fusion primer pair. The second PCR is to amplify Exon-2 with the appropriate cloning-fusion primer pair. These two reactions must be purified before proceeding to the final PCR in order to avoid unwanted side products. The third PCR reaction mixture contains the purified Exon-1 and Exon-2 templates in a 1:1 molar ratio and the cloning primers only. The annealing temperature for this PCR reaction should be the annealing temperature of the overlap regions of the fusion primers. This overlap region serves as a primer for gluing Exon-1 and Exon-2 together.

Please note that the fusion primers as designed for this project are not recommended to emulate. In other words, if you would like to design primers for fusion PCR, please read more

about the technique to design your primers rather than simply copying what I did. The overlap region is extremely small (12 bp). I designed it this way because I didn't know what I was doing. Typical overlap regions span at least 25 bp. I'm really not sure how or why this one worked.

RNA Extraction

When I was trying to isolate OsENODLI, I wanted to avoid fusion PCR altogether. One way I could do this was through amplifying the gene from a library of DNA written only from expressed genes: a cDNA library. A cDNA library is written from extracted RNA from living tissue – which at the same time provides for confirmation of the predicted intron sequences. Once genes are expressed as mRNA to be translated into protein, the introns (unexpressed regions) are spliced from the gene so that the protein can be functional. The trick to RNA extraction, is you must choose the tissue from which you extract the RNA carefully, because if your gene is not expressed in that tissue, you will not be able to exploit the cell's ability to splice your gene for you. The rationale for the tissue I chose to extract RNA from is described in the results section.

I used the TRIzol reagent (Life Technologies, 2012), to extract RNA from kitaake (the cultivar of *Oryza sativa japonica*) callus tissue. Notably, this cDNA approach to obtaining the gene was not ultimately successful.

Chapter 4

Results and Discussion

Gene Amplification from cDNA

My initial approach for trying to isolate OsENODLI was from cDNA (copy DNA as reverse transcribed from RNA). Amplifying OsENODLI from cDNA was a better option than amplifying the gene from gDNA (genomic DNA) because the gene contains an intron. Bastian Minkenburg, Ph.D. Student in Plant Biology, MBBISP Scholar, (814-867-0300, minkenberg@psu.edu) was kind enough to give me 5-day-old seedling cDNA. I tried to amplify OsENODLI from this DNA, but the gene was apparently not present in the tissue he extracted it from, as the cDNA isolation and subsequent PCR effort did not provide the desired 450 bp DNA fragment.

Next, I tried to predict which tissue the gene might be expressed in via GeneVesitigator. This tool is a database of gene expression in common plants and animals. It can predict when a certain gene is expressed in an organism's life cycle as well as co-expressers and regulators. Since PLA is a poorly characterized protein, it was not explicitly in the database, however, I searched similar proteins by binding domain to get an idea of which tissue I might be able to extract PLA mRNA. I searched GeneVestigator's rice database for phytocyanin-like, plastocyanin-like, and cupredoxin proteins and the pertinent results are listed in **Error! Reference source not found.** (Zimmermann, Hirsch-hoffmann, Hennig, & Gruissem, 2004).

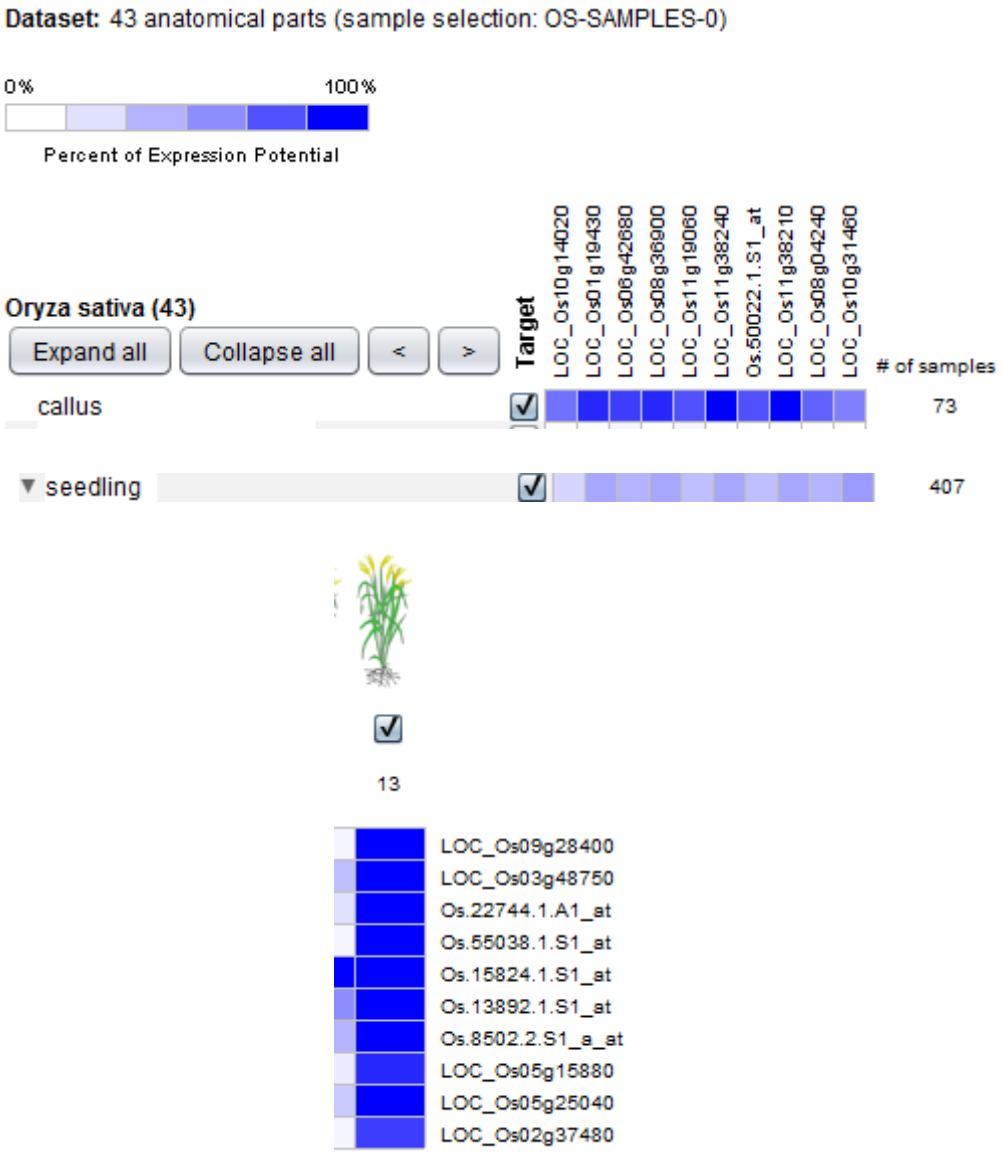

Figure 8: Predicted PLA Expression Tissue. The percent expression potential is listed in blue gradient, the darker the blue, the more likely it is for the gene to be found in that tissue. Gene names start with "LOC" or "Os." The best tissue to find OsPLA in is the flower cDNA (this tissue was unavailable). The next best tissue to find the gene expressed in was callus tissue, which was already in hand.

cDNA from callus tissue was made but the library was compromised by nuclease contamination as evidenced by a smeared gel. Rather than spend time optimizing this procedure

(note that only similar proteins and not PLA itself was predicted to be in callus tissue), we decided to amplify the gene using fusion PCR instead.

The primers used for fusion PCR are listed in 0 as "fusionPCR_fwd," "fusionPCR_rev," "OsPLA_mod_e+c_fwd," and "OsPLA_mod_e+c_rev." The gel with the full gene with intron, Exon-1, Exon-2, and the final fusion product is shown in Figure 9. Note that Exon-1 and Exon-2 were amplified from a gel purified amplicon because amplifying just the gene from genomic DNA had many other amplicons at different band sizes. The whole gene with intron was amplified, gel purified, and used as template for the PCR whose products are in lanes 2 and 3 of Figure 9.

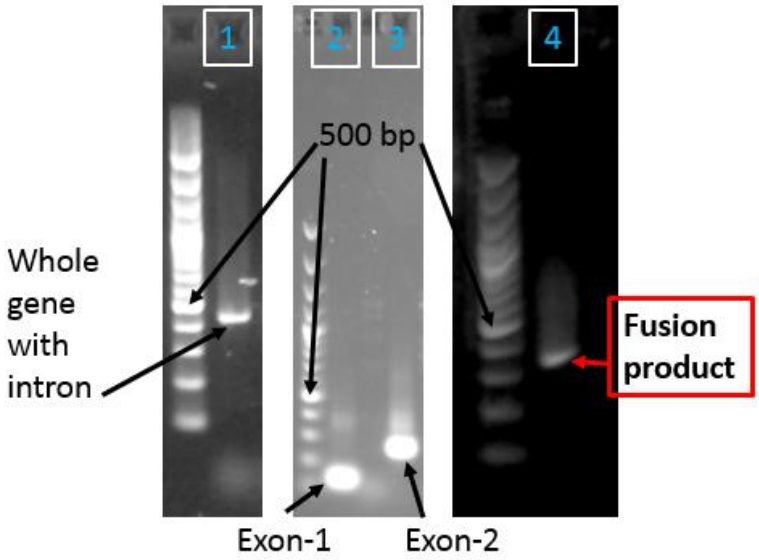

Figure 9: Fusion PCR Progression. Lane 1 shows the whole gene at ~450 bp; lane 2 shows Exon-1 at 117 bp; lane 3 shows Exon-2 at 276 bp; and lane 4 shows the fusion product at 350 bp. Too much DNA was loaded on gels, but the visual serves its purpose. Gels were prepared with 1% agarose, 1X TAE, and 1% V/V GelRed (Biotium, 41003).

The OsPLA_mod_e+c primer set was designed because of unsuccessful cloning using a previous primer set (with notation OsPLA ecoli fwd/rev and OsPLA cyano fwd/rev in my laboratory notebook). The original primers did not contain extra nonsense base pairs after the restriction enzyme recognition site. This lack of bases did not allow the restriction enzymes to

properly cut the PCR-amplified DNA, thus preventing sticky end formation. When designing primers, care must be taken to optimize primer design so that the PCR product to be digested is the optimal sequence for digestion. The requirements for these restriction sites added by PCR can be found on NEB's "Cleavage Close to the End of DNA Fragments." A three bp overhang for each primer was added so that primer integrity was not compromised.

Only one clone was isolated successfully from the transformation of OsPLA into TOP10 *E. coli* cells. This clone was sequenced in replicate (both T7 universal primers were used) and verified to be OsPLA (Figure 11). The SNPs (single nucleotide polymorphisms; variations in DNA sequence relative to the reference) are shown as the four white boxes on the blue strands. The blue strands represent my gene that I cloned, whereas the giant arrow represents the database reference sequence for the gene.

Upon initial inspection, these SNPs could be attributed to cultivar difference: the database sequence is from cultivar nipponbare and my DNA is from kitaake. Variations in genes are not uncommon among higher plants such as rice, particularly when the SNPs represent a 'silent mutation' in which the resulting translated amino acids are not changed. However, since the genomic DNA that was subjected to less PCR error was also sequenced, it was verified that all four of these SNPs were genuine mistakes.

Of the four SNPs, one was in the primer region. This mistake most likely occurred during one of the last PCR amplifications because the SNP is not in the gene with intron clone, and both clones were produced with the same primer set, ruling out the possibility that the primer was made erroneously. Another SNP was a silent mutation, leading to the same amino acid translation due to the code degeneracy. The other two SNPs led to amino acid changes.

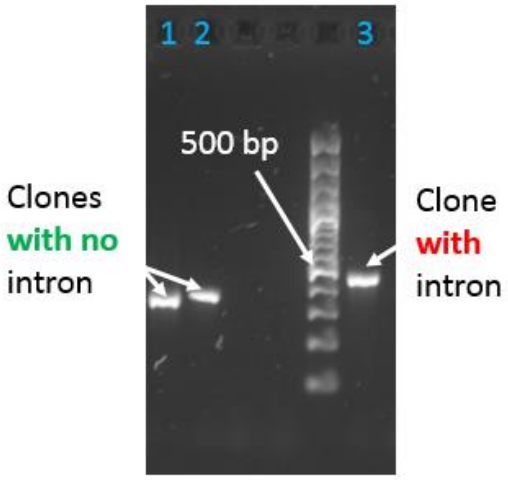

Figure 10: PCR Verification of Clones. Lanes 1 and 2 show a 350 bp band indicative of the spliced version of OsENODLI being cloned into pET14b; lane 3 shows that gene with intron was similarly successfully cloned.

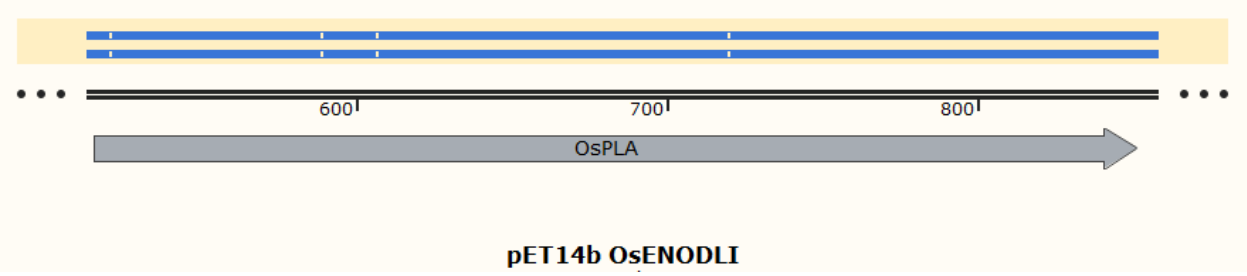

Figure 11 Alignment of Spliced OsENODLI to Database cDNA. The database DNA is from a cDNA library of mRNA from the nipponbare cultuvar. This is confirmation that the correct intron sequence was chosen for fusion PCR as well as the success of the fusion PCR.

However, these SNPs were verified to be errors generated by the PCR process after aligning the separately amplified gene plus the intron (Figure 12). I cloned this genomic DNA *E. coli* because I had difficulty amplifying the gene from genomic DNA and by cloning the genomic DNA, I was assured to have a high concentration of template from which to do subsequent work from plasmid preps, rather than having to rely on amplification from the low copy number genomic DNA. With the direct genomic cloning the intron is included with a minimal number of rounds of PCR. The absence of SNPs (white regions) within the blue show that the sequence aligns perfectly with the predicted except for a single SNP within the intronic

region. Variation within non-coding regions like this intron is extremely common and expected (because these do not affect the amino acids of the translated protein and are therefore not subject to functional selection pressure). Therefore, the SNPs in Figure 11 may be attributed to mutations introduced by polymerase amplification as a result of the enormity of PCR replication (this gene was the one that underwent fusion PCR).

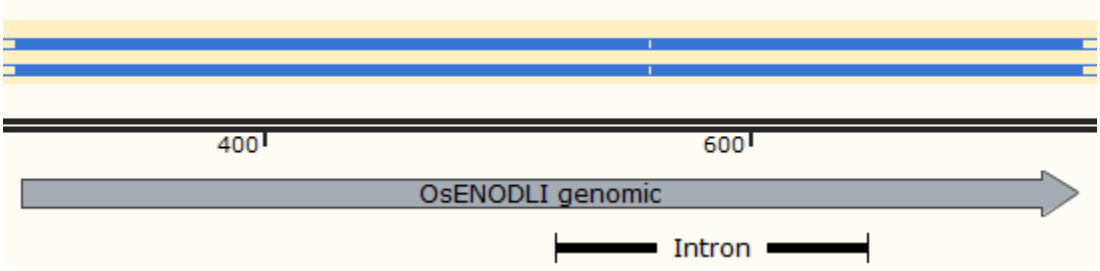

Figure 12: Alignment of Gene Plus intron to Database Genomic DNA. This figure shows the similarity between the kitaake and nipponbare cultivars: only one base pair differs between the two, and this base pair is within the intronic region. This figure also shows that the SNPs in the spliced clone are polymerase error and not cultivar variation.

The protein alignment showed that two of the four SNPs in the gene without intron were significant enough to change the identity of the protein. Amino acid 82 in the database sequence (that matches the database protein per sequencing results) is glycine and the mutant gene encodes for a valine as a result of a SNP. This represents a change from a very small amino acid, to a somewhat larger (though still small) and hydrophobic amino acid. Additionally, amino acid 87 is also changed from a serine in the genomic DNA to an isoleucine in the mutant.

Chapter 5

Future Work

Isolation of a Non-Mutant OsENODLI Clone

The results of the work presented here clearly illustrate that the overall methodology of fusion-based PCR has the potential to produce the clone of interest, however, this is subject to the constraint of PCR-introduce mutations. It is worth noting that the fusion PCR was performed with a non-high fidelity taq. Such errors could be minimized utilizing a high fidelity version of taq such as Phusion or Hi-Fi.

Because the clone with intron was verified to have no SNPs when compared to database genomic DNA, this clone should serve as a template for all future fusion reactions. Amplification from this DNA should be done with a high fidelity polymerase.

On the other hand, a single base change kit could be an option for fixing the two amino-acid-changing bases. However, these are custom designed kits that may be more expensive than just re-doing the fusion PCR.

ELP Intein Purification Method

The protein will be purified using pH and temperature controlled purification that utilizes a combination of self-cleaving intein tag and a solubility tag. Intein proteins were first used to purify proteins analogous to histidine-tag purification. Where the poly-histidine tag binds to a

nickel/cobalt resin for purification, the intein proteins can be purified using chitin resins based on *Bacillus circulans* as an affinity tag at a third of the cost. More importantly, inteins are proteins which undergo self-cleavage so that the protein purification tag can be removed without adding a protease cleavage site (Wu, Mee, Califano, Banki, & Wood, 2006). The mechanism of self-cleavage is given in detail in public domain, where, for example the protein cleaves at its N-terminal at reduced temperature in the presence of the reducing potential of cysteine (Wu et al., 2006). As a result, if the intein tag is placed at the C-terminal of the OsENODLI it can remove itself after purification. Additionally, researchers have taken intein purification a step further, cutting out the chromatography purification altogether, by adding a pH-sensitive solubility tag such as Elastin-Like-protein (ELP) to the C-terminal of the intein (Figure 13: ELP Intein Configuration).

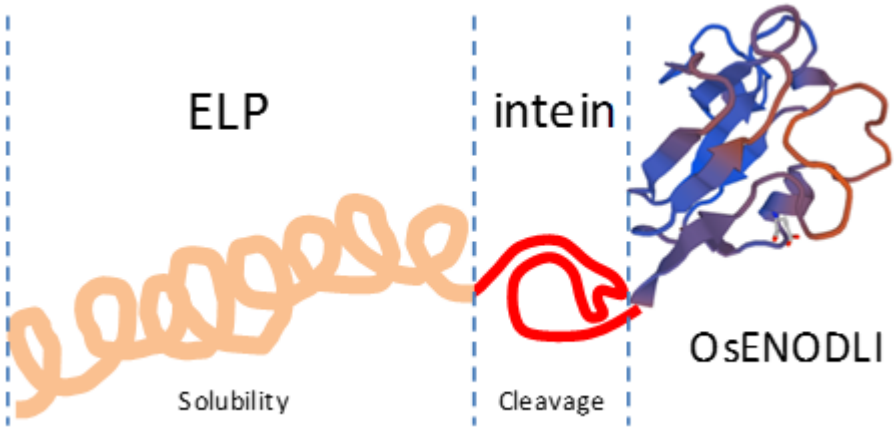

Figure 13: ELP Intein Configuration. The planned low-cost purification scheme for OsENODLI based on a fusion with an ELP solubility tag and a pH activated self-cleaving intein tag.

Thus, an ELP-intein-fused protein will precipitate at the programmed pH, allowing centrifugal purification. Finally the intein tag is cleaved chemically and can be resolubalized and separated from the product (Figure 14). This type of purification is appropriate in context with the demand and scope of use for the protein. The demand for the protein will be very high within

a certain niche of scientists, but will only be required in small quantities. Thus centrifugation is a highly convenient and appropriate method.

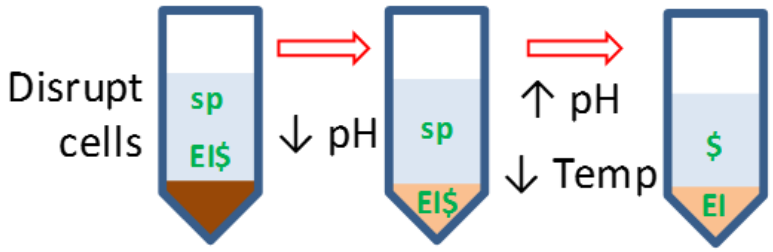

Figure 14: ELP Intein Purification. Low-cost processing flow scheme for purifying the OsENODLI ($) along with at solubility-cleavage tag (EI). Initial soluble expression is separated from cell debris by centrifugation. Reduced pH precipitates the fusion protein to separate it from other soluble protein (sp) and subsequent solubilization with an increase in pH and cleavage with a decrease in temperature facilitates recovery of the purified soluble OsENODLI.

Protein expression data will be monitored via Bradford assays and SDS-PAGE to determine the specific protein production normalized to cost (media inputs). The goal of this project is to make the cheaper clone (*E. coli* or cyanobacteria) available to lead rice researchers around the world. Instead of relying on expensive or time-consuming methods to obtain somatic rice embryos, all they need in the clone from which they can purify their own protein to induce somatic embryos, at only the cost of culturing the organism.

ELP Intein Construct and Cloning

The construct with the DNA to add the ELP intein tag to OsENODLI was given to CurtisLab by Dr. David Wood, Professor of Chemical Engineering at Ohio State University. This construct will be referred to as the pET/ELP-I-CAT vector (Figure 15: pET/ELP-I-CAT). The MCS for this vector is unique in that it cuts into the intein sequence. Care must be taken in designing the primers to insert the gene of interest at the BsrGI site at the end of the intein

sequence. The forward primer for the gene of interest must include the last few basepairs of the intein and the gene must immediately follow the intein sequence. If the sequences do not follow each other exactly one right after the other, the protein of interest will be out of frame or the pH cleavage will not take place.

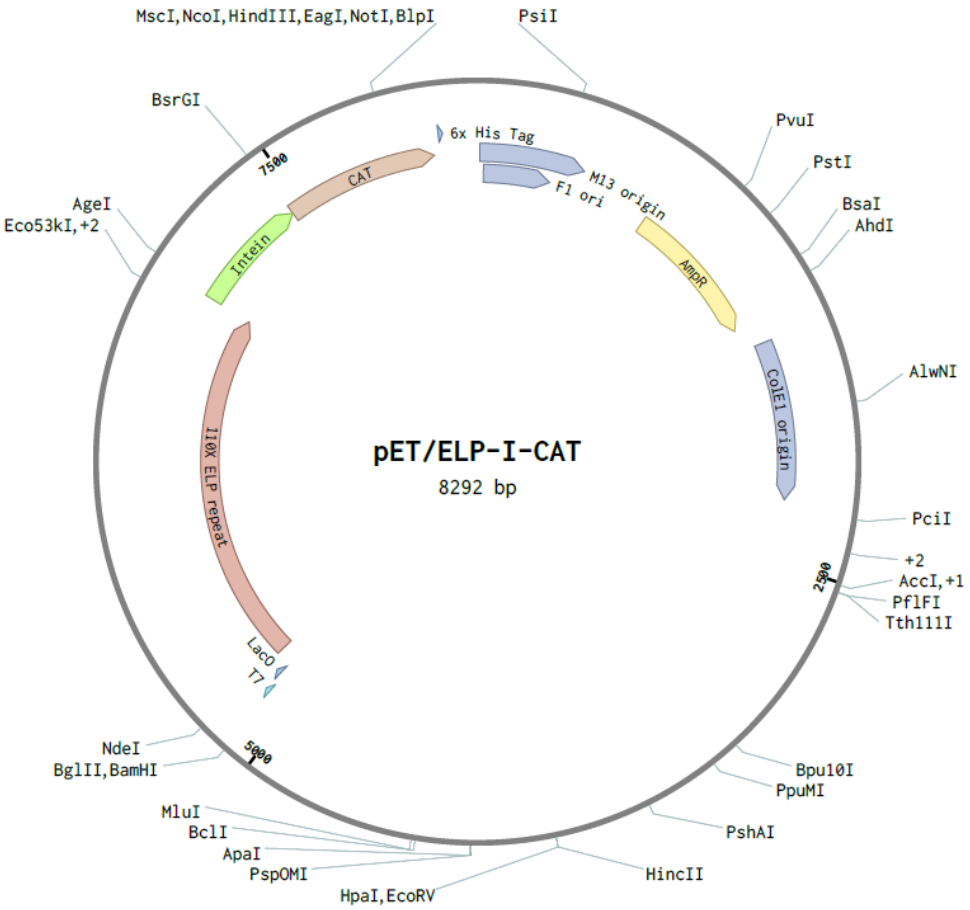

Figure 15: pET/ELP-I-CAT

In the cloning currently taking place, I decided to clone OsENODLI between the BsrGI and BlpI sites. Because BsrGI cuts OsENODLI, the compatible enzyme Acc65I will be used to create sticky ends on the gene. The resulting clone from this procedure will be used to produce and purify OsENODLI for functionality assays.

Completing functionality assays, either through chemical assays such as those typically used for arabinogalactans like the Yariv reagent (Mashiguchi et al., 2004) or through embryogenic properties, are critical to the implementation of OsENODLI. If the protein in its active, correct conformation cannot be isolated post-purification, it may prove to be useless. Further work, out of the scope of my undergraduate thesis, must be completed to ensure the protein is functional after purification.

Appendices

Appendix A: Glossary of Terms and Acronyms

AGP: AragbinoGalactan Protein; a signaling membrane protein

Axenic: a descriptor for a pure organism with no contaminants

cDNA: copy DNA; DNA that is reverse transcribed from RNA; produces a unique library of all genes currently expressed in the tissue from which the RNA was extracted

ELP: elastin-like polypeptide

ENODLI: Early NODulin-Like protein I; an embryogenic protein

Intein: a tag used to purify a protein that can be cleaved by a pH shift

MCS: multiple cloning site; typically in a cloning vector; i.e. “insert gene here”

Monoculture: a culture of a single organism

Os (prefix): referring to *Oryza sativa*

PCR: polymerase chain reaction

PLA: Phytocyanin-Like Arabinogalactan; an embryogenic protein

SE: somatic embryogenesis

SNP: single nucleotide polymorphism, a mistake in DNA that is only one base pair off

Appendix B

Protein and DNA Sequences

Proteins

OsENODLI whole protein as identified by (Mashiguchi et al., 2004)

MEASRRWPYAAWFMAVLGLVAVFSSSEAYVFYAGGRDGWVVDPAESFNYWAERNRFQVNDTIVFLHDDEVGGSVLQVTEGDFDTCSTGNPVQRLEDVA
AGRSVFRFDRSGPFFFISGDEDRCQKGQKLYIIVMAVRPTKPSEAPEPAGAAGPVSSKSWSWQAFPPAGATTPPPLPPSWGSAPEHAQAPGKSSLGGSGGGE
MSRSSSLGAPPPTSGAAGLAGVVASVVVGVLGALLMF

OsENODLI domain that I defined

MYVFYAGGRDGWVVDPAESFNYWAERNRFQVNDTIVFLHDDEVGGSVLQVTEGDFDTCSTGNPVQRLEDVAAGRSVFRFDRSGPFFFISGDEDRCQKGQ
KLYIIVMAVRPTK

Constructs and Inserts

OsENODLI (mRNA)

TACGTCTTCTACGCCGGCGGCCGCGACGGCTGGGTCGTCGACCCCGCCGAGAGCTTCAACTACTGGGCCGAGCGCAACCGGTTCCAGGTGAACGACA
CCATTGTGTTCTTGCACGACGACGAGGTTGGCGGCTCCGTGCTGCAGGTGACGGAGGGGGATTTCGACACGTGCAGCACGGGCAACCCGGTCCAGCG
GCTCGAGGACGTCGCCGCCGGCCGCTCGGTGTTCCGGTTCGACAGGTCCGGCCCTTTCTTCTTCATCAGCGGTGACGAGGACCGGTGCCAGAAGGGG
CAGAAGCTGTACATCATCGTGATGGCGGTGCGCCCGACGAAGTAA

Translation:

MYVFYAGGRDGWVVDPAESFNYWAERNRFQVNDTIVFLHDDEVGGSVLQVTEGDFDTCSTGNPVQRLEDVAAGRSVFRFDRSGPFFFISGDEDRCQKGQ
KLYIIVMAVRPTK

OsENODLI (genomic)

The intron is 104 bp

TCAGAACATCAACAAAGCCCCTAAAACCCCAACCACAACACTGGCAACCACGCCGGCCAACCCTGCCGCGCCGGACGTCGGCGGTGGCGCGCCAAG
CGAGCTGCTGCGCGACATTTCCCCGCCGCCAGAACCACCCAGCGAGCTCTTGCCCGGAGCCTGCGCGTGCTCAGGCGCCGAGCCCCAGGACGGCGGC
AGCGGCGGCGGGGTGGTAGCACCGGCTGGCGGGAACGCCTGCCAGGACCACGACTTGGAAGAGACAGGCCCGGCCGCCCCGGCCGGCTCCGGCGCC
TCGGACGGCTTCGTCGGGCGCACCGCCATCACGATGATGTACAGCTTCTGCCCCTTCTGGCACCGGTCCTCGTCACCGCTGATGAAGAAGAAAGGGCC
GGACCTGTCGAACCGGAACACCGAGCGGCCGGCGGCGACGTCCTCGAGCCGCTGGACCGGGTTGCCCGTGCTGCACGTGTCGAAATCCCCCTCCGTC
ACCTGCAGCACGGAGCCGCCAACCTCGTCGTCGTGCAAGAACACTGCGAGCCAAGAAACGCGCCGCACGTACACTTGAACGATTAAACCAAAAAAA
AAAACACGTAGTTATACATCGATGAGGGGATGATCGAGATGATACACGTACCAATGGTGTCGTTCACCTGGAACCGGTTGCGCTCGGCCCAGTAGTT
GAAGCTCTCGGCGGGGTCGACGACCCAGCCGTCGCGGCCGCCGGCGTAGAAGACGTACGCCTCCGACGATGAAAACACGGCGACCAAACCCAGAAC
GGCCATAAACCAAGCTGCATATGGCCACCTCCTGGAAGCCTCCAT

pET/ELP-I-CAT

OsPLA was cloned into this vector to be expressed with the ELP intein tag.

TGGCGAATGGGACGCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGACCGCTACACTTGCCAGCGCCCTAGCGCCC
GCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCACGTTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTAGTGCTT
TACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGGCCATCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCC
ACGTTCTTTAATAGTGGACTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTATAAGGGATTTTGCCGATTTCGGCCT
ATTGGTTAAAAAATGAGCTGATTTAACAAAAATTTAACGCGAATTTTAACAAAATATTAACGTTTACAATTTCAGGTGGCACTTTTCGGGGAAATGTG
CGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCGCTCATGAGACAATAACCCTGATAAATGCTTCAATAATATTGAAAAAG
GAAGAGTATGAGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCACCCAGAAACGCTGGTGAAAGT
AAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGTTACATCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGT
TTTCCAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACGCCGGGCAAGAGCAACTCGGTCGCCGCATACACTATTC
TCAGAATGACTTGGTTGAGTACTCACCAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCTGCCATAACCATGAGT
GATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCGAAGGAGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTG
ATCGTTGGGAACCGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGCAGCAATGGCAACAACGTTGCGCAAACTATTAA
CTGGCGAACTACTTACTCTAGCTTCCCGGCAACAATTAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTCCGGCT
GGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCATTGCAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGT
TATCTACACGACGGGGAGTCAGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATTAAGCATTGGTAACTGTCAGAC
CAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCATTTTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATC
CCTTAACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTTCTTGAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTG
CAAACAAAAAAACCACCGCTACCAGCGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCTTCAGCAGAGCGCAGA
TACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACTTCAAGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAG
TGGCTGCTGCCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATAAGGCGCAGCGGTCGGGCTGAACGGGGGGTTC
GTGCACACAGCCCAGCTTGGAGCGAACGACCTACACCGAACTGAGATACCTACAGCGTGAGCTATGAGAAAGCGCCACGCTTCCCGAAGGGAGAAA
GGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAGGGAGCTTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGG
GTTTCGCCACCTCTGACTTGAGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGCAACGCGGCCTTTTTACGGTTCC
TGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCGTTATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCG
CCGCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCTGATGCGGTATTTTCTCCTTACGCATCTGTGCGGTATTTCA
CACCGCATATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCGCATAGTTAAGCCAGTATACACTCCGCTATCGCTACGTGACTGGGTCATGGCTG
CGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGTCTGCTCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGC
ATGTGTCAGAGGTTTTCACCGTCATCACCGAAACGCGCGAGGCAGCTGCGGTAAAGCTCATCAGCGTGGTCGTGAAGCGATTCACAGATGTCTGCCT
GTTCATCCGCGTCCAGCTCGTTGAGTTTCTCCAGAAGCGTTAATGTCTGGCTTCTGATAAAGCGGGCCATGTTAAGGGCGGTTTTTTCCTGTTTGGTCA
CTGATGCCTCCGTGTAAGGGGGATTTCTGTTCATGGGGGTAATGATACCGATGAAACGAGAGAGGATGCTCACGATACGGGTTACTGATGATGAACA
TGCCCGGTTACTGGAACGTTGTGAGGGTAAACAACTGGCGGTATGGATGCGGCGGGACCAGAGAAAAATCACTCAGGGTCAATGCCAGCGCTTCGTT
AATACAGATGTAGGTGTTCCACAGGGTAGCCAGCAGCATCCTGCGATGCAGATCCGGAACATAATGGTGCAGGGCGCTGACTTCCGCGTTTCCAGAC
TTTACGAAACACGGAAACCGAAGACCATTCATGTTGTTGCTCAGGTCGCAGACGTTTTGCAGCAGCAGTCGCTTCACGTTCGCTCGCGTATCGGTGAT
TCATTCTGCTAACCAGTAAGGCAACCCCGCCAGCCTAGCCGGGTCCTCAACGACAGGAGCACGATCATGCGCACCCGTGGGGCCGCCATGCCGGCGA
TAATGGCCTGCTTCTCGCCGAAACGTTTGGTGGCGGGACCAGTGACGAAGGCTTGAGCGAGGGCGTGCAAGATTCCGAATACCGCAAGCGACAGGCC
GATCATCGTCGCGCTCCAGCGAAAGCGGTCCTCGCCGAAAATGACCCAGAGCGCTGCCGGCACCTGTCCTACGAGTTGCATGATAAAGAAGACAGTC
ATAAGTGCGGCGACGATAGTCATGCCCCGCGCCCACCGGAAGGAGCTGACTGGGTTGAAGGCTCTCAAGGGCATCGGTCGAGATCCCGGTGCCTAAT
GAGTGAGCTAACTTACATTAATTGCGTTGCGCTCACTGCCCGCTTTCCAGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCG
GGGAGAGGCGGTTTGCGTATTGGGCGCCAGGGTGGTTTTTCTTTTCACCAGTGAGACGGGCAACAGCTGATTGCCCTTCACCGCCTGGCCCTGAGAGA
GTTGCAGCAAGCGGTCCACGCTGGTTTGCCCCAGCAGGCGAAAATCCTGTTTGATGGTGGTTAACGGCGGGATATAACATGAGCTGTCTTCGGTATCG
TCGTATCCCACTACCGAGATATCCGCACCAACGCGCAGCCCGGACTCGGTAATGGCGCGCATTGCGCCCAGCGCCATCTGATCGTTGGCAACCAGCA
TCGCAGTGGGAACGATGCCCTCATTCAGCATTTGCATGGTTTGTTGAAAACCGGACATGGCACTCCAGTCGCCTTCCCGTTCCGCTATCGGCTGAATTT
GATTGCGAGTGAGATATTTATGCCAGCCAGCCAGACGCAGACGCGCCGAGACAGAACTTAATGGGCCCGCTAACAGCGCGATTTGCTGGTGACCCAA
TGCGACCAGATGCTCCACGCCCAGTCGCGTACCGTCTTCATGGGAGAAAATAATACTGTTGATGGGTGTCTGGTCAGAGACATCAAGAAATAACGCC
GGAACATTAGTGCAGGCAGCTTCCACAGCAATGGCATCCTGGTCATCCAGCGGATAGTTAATGATCAGCCCACTGACGCGTTGCGCGAGAAGATTGT
GCACCGCCGCTTTACAGGCTTCGACGCCGCTTCGTTCTACCATCGACACCACCACGCTGGCACCCAGTTGATCGGCGCGAGATTTAATCGCCGCGACA
ATTTGCGACGGCGCGTGCAGGGCCAGACTGGAGGTGGCAACGCCAATCAGCAACGACTGTTTGCCCGCCAGTTGTTGTGCCACGCGGTTGGGAATGT
AATTCAGCTCCGCCATCGCCGCTTCCACTTTTTCCCGCGTTTTCGCAGAAACGTGGCTGGCCTGGTTCACCACGCGGGAAACGGTCTGATAAGAGACA
CCGGCATACTCTGCGACATCGTATAACGTTACTGGTTTCACATTCACCACCCTGAATTGACTCTCTTCCGGGCGCTATCATGCCATACCGCGAAAGGTT
TTGCGCCATTCGATGGTGTCCGGGATCTCGACGCTCTCCCTTATGCGACTCCTGCATTAGGAAGCAGCCCAGTAGTAGGTTGAGGCCGTTGAGCACCG
CCGCCGCAAGGAATGGTGCATGCAAGGAGATGGCGCCCAACAGTCCCCCGGCCACGGGGCCTGCCACCATACCCACGCCGAAACAAGCGCTCATGA
GCCCGAAGTGGCGAGCCCGATCTTCCCCATCGGTGATGTCGGCGATATAGGCGCCAGCAACCGCACCTGTGGCGCCGGTGATGCCGGCCACGATGCG
TCCGGCGTAGAGGATCGAGATCTCGATCCCGCGAAATTAATACGACTCACTATAGGGGAATTGTGAGCGGATAACAATTCCCCTCTAGGATCCAATA
ATTTTGTTTAACTTTAAGAAGGAGATATACATATGGGCCACGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGTGGCGGTGTGCCGGGCGCAGGTGTT
CCTGGTGTAGGTGTGCCGGGTGTTGGTGTCCGGGTGCAGGCGTTCCTGGTGTAGGTGTGCCGGGTGGCGGTGTGCCGGGCGT
GGGTGTTCCGGGCGTGGGTGTTCCGGGTGCCGGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGTGTTGGTGTAC
CAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGTGGCGGTGTGCCGGGCGC
AGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGC
CGGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGTGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGCGCAGGTGTT
GGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGCGGTGTGC
CGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGC
GGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGTGGCGGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCC
GGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGTGGC
GGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCC
GGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGTGGCGGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTT
GGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGGTGTTC
CGGGTGGCGGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGC
AGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCGTGGGTGTTCCGGGCGCAGGTGTTCCTGGTGTAGGTGTG
CCGGGTGTTGGTGTGCCGGGTGTTGGTGTACCAGGTGGCGGTGTTCCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGTGGGTGTTCCGGGCG
TGGGTGTTCCGGGTGGCGGTGTGCCGGGCGCAGGTGTTCCTGGTGTAGGTGTGCCGGGTGTTGGTGTGCCCGGGTGTTGGTGTACCAGGTGGCGGTGTT
CCGGGTGCAGGCGTTCCGGGTGGCGGTGTGCCGGGCGGGCTGGTGAGCTCGAACAACAACAACAATAACAATAACAACAACCTCGGGATCGAGGGA
AGGATTTCAGAATTCGCCCTCGCAGAGGGCACTCGGATCTTCGATCCGGTCACCGGTACAACGCATCGCATCGAGGATGTTGTCGgTGGGCGCAAGCC
TATTCATGTCGTGGCTGCTGCCAAGGACGGAACGCTGCATGCGCGGCCCGTGGTGTCCTGGTTCGACCAGGGAACGCGGGATGTGATCGGGTTGCGG
ATCGCCGGTGGCGCCATCcTGTGGGCGACACCCGATCACAAGGTGCTGACAGAGTACGGCTGGCGTGCCGCCGGGGAACTCCGCAAGGGAGACAGG
GTGGCGCAACCGCGACGCTTCGATGGATTCGGTGACAGTGCGCCGATTCCGGCGCGCGTGCAGGCGCTCGCGGATGCCCTGGATGACAAATTCCTGC
ACGACATGCTGGCGGAAGAACTCCGCTATTCCGTGATCCGAGAAGTGCTGCCAACGCGGCGGGCACGAACGTTCGGCCTCGAGGTCGAGGAACTGCA
CACCCTCGTCGCCGAAGGGGTTGTTGTACACAACATGGAGAAAAAAATCACTGGATATACCACCGTTGATATATCCCAATGGCATCGTAAAGAACAT

TTTGAGGCATTTCAGTCAGTTGCTCAATGTACCTATAACCAGACCGTTCAGCTGGATATTACGGCCTTTTTAAAGACCGTAAAGAAAAATAAGCACAA
GTTTTATCCGGCCTTTATTCACATTCTTGCCCGCCTGATGAATGCTCATCCGGAATTTCGTATGGCAATGAAAGACGGTGAGCTGGTGATATGGGATA
GTGTTCACCCTTGTTACACCGTTTTCCATGAGCAAACTGAAACGTTTTCATCGCTCTGGAGTGAATACCACGACGATTTCCGGCAGTTTCTACACATAT
ATTCGCAAGATGTGGCGTGTTACGGTGAAAACCTGGCCTATTTCCCTAAAGGGTTTATTGAGAATATGTTTTTCGTCTCAGCCAATCCCTGGGTGAGTT
TCACCAGTTTTGATTTAAACGTGGCCAATATGGACAACTTCTTCGCCCCCGTTTTCACCATGGGCAAATATTATACGCAAGGCGACAAGGTGCTGATG
CCGCTGGCGATTCAGGTTCATCATGCCGTCTGTGATGGCTTCCATGTCGGCAGAATGCTTAATGAATTACAACAGTACTGCGATGAGTGGCAGGGCGG
GGCGTAATGAGCAAGCTTGCGGCCGCACTCGAGCACCACCACCACCACCACTGAGATCCGGCTGCTAACAAAGCCCGAAAGGAAGCTGAGTTGGCT
GCTGCCACCGCTGAGCAATAACTAGCATAACCCCTTGGGGCCTCTAAACGGGTCTTGAGGGGTTTTTTGCTGAAAGGAGGAACTATATCCGGAT

Custom Designed Primers

Rand_fwd; Rand_rev

Amplifies a 754bp region from a histidine kinase that is specific to our BBrady (primer blast showed no other amplicons within the whole family Bradyrhizobiciae).

fwd: CGGCAGCGGCGTTAGGTATTC

rev: AAATTGCCCGATCTCTTGCCTCATC

fusionPCR_fwd

Amplifies a 117 bp region from the OsENODLI gene when paired with PCR2_rev.

GTGAACGACACCATTGTGTTCTTG

fusionPCR_rev

Amplifies a 246 bp region from the OsENODLI gene when paired with PCR2_fwd.

CAAGAACACAATGGTGTCGTTCAC

OsPLA_mod_e+c_fwd

For adding the NdeI recognition site to OsENODLI with an additional 3 bp to optimize restriction digest efficiency. Used for cloning OsENODLI into pET14b, pAQ1, and pAQ3.

TATCATATGTACGTCTTCTACGCCGG

OsPLA_mod_e+c_rev

For adding the BamHI recognition site to OsENODLI with an additional 3 bp to optimize restriction digest efficiency. Used for cloning OsENODLI into pET14b, pAQ1, and pAQ3

ACAGGATCCTTACTTCGTCGGGCGCAC

PCR2

Amplifies a 450 bp region from rice gDNA. A primer set with complete homology to OsENODLI. Used to amplify the gene from genomic DNA.

fwd: GAGGCGTACGTCTTCTACGC

rev: CGTCGGGCGCACCGCCAT

OsPLA_int_rev

CACCTGGAACCGGTTGC

OsPLA int fwd

CTTCAACTACTGGGCCG

Universal Primers

T7 rev

GCTAGTTATTGCTCAGCGG

T7 fwd

TAATACGACTCACTATAGGG

Appendix C: Additional Protocols

Heat Shock Transformation

This is the protocol I have successfully used multiple times to transform chemically competent cells. It is a compilation of many open source protocols that may be found on the Drive and sites like OpenWetware.

1. Put frozen cells on ice and aliquot ligation reaction mixture into tube immediately.

Note: you may add up to 10% of total competent cell volume of your ligation reaction

2. When cells have thawed (~1 minute), gently flick tube to mix.
3. Set a timer for 30 minutes, turn on hot water bath to 42°C.
4. After 30 minutes, heat shock cells for 45 s at 42°C. Do not jiggle cells once they have entered the water bath.
5. Recover cells on ice for two minutes.
6. Resuspend cells in 500 µL of SOC.
7. Incubate epitube on a 37°C shaker for 1 hour so that the resistance genes may be expressed.
8. Plate entire reaction on selective plate.

LB Amp 100 Selective Medium

10 g tryptone

5 g yeast extract

10 g NaCl

1.5 g agar

100 µL ampicillin (100 mg/mL stock)

BIBLIOGRAPHY

Arey, B. J. (2012). The Role of Glycosylation in Receptor Signaling. *Glycosylation*, 273 – 286. http://doi.org/10.5772/50262

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. a., Dvorkin, M., Kulikov, A. S., … Pevzner, P. a. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. http://doi.org/10.1089/cmb.2012.0021

Bawono, P., Heringa, J., & Editor, D. J. R. (2014). Multiple Sequence Alignment Methods. *Methods in Molecular Biology (Clifton, N.J.)*, *1079*, 245–62. http://doi.org/10.1007/978-1-62703-646-7

Boers, S. A., Burggrave, R., van Westreenen, M., Goessens, W. H. F., & Hays, J. P. (2014). Whole-genome mapping for high-resolution genotyping of Pseudomonas aeruginosa. *Journal of Microbiological Methods*, *106*, 19–22. http://doi.org/10.1016/j.mimet.2014.07.020

Life Technologies. (2012). TRIzol ® Reagent, (15596026), 18–21. http://doi.org/10.1101/pdb.caut2701

Mashiguchi, K., Yamaguchi, I., & Suzuki, Y. (2004). Isolation and identification of glycosylphosphatidylinositol-anchored arabinogalactan proteins and novel beta-glucosyl Yariv-reactive proteins from seeds of rice (Oryza sativa). *Plant & Cell Physiology*, *45*(12),

1817–1829. http://doi.org/10.1093/pcp/pch208

NEB. (2015). Traditional Cloning Quick Guide, 1–3.

Poon, S., Heath, R. L., & Clarke, A. E. (2012). A chimeric arabinogalactan protein promotes somatic embryogensis in cotton cell culture. *Plant Physiology*, *160*(2), 684–95. http://doi.org/10.1104/pp.112.203075

Roberts, R. J., Carneiro, M. O., & Schatz, M. C. (2013). The advantages of SMRT sequencing. *Genome Biology*, *14*(6), 405. http://doi.org/10.1186/gb-2013-14-6-405

Tropea, J. E., Cherry, S., & Waugh, D. S. (2009). High Throughput Protein Expression and Purification. *Methods in Molecular Biology (Clifton, N.J.)*, *498*(1), 297–307. http://doi.org/10.1007/978-1-59745-196-3

Wu, W.-Y., Mee, C., Califano, F., Banki, R., & Wood, D. W. (2006). Recombinant protein purification by self-cleaving aggregation tag. *Nature Protocols*, *1*(5), 2257–62. http://doi.org/10.1038/nprot.2006.314

Yoo, J. (2013). *Establishment and maintenance of axenic Botryococcus braunii race B algae culture*. The Pennsylvania State University. Retrieved from https://honors.libraries.psu.edu/paper/19923/

Zimmermann, P., Hirsch-hoffmann, M., Hennig, L., & Gruissem, W. (2004). GENEVESTIGATOR . Arabidopsis Microarray Database and Analysis Toolbox 1 [w], *136*(September), 2621–2632. http://doi.org/10.1104/pp.104.046367.1

.

ACADEMIC VITA

Academic Vita of Katie Legenski
katie.legenski@gmail.com

B.S. Chemical Engineering (Bioprocess Option)
Honors: Chemical Engineering

Thesis Title: Rice Embryogenic Protein Production and Distribution
Thesis Supervisor: Wayne Roger Curtis

Work Experience

Spring 2016

Penn State **teaching assistant**, Ch E 480: Unit Operations Practicum; TA for distillation, fluid flow, and heat exchange labs, grade course work
Supervisor: Dr. Mechteld Hillsley

Fall 2015

Penn State **grader**, Ch E 340: Introduction to Biomolecular Engineering; teach in professor absence, hold help sessions, proctor exams, grade exams and home works
Supervisor: Dr. Howard Salis

Summer 2015

Penn State Biofellow and **National Science Foundation fellow**, University Park: bioinformatician, plant and molecular biologist research intern
Supervisor: Dr. Wayne Curtis

Summer 2014

National Science Foundation research experience for undergraduate recipient, University of Kansas: environmental and civil engineering research intern
Supervisor: Dr. Belinda Sturm

May-December 2013

Peer tutor, learning assistant (chemistry, mathematics, physics)
Supervisor: Mrs. Lisa Zackowski

Grants Received

Schreyer Travel Ambassador Grantee, February 2014

Awards:

Professional Memberships: American Institute of Chemical Engineers, Tau Beta Pi, Society of Women Engineers

Presentations:

Middle Atlantic Discovery Chemistry Project presenter, May 2013

Bioinformatics and Genomics Retreat poster presenter, Penn State September 2015

2nd place: 2015 AIChE national student conference poster competition; Biotech division VII

Community Service Involvement:

Tau Beta Pi community service chair, August 2015-present

Schreyer Honors College alumni event liaison, January 2014-present

Peer mentor and orientation leader for STEM students (Penn State Berks), January-December 2013

Benefit Piano Concert Series coordinator, August 2012-December 2013

Honors Club community service chair, August 2012-December 2013

International Education:

March 2013: ENGL 297H Istanbul: Orientalism and the Other; week long study abroad

March 2014: ENGL 297H Health in South Africa: week long service learning project in Johannesburg

January-May 2015: Semester study abroad at the National University of Singapore