OEP15-1 and OEP7.3 Localize to the Outer Envelope of Chloroplasts Using a Novel and Uncharacterized Targeting Pathway

by

Alyssa Kate Overton

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Master of Science

in

Biology

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Chloroplasts originated from an endosymbiotic event where a Gram-negative cyanobacterium was engulfed by an ancestral eukaryotic cell. The genome of the endosymbiont has since gone through extensive gene transfer to the host cell nucleus during the evolutionary transition to an organelle. The majority of the plastid proteome is now encoded in the nucleus of plant cells. This means that plastid targeted proteins are translated in the host cytosol and must be translocated across the two membranes surrounding the plastids before they can perform their function. Stroma-targeted proteins posses transit peptides that are recognized for translocation by the TOC and TIC complexes. The proteins which make up the TOC complex are just some of a wide variety of OEPs which must also be targeted to the chloroplast outer membrane to perform their functions, but most OEPs use targeting strategies which do not involve transit peptides.

Novel strategies for outer envelope targeting and localization are still being added to our understanding of OEPs. In the Chuong and Smith labs, a C-terminal TP-like signal was recently added to the list of known pathways and was discovered in TOC159. In an effort to test other OEPs which may posses a similar signal all OEPs were input into an N-terminal TP prediction tool, ChloroP, in their reverse orientation. The protein annotated at At4G02482 at the time of this ChloroP analysis was OEP15-1 and was one of the proteins which scored a high enough likelihood of C-terminal TP presence to be considered a candidate for this novel pathway.

The research outlined in this thesis aims to determine the localization pathway or targeting strategy of OEP15-1 and the more recently annotated gene product of the same accession number, OEP7.3. The majority of the thesis focusses on the expression patterns of transiently expressed recombinant fusion proteins in onion epidermal cells and *A. thaliana* protoplasts which were analyzed by fluorescence microscopy and Western blotting. The

expression patterns confirm that OEP7.3 and OEP15-1 are both targeted to the outer envelope of chloroplasts and suggest that they may achieve this using β-strands in a pathway similar to both mitochondria-targeting β-barrel proteins and a few studied OEP β-barrel proteins.

It is not currently known if both proteins annotated at At4G02482 are produced in *A. thaliana,* but the data in this thesis suggests that they are. The function(s) of OEP15-1 and OEP7.3 is also unknown, but bioinformatic analyses described in this thesis point towards OEP15-1 existing as a β-barrel protein in the outer envelope of chloroplasts which may be involved in cross-membrane transport.

Acknowledgments

I would first like to thank my co-supervisory team of Dr. Simon Chuong and Dr. Matthew Smith for their continual support and their limitless answers to my limitless questions. Your generosity with lab spaces, resources and time made this roller coaster of a research project manageable. Thank you both for sharing all your wisdom with me, and for providing my constructs, both ordered and hand-cloned.

Next, I would like to thank my committee members Dr. Barb Moffatt and Dr. Susan Lolle for their support and flexibility as this project took shape. You are both so good at making me feel hopeful when my project takes sharp turns by supplying new points of view. Thank you both for reminding me to have fun with this thesis. I think I did, and I hope you enjoy it!

I would also like to thank all members of Chuong, Smith, Moffatt and Lolle lab for camaraderie and creating a rich environment for exchanging ideas. Thank you Nilanth Yogadasan and George Zhou for training me in techniques. Nilanth and Delaney Nash for answering so many questions in the lab. Thank you, Delaney, for being a sounding board for all my wild ideas towards the end. Thank you, Mike Fish, for helping me orient myself at Laurier, and Saeer Adeel for so much plant related advice. Thank you as well to Kelly Yeung who began exploration into the proteins of interest in this thesis and generated the construct OEP15-1NT:GFP.

Lastly but not least I would like to thank my parents Debby and Rod Overton for their ongoing love and support always and for keeping my student stress low during this journey, as well as Alex Ng for his love and support during a whirlwind of emotions as this project evolved.

Table of Contents

Chapter 1. Introduction

Chapter 2. Materials and Methods

Chapter 3. Results

Chapter 4. Discussion

List of Figures

List of Tables

List of Abbreviations

AKR2A – Ankyrin Repeat Protein 2A

AP – Adaptor Protein

ARD – Ankyrin Repeat Domain

ATP – Adenosine TriPhosphate

BLAST – Basic Local Alignment Search Tool

BSA – Bovine Serum Albumin

cDNA – complementary DNA

CD – Circular Dichroism

CDS – Coding Sequence

(C)OM – (Chloroplast) Outer Membrane

C-terminus – Carboxyl Terminus

DHAP – DiHydroxyAcetone Phosphate

DNA – DeoxyriboNucleic Acid

(E)GFP – (Enhanced) Green Fluorescence Protein

FdTP – Ferredoxin Transit Peptide

G3P – Glyceraldehyde-3-Phosphate

GTP – Guanidine Triphosphate

hFH – human Fumarate Hydratase

IMS – Inter Membrane Space

MGD1 – Monogalactosydiacylglycerol synthase 1

MGDG – MonoGalactosylDiacylGlycerol

mRNA – messenger RNA

NADPH – Nicotinamide Adenine Dinucleotide Phosphate

NCBI – National Center for Biotechnology Information

NPC – Nuclear Pore Complex

N-terminus – Amino Terminus

pBS – pBluescript vector

PCR – Polymerase Chain Reaction

PEG – PolyEthylene Glycol

PG – PhosphatidylGlycerol

PGA – 3-PhosphoGlyceric Acid

PPDB – Plant Protein DataBase

PSI – PhotoSystem 1

PSII – PhotoSystem 2

RbcS – small subunit of Rubisco

RbcL – large subunit of Rubisco

RE – Restriction Enzyme

RNA – RiboNucleic Acid

RT-PCR – Reverse Transcriptase Polymerase Chain Reaction

RuBisCO – Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase

RuBP – Ribulose-1,5-BisPhosphate

rTP – reverse Transit Peptide

SA – Signal Anchored

SAM – Sorting and Assembly Machinery complex

SEC – general SECretion pathway

SRP – Signal Recognition Particle

TA – Tail Anchored

TAIR – The Arabidopsis Information Resource

TAT – Twin-Arginine Translocon

TIC – Translocon at the Inner envelope of Chloroplasts

TMD – Trans-Membrane Domain

TOC – Translocon at the Outer envelope of Chloroplasts

TP – Transit Peptide

UTR – Un-Translated Region

WT – Wild Type

Chapter 1. Introduction

1.1 Plastids

1.1.1 Plastid Origin

Plastids are a diverse group of plant cell organelles, which include chlorophyll-containing chloroplasts, the organelle responsible for photosynthesis (Thomas et al., 2009; Wise and Hoober, 2006). Chloroplasts and other plastid types also perform roles in other metabolic pathways such as the synthesis of certain lipids and amino acids (Thomas et al., 2009). The evolution of chloroplasts, as well as other plastids, is thought to have begun with an endosymbiotic relationship, established when an ancestral eukaryotic single-celled organism phagocytosed a cyanobacterial cell (Lopez-Juez and Pyke, 2004). Their phagocytotic origin provides an explanation for why chloroplasts are surrounded by two membranes (Jarvis et al., 2008). There has been some debate on the origin of these two membranes. Past theories posit that one membrane originated from a single bacterial membrane and the other from the original phagosome membrane or that both membranes originated from gram-negative bacteria (Wise and Hoober, 2006). Due to the similarity in lipid composition between the two membranes themselves, as well as residual peptidoglycan synthesizing machinery that is located just underneath the membranes, it is now widely accepted that both membranes arose from the ancestral gram-negative cyanobacterial endosymbiont (Gould et al., 2008).

The endosymbiotic origin of plastids is similar to that of mitochondria, the other double membrane surrounded organelle in plant cells (and all eukaryotes). Mitochondria evolved instead from an α-proteobacteria that was engulfed by an archaeal host in the event that established the eukaryotic lineage (Garg and Gould, 2016). Due to the mitochondria's earlier evolutionary origin

and therefore their presence in eukaryotes other than plants, they are more studied than plastids, and can inform hypotheses in some cases when studying plastids (Garg and Gould, 2016).

1.1.2 Plastid Types and Development

Different cell types within plants contain and support many copies of one of a large variety of plastid types, each of which carries out a unique function (Wise and Hoober, 2006). All nucleated cells in a plant contain plastids. All plastid types develop from proplastids, which are undifferentiated precursors without a defined function and with minimal internal structure or organization (Pogson and Albrecht-Borth, 2014). Along with host cell type, the fate of these proplastids' development is also sometimes defined by environmental cues. For example, if a plant is grown in the dark, the proplastids of shoot or leaf cells will develop into etioplasts, which are devoid of pigments (Pogson and Albrecht-Borth, 2014). There are different pigment-containing plastids which are collectively known as chromoplasts of which chloroplasts are the most well-known (Jarvis, 2008; Taiz and Zeiger, 2006). In most plants, proplastids in the shoots and leaves will develop directly into chloroplasts if light is present (Wise and Hoober, 2006). In storage organs such as roots and seeds, proplastids often differentiate into colourless plastids called leucoplasts which include amyloplasts that store starch, elaioplasts that store oil, or proteinoplasts that store proteins (Lopez-Juez and Pyke, 2004; Wise and Hoober, 2006). In columella cells at the tips of roots there are special amyloplasts called statoliths whose starch granules fall in response to gravity allowing the roots to sense gravity and grow in its direction (Edelmann, 2018).

Once differentiated, plastids can re-differentiate into other plastid types (Taiz and Zeiger, 2006). For example, old chloroplasts in the process of senescence will re-differentiate into a plastid type specialized in dismantling the photosynthetic machinery, called a gerontoplast

(Thomas et al., 2009). An etioplast or amyloplast exposed to light will produce chlorophyll and differentiate into a chloroplast so the plant can photosynthesize and therefore continue to grow under the new source of light (Taiz and Zeiger, 2006). Differentiated plastids can also duplicate through binary fission, using cellular machinery and a mechanism that is very similar to the bacterial cells from which they originated. This ensures there are adequate numbers of mature chloroplasts that can be distributed evenly between daughter cells when the host plant cell undergoes cell division (Pogson and Albrecht-Borth, 2014; Wise and Hoober, 2006).

1.2 Chloroplast Structure and Function

A chloroplast's structure supports its primary function, photosynthesis. The two outer membranes of the chloroplast define an internal soluble compartment called the stroma (Nelson and Ben-Shem, 2004; Taiz and Zeiger, 2006). Within the stroma many thylakoids are arranged in grana and are connected by lamellae. The thylakoid membranes also define a soluble compartment called the lumen (Taiz and Zeiger, 2006). A diagram of these structures can be seen in Figure 1.1.

1.2.1 Light-Dependent Reactions of Photosynthesis

Photosynthesis can be divided into two sets of reactions. The first, termed the light-dependent reactions, take place in the thylakoid membranes (Nelson and Ben-Shem, 2004). In the light-dependent reactions, a series of pigments and proteins held in close association within the thylakoid membrane absorb energy from light and transform it into energy that is accessible to the organelle, ATP and NADPH. There are four protein complexes which function in the light- dependent chain of reactions: Photosystems I and II (PSI and PSII, respectively), the
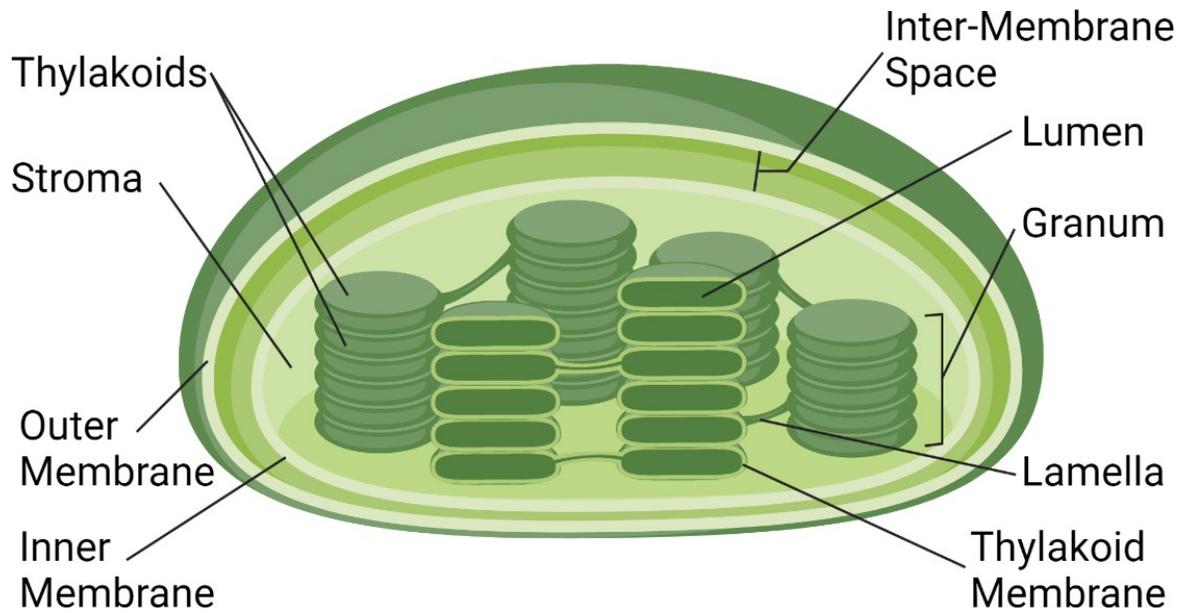
Figure 1.1: Chloroplast Structure. A diagram of a chloroplast showing membranes and soluble compartments. Created with BioRend

cytochrome-b_6f complex and an ATP synthase (Taiz and Zeiger, 2006). PSII reduces water to oxygen and protons, releasing electrons which, when excited by light energy harvested by PSII, pass through plastoquinone, the cytochrome-b_6f complex and plastocyanin to reach PSI (Nelson and Ben-Shem, 2004). Light energy captured by pigments in PSI is concentrated at two chlorophyll molecules at its reaction centre which function as primary electron donors and reduce a ferredoxin molecule in the chloroplast stroma (Taiz and Zeiger, 2006). This ferredoxin is in turn used to reduce $NADP^+$ to NADPH by ferredoxin-$NADP^+$ reductase. As high energy electrons move through the two photosystems and the cytochrome-b_6f complex via a series of redox reactions, protons are released into the lumen of the thylakoid producing an electrochemical potential gradient called the Proton-Motive Force (Taiz and Zeiger, 2006) which is utilized by chloroplast ATP synthase in the thylakoid membrane to drive ATP synthesis. The resulting NADPH and ATP molecules which are generated in the stroma are utilized by the second set of photosynthetic reactions, referred to as the Calvin-Benson cycle (Nelson and Ben-Shem, 2004).

1.2.2 Calvin-Benson Cycle of Photosynthesis

The second set of reactions, collectively called the Calvin-Benson cycle (or the reductive pentose phosphate cycle), take place in the stroma of the chloroplast and are dependent on the ATP and NADPH generated as a result of the light-dependent reactions to fuel fixation of carbon from CO_2 into carbohydrates, which function as energy storage molecules (Nelson and Ben-Shem, 2004; Biel and Fomina, 2015). These reactions are not directly dependent on light so they have often been referred to as the light-independent reactions of photosynthesis (West-Eberhard et al., 2011).

The Calvin-Benson cycle can be broken into three phases (Taiz and Zeiger, 2006). First, the CO_2 acceptor protein Ribulose-1,5-Bisphosphate Carboxylase Oxygenase (Rubisco) fixes carbon to a molecule of ribulose-1,5-bisphophate (RuBP), creating an unstable 6-carbon molecule which is quickly split into two stable molecules of 3-phosphoglycerate (3-PGA) (Calvin and Bassham, 1955; Taiz and Zeiger, 2006). Next, the 3-PGA is reduced to glyceraldehyde-3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP) at the cost of ATP and NADPH (Biel and Fomina, 2015; Taiz and Zeiger, 2006). G3P and DHAP are 3-carbon carbohydrates which are processed by a series of enzymes to yield 5-carbon carbohydrates ending in the formation of ribulose-5-phosphate which in the third stage is regenerated by phosphorylation at the cost of ATP to regenerate the CO_2 acceptor RuBP (Biel and Fomina, 2015; Calvin and Bassham, 1955; Taiz and Zeiger, 2006). G3P can be converted into sucrose molecules which are allocated to other parts of the plant or converted into starch in by the chloroplast and stored there for later distribution (Taiz and Zeiger,2006). To ensure the continual turning of the cycle, of the 6 molecules of G3P produced from every 3 CO_2 molecules fixed, only one will be used to produce sucrose, the other 5 are recycled to regenerate 3 RuBP molecules which will in turn each have a CO_2 molecule linked to them beginning the cycle again (Calvin and Bassham, 1955; Taiz and Zeiger, 2006).

All photosynthetic organisms fix carbon via the Calvin-Benson cycle, but once the CO_2 is fixed into an organic molecule, the metabolism of that molecule can vary resulting in different carbohydrate molecules as a final product (Biel and Fomina, 2015).

1.2.3 Light-Dependent and Light-Independent Reaction Separation

The light-dependent and light-independent reactions of photosynthesis are not spatially separated in so-called C3 plants, which employ this original version of Calvin-Benson cycle

(West-Eberhard et al., 2011). Though Rubisco is primarily a carboxylase, it also has affinity for O_2, and the resulting oxygenase reaction of Rubisco leads to a loss of carbon that had been previously fixed by the plant (Taiz and Zeiger, 2006). Rubisco is naturally exposed to both carbon dioxide and oxygen in the atmosphere in C_3 plants. Several strategies have evolved in groups of land plants to keep oxygen away from Rubisco and concentrate carbon dioxide for fixation (West-Eberhard et al., 2011) as a way of minimizing the oxygenase reaction. CAM plants (crassulacean acid metabolism) and C_4 plants both fix carbon twice, first by the enzyme phosphoenolpyruvate carboxylase (PEPC), which does not react with oxygen, to produce four-carbon molecules such as malate (West-Eberhard et al., 2011). CAM plants are found in hot and arid environments where stomata of leaves must be kept closed to prevent water loss during the day; so, they have evolved a temporal separation of oxygen and Rubisco (West-Eberhard et al., 2011). PEPC is active at night and the malate is converted into malic acid which is stored in the vacuoles until daytime when Rubisco is active but no gas exchange is occurring, allowing it to re-fix the carbon released from the malic acid without the interference of oxygen molecules (West-Eberhard et al., 2011). C_4 plants evolved a spatial separation of oxygen and Rubisco. PEPC fixes carbon into malate in the mesophyll cell layer which performs gas exchange as well. The malate then diffuses into the bundle sheath cells deep in the leaf which are not involved in gas exchange and in which rubisco is expressed (West-Eberhard et al., 2011). In both CAM and C_4 systems, the release of CO_2 from malate or malic acid re-generates the pyruvate molecule for PEPC to carboxylate (Taiz and Zeiger, 2006).

1.3 Protein Trafficking in Plant Cells

The majority of the genetic material of eukaryotic cells is contained centrally in the nucleus (Jarvis, 2008; Taiz and Zeiger, 2006). Cellular organelles contain many proteins which

are encoded by nuclear genes and must be directed to the organelle from the cytosol where they are translated (Lee and Hwang, 2014). Depending on the target organelle, proteins are trafficked through the cell in widely different ways, some of which are more well investigated than others.

1.3.1 Endomembrane System

For some organelles, protein trafficking is achieved through the movement of endomembrane vesicles which arise from the endoplasmic reticulum (ER) after proteins are co-translationally incorporated into the organelle (Lee and Hwang, 2014). Organelles to which proteins are delivered in this fashion are part of the endomembrane system and include the nuclear envelope, ER, Golgi, lysosomes and the vacuole (Gould et al., 2016; Garg and Gould, 2016). The endomembrane system is unique to and a common feature between eukaryotes. Though protein targeting to compartments of the endomembrane system is mostly studied in mammals and yeast, studies examining endomembrane protein trafficking in plants revealed many similarities between mammals and plants. Such similarities include the targeting sequences of cargo proteins and proteins involved with recognition (Wang et al., 2014).

The first step in the targeting pathway for proteins targeted to most organelles in the endomembrane system is co-translational insertion of the polypeptide through the multi-component SEC (general secretion pathway) complex into the ER (Garg and Gould, 2016; Natale et al., 2008). Proteins that are not destined to remain in the ER are then sorted into coat protein complex transport vesicles which deliver them to the Golgi apparatus (Wang et al., 2014). After the Golgi apparatus, soluble proteins are trafficked via a variety of endosomal vesicles which deliver proteins to specific compartments of the endomembrane system. For example, on its way to the vacuole a protein is passed through a series of endosomes ending in the pre-vacuolar compartment (Wang et al., 2014). On the membrane of the pre- and post-Golgi

8

vesicles are adaptor proteins (APs) which, in complexes with each other, control cytosolic proteins' journeys by translocating soluble proteins into their vesicle or catalyzing coat protein insertion into the membrane (Hwang and Robinson, 2009; Lee and Hwang, 2014). Differently composed AP complexes will reside in different endosomes and recognize proteins via targeting signals which are unique to each destination within a cell but which, interestingly, do not vary greatly between lineages of eukaryotic cells (Wang et al., 2014). For example, some tonoplast targeted proteins including *Arabidopsis* vacuolar ion transporter1 (VIT1), a protein with many transmembrane domains (TMDs), are recognized by the presence of a dileucine motif which is conserved in yeast and mammalian cells as well (Wang et al., 2014).

Secreted proteins follow a trafficking system which employs the endomembrane system and many excreted proteins are trafficked via a pathway homologous to the SEC mediated endomembrane pathway described previously (Garg and Gould, 2016). In some cases, protein excretion is instead achieved through the twin-arginine translocon (TAT) system which contains 2-3 membrane integrated proteins and allows translocation of mainly co-factor bound proteins (Garg and Gould, 2016; Natale et al., 2008). Which pathway a protein will take is informed by which membrane the protein must cross. The SEC transports unfolded proteins across the cell membrane while the TAT pathway tends to be involved in the transport of folded proteins across the cell membrane. A version of the TAT pathway also operates in chloroplasts and is responsible for transport of some thylakoid membrane proteins; there is also evidence of a similar pathway operating at plant mitochondrial membranes (Natale et al., 2008). Both the SEC and TAT pathways function in protein secretion in bacteria and archaea as well as in eukaryotes. Interestingly, proteins belonging to the two pathways are also found in the thylakoids of plant plastids as a consequence of their bacterial endosymbiotic origin (Natale et al., 2008).

Other organelles which are a part of the endomembrane system but which do not receive proteins through the SEC system include the nucleus and peroxisomes (Garg and Gould, 2016). Peroxisome matrix proteins possess peroxisomal targeting signals (PTS) which are recognized and translocated into the matrix by a complex of membrane associated proteins called peroxins (Smith and Aitchison, 2013). This type of translocation occurs post-translationally as opposed to proteins targeted to the ER via the signal recognition particle (SRP)-dependent pathway which begin translocation co-translationally (Garg and Gould, 2016). The membrane bound peroxins are trafficked to the peroxisome either via this peroxin complex during direct peroxisome targeting, which is similar to the matrix proteins, or through the ER but without a cleavable targeting peptide (Smith and Aitchison, 2013).

Because translation occurs in the cytoplasm of cells, proteins which function in the nucleus must also be post-translationally targeted and do so by travelling through selective channels called the Nuclear Pore Complexes (NPCs) composed of many different nucleoporin proteins on the membrane of the nucleus (Marfori et al., 2011). These NPCs form bi-directional pores, which allow the passive translocation of smaller macromolecules. Larger nuclear proteins are transported actively across the nuclear envelope through NPCs with the help of nuclear transport factors or carrier proteins (Marfori et al., 2011). These nuclear proteins also possess variable nuclear targeting signals which are recognized by specific proteins to form complexes that will be allowed through the NPC (Marfori et al., 2011).

1.3.2 Mitochondria and Plastids

Plants possess two types of organelles which are not part of the endomembrane system: the mitochondria and plastids (Garg and Gould, 2016). Mitochondria and plastids are surrounded by two membranes due to their endosymbiotic origins and have membrane associated proteins

which direct proteins for import or insertion into these organelles post-translationally (Lee et al., 2014; Pan et al., 2005). Mitochondria matrix-targeted and chloroplast stroma-targeted proteins have cleavable presequences and transit peptides, respectively, located at their N-termini (Lee et al., 2019; Pan et al., 2005). These cleavable peptides each contain amino acid sequences which are recognized by proteins located on the cytosolic face of the outer membrane of the organelle to which each is targeted (Pan et al., 2005). While the cleavable peptide is attached, the protein is termed a pre-protein and is often unfolded in order to pass through the size restrictive protein pores at each membrane. Only after the targeting sequence is cleaved do the proteins fold into their mature and functional conformation (Pan et al., 2005). The recruitment and translocation of pre-proteins into these organelles is most often initiated by membrane bound receptor proteins with GTPase activity (Shiota et al., 2015; Hwang and Robinson, 2009). These receptors belong to a complex of proteins which also include a β-barrel protein that forms the channel through which the pre-protein will pass to reach the intermembrane space (Shiota et al., 2015; Hwang and Robinson, 2009). An associated protein complex is present at the inner membrane of each of these double membrane organelles and mediates the passage of the pre-protein across the inner membrane into the stroma or matrix where the targeting information will be cleaved, allowing the protein to fold into its mature conformation or be sorted to a sub-organellar compartment (Jarvis et al., 2008; Garg and Gould, 2016).

1.4 Plastid Proteins

Extensive horizontal gene transfer from the original endosymbiont to the host cell nucleus has occurred since the original endosymbiotic event, such that plastids and the host cell are now entirely interdependent (Sjuts et al., 2017; Zimorski et al., 2014). Though the vestigial plastid genome still contains some protein coding information, 95% of the approximately 3000

proteins that make up the proteome of chloroplasts are encoded by the nuclear genome and translated on free ribosomes in the cytosol as precursor proteins (Jarvis, 2008; Sjuts et al., 2017). Consequently, the vast majority of the chloroplast proteome must be post-translationally translocated into the chloroplast across its two envelope membranes to reach their functional location inside the organelle (Sjuts et al., 2017). In addition to a developing chloroplast's initial need for the import of proteins to support the wide variety of functions carried out by plastids, developmental and environmental factors require mature chloroplasts to undergo subtle changes in protein content composition, meaning that protein trafficking is continuously needed and must be adaptable (Richardson et al., 2014).

1.4.1 Transit Peptides

Most nuclear-encoded and chloroplast-targeted pre-proteins possess an N-terminal extension called a transit peptide (TP) that targets them to chloroplasts and is cleaved after translocation into the stroma (Jarvis et al., 2008). TPs are recognized by the Translocon at the Outer envelope of the Chloroplast (TOC) complex and although they possess some amino acid composition similarities and identifiable motifs, overall TPs are highly variable (Lee et al., 2019). Some conserved characteristics include a lack of acidic residues, an abundance of hydroxylated residues and proline residues, and an overall unstructured nature in aqueous solutions (Bruce, 2000). This disordered structure of TPs changes to include random coils and one or more alpha helical secondary structures when interacting with import receptors or membrane lipids (Bruce, 2000). The order in which the random coil and α-helical structures appear is reversed between the TPs of ferredoxin and Rubisco activase, suggesting the order of these structural components does not influence their ability to localize to the stroma (Bruce, 2000). In higher plants, the TP often takes a helix-coil-helix structure and at least one of these

12

helices is amphipathic due to a high concentration of hydroxylated residues. The helix of the TP which will be amphipathic, is not consistent and may be the helix most N-terminal, the helix more internal to the protein or both (Bruce 2000).

Additional to those few common characteristics, TPs appear to be made up of multiple sequence motifs whose removal inhibits protein targeting to varying degrees (Lee et al., 2015). In some cases, these motifs can be introduced into unrelated sequences and confer plastid targeting; however, in most cases the context that the surrounding residues in the TP contribute is essential for their function. It does not appear that TPs recognized by the same import receptors contain conserved sequences, causing Lee and colleagues (2015) to hypothesize that each TOC or outer envelope protein (OEP) that interacts with a pre-protein during import recognizes multiple motifs, and that each TP contains at least one specific motif per import protein interacted with during translocation (Lee et al., 2015). One example of a sequence motif in chloroplast TPs is the semi-conserved FGLK domain in ferredoxin and small subunit of Rubisco TPs, which mediates these preproteins' interactions with Toc34 (Holbrook et al., 2016). This motif may also serve as a recognition site for a family of chaperone proteins 14-3-3 which have a well understood role in mitochondrial pre-protein import (Bruce, 2001).

Some of the structural differences observed in TPs could be associated with the protein's final destination as some pre-proteins are targeted to the stroma while others are targeted to the lumen of the thylakoid (Bruce, 2001). A general pattern that differentiates these two groups is that the TPs of stroma-targeted pre-proteins seem to contain three distinct regions, whereas thylakoid lumen-targeted pre-proteins have these three features plus a C-terminal extension that directs the proteins across the thylakoid membrane after the common TP has been cleaved off in the stroma. The three common regions of the TP are an uncharged N-terminal ~10AA peptide,

followed by a heavily hydroxylated region and then an arginine rich and possibly amphipathic C-terminus (Bruce, 2001).

Importantly, the TP is always followed by a stromal processing peptidase (SPP) cleavage site just C-terminal of the targeting domains (Bruce, 2001). The number of different stromal processing peptidases is not known. Two well defined cleavage sites which are recognized by different peptidases for cleavage do not cover all chloroplast pre-proteins suggesting the existence of many other processing site sequences and possibly other proteins to recognize them (Bruce, 2001). SPPs belong to a family of zinc-dependent metallopeptidases which also includes the mitochondrial processing peptidase (MPP) which cleaves the presequences of mitochondria targeted proteins after their translocation into the matrix (Park et al., 2018). In *Arabidopsis,* the vast majority of pre-proteins involved in the light-dependent photosynthesis reactions as well as pre-RbcS are all cleaved by one SPP encoded in the nucleus which must also be translocated across the double membrane and processed before it is functional (Park et al., 2018).

1.4.2 General Protein Import Pathway; TOC and TIC

The targeting and translocation of TP-containing pre-proteins is mediated by the translocon at the outer envelope of chloroplasts (TOC complex) (Lee et al., 2017; Sjuts et al., 2017). The core TOC complex is made up of three membrane-bound proteins: two GTPases, Toc34 and Toc159 and one β-barrel protein channel, Toc75 (Richardson et al., 2014; Sjuts et al., 2017). The Toc proteins were initially discovered in pea plants and are conserved in *Arabidopsis* with homologous proteins in all land plants examined to date (Kessler et al., 1994; Wise and Hoober, 2006). The function and structure of each homolog to one of the known Toc proteins is conserved between all species of land plants examined to date (Wise and Hoober, 2006).

Two GTP binding proteins, Toc34 and Toc159, recognize cytosolic pre-proteins for import via their TP (Kessler et al., 1994; Richardson et al., 2014). A striking difference between these two proteins is the presence of a long disordered acidic domain (A-domain) at the N-terminus of Toc159. This domain, together with the central G-domain, is exposed to the cytosol of the cell and the A-domain appears to be responsible for the recognition specificity for certain TPs favoured by Toc159 or by its minor isoforms (Dutta et al., 2014; Richardson et al., 2009; Schnell 2019). Toc159 shows a higher affinity for the N-terminus of TPs *in vitro* while Toc34 has the highest affinity for phosphorylated C-termini of TPs *in vitro* so it is likely that the two receptors recognize these separate peptides in plant cells (Wiesemann et al., 2019). The recognition of both segments on the TP must then be necessary for translocation across the chloroplast outer membrane, explaining the necessity of two GTPases (Wiesemann et al., 2019).

These two GTPases each belong to a family of homologs. Toc34 and Toc33 make up one family while there are four homologs in the Toc159 family: Toc159, Toc120, Toc132 and Toc90 (Richardson et al., 2014). At least one protein from each family is necessary in a TOC complex and different combinations of these recognize different but overlapping subsets of stroma-bound pre-proteins (Ivanova et al., 2004; Richardson et al., 2009; Wise and Hoober, 2006). For example, a TOC complex with Toc159 has been hypothesized to import proteins involved in photosynthesis because Toc159 is shown to preferentially bind them, whereas family member Toc132 selectively binds proteins involved in biogenesis and constitutively expressed housekeeping proteins (Smith et al., 2004; Bauer et al., 2000; Ivanova et al., 2004; Richardson et al., 2014). Selective recognition by different GTPase receptors is not conferred by the cargo protein's function, rather by the presence of different motifs in the TPs which may be shared between pre-proteins with related functions (Bruce, 2000; Dutta et al., 2014). Closely associated,

15

Toc75 is a deeply embedded β-barrel protein which acts as a protein channel to conduct stroma-targeted proteins recognized by the GTPases through the outer membrane (Sjuts et al., 2017).

After passing through the TOC complex, the translocon at the inner envelope membrane of chloroplasts (TIC complex) mediates the passage of pre-proteins across the inner membrane and into the chloroplast stroma (Jarvis, 2008; Sjuts et al., 2017). The components of the TIC complex are not yet universally agreed upon, but a proteinaceous pore must be present in the complex and it is known that translocation across this pore is driven by an ATPase 'motor protein' (Ganesan and Theg, 2019; Kikuchi et al., 2018).

There are two proposed models involving entirely separate sets of proteins. The original model involves a pore forming complex of Tic110 and Tic40, with close interaction of stromal chaperones Hsp70 and Hsp93 (Huang et al., 2016). The newly proposed 1-MDa complex consists of Tic20, Tic56, Tic100 and Tic214 (Kikuchi et al., 2018). Tic20 is now widely regarded as the channel-forming protein, but there is ongoing debate regarding its interaction with Tic110 and Tic40 (Chen et al., 2002; Kikuchi et al., 2018; Li et al., 2020). The 1-MDa complex is proposed to contain 3 copies of Tic20, a protein which has 4 hydrophobic α-helices, potentially creating a 12 transmembrane-helix pore if it contributes all 4 helices to pore formation (Campbell et al., 2014a; Ganesan et al., 2019). This would be a larger pore than the 8-helix pore which would form from the two Tic110 proteins believed to be present in the Tic110 complex, causing Ganesan and colleagues to hypothesize that a Tic20 containing, 1-MDa TIC complex is the complex which performs the second half of stromal pre-protein translocation (2019). Data which supports the role of Tic20 as the channel-forming protein of the TIC includes chemical cross-linking experiments which show that Tic20 interacts with pre-proteins in transit, the presence of an intrinsically disordered region at its N-terminus which could perform peptide

recognition functions like the N-terminal A-domain of Toc159, and the self-association of Tic20 in lipid membranes (Kouranov et al., 1998; Chen et al., 2002; Campbell et al., 2014a). The function of the other proteins proposed to associate with Tic20 in this 1 MDa complex are not yet known (Ganesan et.al., 2019).

Once the pre-protein is translocated across the inner membrane by the TIC complex and emerges in the stroma, the TP is removed from the pre-protein by stromal processing peptidase so that the protein may fold into its native form and function without this extra peptide (Bölter and Soll, 2016).

1.4.3 Protein Targeting to the *Chloroplast Outer Membrane*

The proteins of the TOC complex belong to a larger group of chloroplast outer envelope proteins (OEPs) (Lee et al., 2017). Currently, the number of OEPs that are known or predicted to reside in the outer envelope is 117 (Inoue, 2015). All 117 OEPs are nuclear encoded and rely on post-translational targeting pathways (Inoue, 2015; Kim et al., 2019). There are at least four well-studied pathways utilized by OEPs for targeting to the chloroplast outer membrane. OEPs can be categorized based on which of the four pathways they use. The four groups include the signal-anchored (SA) proteins, tail-anchored (TA) proteins, β-barrel proteins, and one (or two) members of the Toc75/OEP80 family that use an N-terminal TP for targeting (Gross et al., 2020; Hofmann and Theg, 2005).

Signal-anchored (SA) proteins have a hydrophobic, α-helical transmembrane domain (TMD) near their N-terminus and a positively charged C-terminal region (Hofmann and Theg, 2005; Kim et al., 2019; Lee et al., 2017). OEP64, a protein which receives a subset of pre-proteins from chaperone protein HSP90, is an example of a signal-anchored protein (Qbadou et

al., 2006). Ankyrin repeat protein 2A (AKR2A) is a cytosolic factor involved in the targeting of a subset of SA proteins to the chloroplast outer envelope (Dhanoa et al., 2010; Kim et al., 2014). AKR2A binds a site on the ribosome near the exit tunnel while a SA protein is being translated. The N-terminus of AKR2A binds the TMD and flanking C-terminal positively charged region co-translationally as they emerge from the ribosome exit tunnel to transport them to the outer envelope (Kim et al., 2019). At the outer envelope the C-terminal ankyrin repeat domain (ARD) of AKR2A recognizes two lipids, monogalactosyldiacylglycerol (MGDG) and phosphatidylglycerol (PG), as a receptor which it binds to complete the delivery of its cargo OEP (Kim et al., 2014; Kim et al., 2019). That MGDG serves as part of the receptor for AKR2A is interesting because the lipid is an evolutionary product of the chloroplast's origin as a cyanobacteria and is unique to chloroplasts while AKR2A is evolutionarily derived from ankyrin repeat domain of the host eukaryotic cell (Kim et al., 2014).

Tail-anchored (TA) proteins have a single hydrophobic TMD near the C-terminus, with the extreme end of this terminus, the C-terminal sequence or CTS, providing most of the information for where this TMD is to be inserted (Teresinski et al., 2019). Because the targeting information is in the CTS and the adjacent TMD, the proteins must be targeted to the appropriate organelle only post-translationally, as the C-terminus is the last to be synthesized and is only exposed to the cytosol after the whole protein is released from the ribosome (Dhanoa et al., 2010). The region flanking the other side of the TMD is also necessary to target some plastid TA proteins while just the CTS and TMD are enough in others, including OEP9 (Dhanoa et al., 2010).). In the case of Toc33 nearly all of the protein including the G-domain is necessary for its targeting to the chloroplast outer membrane (COM) (Dhanoa et al., 2010). In a subset of tail-anchored OEPs, including OEP9 and Toc34, though not Toc33 due to its much shorter CTS, the

targeting information at the C-terminus is rich in lysine or arginine and serine residues (Dhanoa et al., 2010; Qbadou et al., 2003; Teresinski et al., 2019). Positively charged R/K residues and S residues in these proteins form a 7-9 AA length motif which may be found anywhere within the C-terminus, while elsewhere in the signalling information negatively charged residues are also enriched. Both positively and negatively charged residues appear to be necessary for targeting of TA proteins, even if they do not possess the RK/ST-rich motif (Teresinski et al., 2019). This conserved RK/ST enriched motif may allow them to utilize a single protein-mediated targeting pathway while other TA OEPs appear to target to the membrane with the help of entirely separate proteins (Dhanoa et al., 2010; Teresinski et al., 2019). A separate subset of TA proteins, including Toc33/34 and OEP9 have been shown to interact with AKR2A, though the specifics of the role AKR2A plays in TA protein targeting is not yet fully elucidated (Dhanoa et al., 2010; Kim et al., 2019). TA proteins in other organisms and organelles are inserted with the help of the Guided Entry of TA proteins (GET) system which is regulated by an ATPase named GET3 (Lin et al., 2019). Recently a homolog of GET3, ArsA1, was found in plants which is proposed to play a role in plastidic TA protein insertion in algae (Lin et al., 2019).

β-barrel proteins are characterized by a series of β-sheet structures spanning the membrane to create a channel (Hofmann and Theg, 2005; Lee et al., 2017). In eukaryotes, β-barrel proteins are found only in the membranes of mitochondria and plastids, which is consistent with their evolutionary bacterial history (Klinger et al., 2019). Specifically, β-barrel proteins are also present in the outer membrane of all bacteria but not in the cell membranes of archaea or eukaryotes. Both mitochondria and plastids developed from engulfed bacteria and share membrane characteristics with bacteria as a result (Klinger et al., 2019). In the case of β-barrel proteins which reside in the outer envelope of mitochondria, targeting of the pre-protein to

the mitochondrial inter-membrane space is directed by a C-terminal β-hairpin signal where it can be recognized and assembled into the membrane by the Sorting and Assembly Machinery (SAM) complex (Jores et al., 2016). The details of the β-barrel targeting pathway to the plastid outer membrane remain unclear. It is expected that a similar complex to the SAM found in the mitochondrial outer membrane, which has a β-barrel as a central component, is responsible for the insertion of this type of protein in plastid membranes (Richardson et al., 2014). The targeting differentiation between mitochondria and chloroplast localization appears to be related to the hydrophobicity of the last β hairpin turn, where a higher hydrophobicity favours plastid targeting and if the hydrophobicity drops too low the protein stops targeting effectively and remains in the cytosol (Klinger et al., 2019). Klinger and colleagues (2019) also studied the second last β-strand of a subset of plant β-barrel proteins and found a pattern wherein the most hydrophilic amino acids in this strand were towards the N-terminus in plastid β-barrel proteins and closer to the turn (more C-terminal) in mitochondrial β-barrel proteins. These two observations suggest that β-barrel plastid proteins may be recognized for insertion via a β-hairpin signal-like peptide composed of a hydrophilic N-terminus of the second last β-strand and a very hydrophobic final hairpin turn (Jores et al., 2016; Klinger et al., 2019).

Unlike most other β-barrel proteins of the COM, Toc75 is a β-barrel protein targeted to the chloroplast by an N-terminal TP; however, the β-barrel TMDs and a polyglycine region allow it to be incorporated into the membrane instead of translocating across the CIM all the way to the stroma, as is the case for other known TP-containing proteins (Hofmann and Theg, 2005; Richardson et al., 2014; Sjuts et al., 2017). Toc75 may not be completely unique in its targeting as one of the other members of its protein family, OEP80, was recently discovered to also posses a cleavable N-terminal domain which is exposed to the intermembrane space after insertion into

the outer membrane (Gross et al., 2020). It is proposed that this domain could function as the signal for target specificity (Gross et al., 2020).

1.4.4 A New Chloroplast OEP Targeting Pathway

Toc159 is predicted to contain an α-helix near its C-terminus which could function as a TMD making it similar to tail-anchored proteins (Teresinski et al., 2019). This helix has been shown to interact with TA insertion machinery but its function as a TMD is not universally agreed upon (Lin et al., 2019; Teresinski et al., 2019). However, the C-terminal sequence following the TMD is relatively shorter and less positively charged than in other TA proteins characterized (Teresinski et al., 2019). Additionally, Toc159 does not contain an RK/ST enriched motif or appear to interact with AKR2A as a chaperone, and instead utilizes a so-far uncharacterized strategy for targeting to the outer envelope of chloroplasts (Teresinski et al., 2019). Toc159 is targeted to the outer membrane of chloroplasts by direct binding of its own central GTPase domain to the GTPase domain of Toc34 in a process requiring the hydrolysis of GTP (Smith et al., 2002a). Recently, the bioinformatics tool Chloro-P was used to predict a novel transit peptide-like chloroplast targeting signal in the C-terminus of the *Bienertia sinuspercici* isoform of Toc159 (BsToc159) as well as a sequence that was recognizable as a peptidase processing site (Lung and Chuong, 2012). Chloro-P was capable of generating this result because the BsToc159 amino acid sequence was input in reverse so that the program would look for a TP at the C-terminus as opposed to the N-terminal TPs it was developed to recognize (Lung and Chuong, 2012). Protoplast transient expression assays were used to verify that the predicted 56AA putative transit peptide could direct EGFP to the outer membrane of chloroplasts (Lung and Chuong, 2012). Transient expression analyses were performed with truncation constructs including various regions of the most C-terminal 100 amino acids (100AA)

of BsToc159 which resulted in the conclusion that the predicted most C-terminal 56AA TP-like signal contained the necessary and sufficient targeting information; however, the inclusion of the remainder of the most C-terminal 100AA increased targeting efficiency, suggesting the presence of an element allowing stable interaction with the membrane is present in that region (Lung et al., 2014). Equivalent and very conserved ~100AA regions at the C-terminus of AtToc159, its minor isoforms and homologs in other species were identified using alignment software (Lung et al., 2014). Species specificity was inferred due to divergence localized to the extreme end of the C-terminus (Lung et al. 2014). When transient expression was performed to verify this targeting region prediction, targeting was less efficient when the TP from *Bienertia sinuspersici* Toc159 was expressed in *A. thaliana* and vice versa. Conversely, changing the C-terminal tail to match the host organism resulted in targeting efficiency that matched the expected rate, confirming that species-specific targeting information is included at the extreme end of the C-terminus (Lung et al., 2014).

1.4.5 Prediction Tools

Subsequent bioinformatics analysis of 117 OEPs identified by Inoue (2015) using ChloroP to identify C-terminal TP-like sequences in the reverse orientation of proteins identified seven additional OEPs with a predicted reverse transit peptide-like signal (rTP) at their C-termini (Grimberg, 2016). These seven OEPs were selected based on a ChloroP score of above 0.5 (on a scale of 0-1) when their amino acid sequences were input into the program in reverse (Grimberg, 2016). It is important to note that this means that the C-terminus of these OEPs have similar charge, hydrophobicity, amino acid composition and primary structure predictions to known N-terminal TPs; however, the use of Chloro-P with reversed inputs means there are considerations to make before accepting the presence of a TP. For example, the exposure of the carboxylic acid

group at the terminus of the rTP-like signal as opposed to the regularly exposed amino group at the N-terminus of true TPs may influence their ability to function as a true transit peptide. These bioinformatic results have been taken only as leads towards identifying proteins which may utilize this new C-terminal TP-like signal localization path and each protein must be studied using wet lab techniques as a follow-up.

1.5 A C-terminal TP Candidate; At4G02482

One of the protein candidates predicted to contain C-terminal reverse TP-like targeting signal was a protein encoded on the 4th chromosome of *Arabidopsis thaliana* with accession number At4G02482; its ChloroP score when the AA sequence was analyzed in reverse was 0.548 (Grimberg, 2016). The structure and function of this protein has not been previously studied and does not appear to have any known homologues in other plant species. At4G02482 is a small protein with an apparent mass of only 14.74 kDa so the TP length of 44AA predicted by ChloroP would cover a large portion of the protein's total length (Grimberg, 2016). It was selected for further studies that were performed by past students in the lab due to its high ChloroP score and annotated putative function as a GTPase, an interesting potential pattern considering Toc159 is also a GTPase.

At4G02482 was predicted to function as a GTPase in databases such as The Arabidopsis Information Resource (TAIR, https://www.arabidopsis.org) and Araport (https://www.araport.org) due to its sequence similarity to Toc159, a GTPase which performs a role in the TOC complex. When aligned with Toc159, however, all of the sequence identity between the two proteins occurs in the membrane associated domain (M-domain) of Toc159; that is, there is no similarity between the G-domain of Toc159 (that contains the GTPase activity) and any part of At4G02482 (analysis shown in section 3.8). The highest homology

occurs in a 103 AA region which corresponds to residues 1383-1485 of TOC159 and residues 32-134 of OEP15-1 in which the pair have 56.3% identity and 81.6% similarity (Huang and Miller, 1991). This membrane associated domain of Toc159 was named due to its resistance to proteolysis in intact chloroplasts suggesting that this region of the protein is within the protective hydrophobic membrane of the chloroplast (Chen et al., 2000). It is also possible that much of this region resides in the intermembrane space instead of being a large transmembrane domain (Chen et al., 2000; Lung et al., 2014). Pre-protein substrates at the intermediate stages of import have been found to interact with some peptides in the M-domain, supporting the idea that some of this region may be exposed to the intermembrane space or that it may contribute to the pore formed by Toc75 and exposing it to translocating pre-proteins in this way (Ganesan and Theg, 2019). The alignment of At4G02482 to the M-domain instead of the G-domain of Toc159 means it is inaccurate to refer to this protein as a putative GTPase, but does not detract from the intrigue of studying this small protein due to the limited understanding of the Toc159 M-domain it resembles.

The PPDB does not include At4G02482 in its proteomics list for 'plastid (all)' or 'outer envelope' (Sun et al, 2008). There are many 2D-PAGE experiments which may have resolved At4G02482, but there is no published identification by mass spectrometry of this protein from proteomic studies to date (Sun et al., 2008). Instead, the classification of At4G02482 as an OEP by TAIR appears to be based solely on its resemblance to Toc159. There are also a few ESTs which align to its general gene location on chromosome 4; however, these have been found in RNA extractions from mixed tissues, such as 'above ground tissues or 'ovule', for example. This means that while at least one isoform of At4G02482 appears to be transcribed in *A. thaliana*, more work is necessary to experimentally confirm its localization.

1.5.1 History of Annotations for At4G02482

At the time in which research was last performed on At4G02482, prior to the start of the current study, it was annotated to contain two exons and one intron, corresponding to a predicted 134 AA protein. For the purpose of the current project, I propose naming this annotation of At4G02482 as "OEP15-1" for its predicted presence in the outer envelope of chloroplasts and its estimated molecular weight of 14.74 kDa, which rounds up to 15.

The original annotation of OEP15-1 found in the *A. thaliana* genome assemblies TAIR9 and TAIR10 was discarded recently from the 2016 assembly of Araport11 in favour of a single exon encoding a 66 AA protein (Lamesch et al., 2012; Cheng et al., 2017). I will refer to the two potential isoforms as OEP15-1 (2 exon, 14.74-kD isoform) and OEP7.3 (single exon, 7.26-kD isoform), respectively, named for their molecular weights (see Figure 1) as is common practice when naming OEPs. Expressed sequence tags (ESTs) of OEP15-1 in online databases are limited and either span only the ORF of OEP7.3 or include the intron of OEP15-1, which contains a STOP codon, making it unlikely that these encode for the predicted 134-AA protein (see Figure 2.1). An approximately 500 bp product can be amplified from mRNA isolated from Arabidopsis plants using OEP15-1-specific primers, which fits the length this gene takes up on the chromosome, intron included.

1.5.2 OEP15-1 and OEP7.3 as Research Candidates

OEP15-1 and OEP7.3 are an intriguing pair of candidates for research as they have the potential to help characterize the new C-terminal TP targeting pathway that has already been shown to be used by Toc159. Based on previous research by lab members and into the localization of OEP15-1, I hypothesized that OEP15-1 and OEP7.3 are localized to the

chloroplast outer membrane (COM) and targeted there using the common C-terminal transit peptide-like signal they share.

1.6 Objectives

The current study aims to verify the presence of both OEP15-1 and OEP7.3 in *Arabidopsis thaliana*, compile any evidence of their function, and determine which regions are necessary for targeting to the outer envelope of chloroplasts. mRNA and ESTs were studied as an indication of two isoforms' presence, while homology to TOC159 was studied to give an evolutionary context to the gene's existence. Bioinformatic tools for tertiary structure prediction provide a model which serves as a first clue towards possible function as well as important context when analyzing construct localization. The regions necessary for targeting were studied using truncated proteins fused to GFP for visualization under fluorescence microscopy and Western blots.
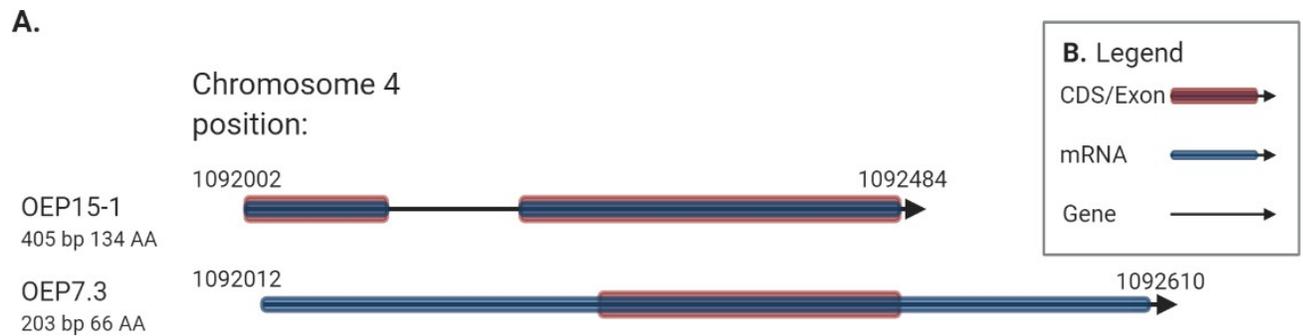
Figure 1.2: Annotations of At4G02482. A: Two annotations of At4G02482 in the context of chromosome 4 of *A. thaliana*. OEP15-1 begins upstream of OEP7.3. OEP15-1 has two exons and one predicted intron while OEP7.3 has one exon and the remainder of the mature mRNA is a UTR. B: Legend for understanding OEP15-1 and OEP7.3 annotations at the bottom of frame A. A black arrow indicates the location of the gene. Blue highlighting indicates presence in the mature mRNA, therefore internal un-highlighted black lines indicate introns. Red highlighting indicates coding sequence. Image created in Biorender.com

Chapter 2. Materials and Methods

2.1 Chemicals and Supplies

All chemicals were analytical grade and purchased from Sigma-Aldrich (Oakville, ON, Canada), BioShop Canada Inc. (Burlington, ON, Canada) or Honeywell Fluka through Fisher Scientific (Ottawa, ON, Canada), unless otherwise specified. Two antibodies were used. One primary antibody against EGFP was made by the Southern Alberta Cancer Research Institute in Calgary and one horseradish peroxidase-conjugated anti-rabbit antibody raised in goat used as the secondary antibody (Sigma-Aldrich, Oakville, ON, Canada, cat. no. A6154). DNA sequencing service was provided by the Sanger Sequencing Facility at The Centre for Applied Genomics (The Hospital for Sick Children, Toronto, ON, Canada). Primers were ordered from Eurofins Genomics (Eurofins Genomics LLC, 12701 Plantside Drive, Louisville, KY 40299).

2.2 Growth of *Arabidopsis thaliana* Plants

Wild Type *A. thaliana* Col-0 seeds were suspended in 0.1% (w/v) agar and incubated at $4°C$ for 3 days before planting at a concentration of ~2 seeds per cm^2 on flats of Sungro professional potting mix Sunshine #1 for *Arabidopsis* ordered from Eddi's wholesale (Cambridge, ON.). The planted flats were kept in a growth chamber with a day/night photoperiod of 16/8h at 22°C with a photon flux density of 130-150 μmol m-2 s-1. The flats were covered with a translucent plastic lid until cotyledons were visible at which point the lid was removed. Once the lid was removed all plants were watered twice a week from the bottom. Once the plants showed signs of bolt development, they were fertilized with a 20-20-20 fertilizer once a week throughout their bolting stage until seed collection.

2.3 RNA Extraction

Rosette leaves of 4- to 6-week-old WT *A. thaliana* plants were placed into a 1.5 mL microfuge tube, taking care not to crowd the tube. The closed tube was frozen in a bath of liquid nitrogen and then manually crushed to fine powder with a small plastic pestle. The powdered tissue was then frozen again with liquid nitrogen and stored at -80°C until needed. Multiple microfuge tubes of powdered tissue were prepared at once up to the desired amount of tissue.

RNA was extracted using the acid-guanidine-phenol-chloroform (AGPC) RNA extraction method (Chomczynski and Sacchi, 1987). Briefly, 100 mg of powdered WT *A. thaliana* rosette leaf was added to 1 mL of acid-guanidine-phenol extraction solution [38% (v/v) buffer-saturated phenol, 0.8 M guanidine thiocyanate, 0.4 M ammonium thiocyanate, 0.1 M sodium acetate (pH 5), and 5% (v/v) glycerol] in a 1.5 mL tube and vortexed until homogenous. The tube was left at RT for at least 5 min before adding 200 µL chloroform, vortexing for 15 sec and again leaving at RT for 10 min. The tube was then centrifuged at 14,000 rpm for 15 min at 4°C and 600 µL of the supernatant was transferred to a new 1.5 mL tube containing 600 µL of 100% isopropanol. This tube was mixed by inversion and left at RT for 5 min before centrifuging at 14,000 rpm for 10 min at 4°C. The supernatant was discarded and the RNA pellet was washed with 800 µL 70-80% ethanol by inverting the tube until the pellet was dislodged from the wall of the tube and recentrifuged at RT for 5 min at 7,500 rpm. The supernatant was again discarded and the tube was inverted on a paper towel to allow the pellet to air dry at RT for 30 min. 88 µL of diethyl pyrocarbonate (DEPC)-treated water was added to the air-dried pellet and the tube was incubated in a 65°C water bath until the pellet had dissolved (roughly 10 min with intermittent mixing by inversion or gentle vortexing).

To remove contaminating DNA from the RNA sample an Ambion DNA-free Kit was used for DNase treatment. 10 µL of 10X DNase buffer was added to the tube containing the re-suspended RNA and mixed by flicking the bottom of the tube before adding 2 µL recombinant DNase I enzyme. This solution was mixed gently by flicking the bottom of the tube and then incubated at 37ºC for 30 min. To heat inactivate rDNase I, 1 µL of 0.5% EDTA was mixed into the tube before incubating at 75ºC for 10 min. The resulting RNA was stored in this solution at -80ºC until used for RT-PCR.

2.4 Detection of Endogenous OEP15-1 and OEP7.3 Transcripts

When RT-PCR was first performed using primers designed to amplify the ORF of *OEP15-1* based on the annotation shown as TAIR 10 protein coding gene in Figure 2.1 by Yeung and Chuong (2016), the amplified fragment was cloned into pBluescript (construct pBS-OEP15-1) and used as the starting template for EGFP fusion constructs. Recently, pBS-OEP15-1 was sent for sequencing as part of a troubleshooting process, which revealed that the intron of *OEP15-1* was present in the clone. There is an in-frame stop codon located in this intron. Therefore, any clones designed to include all or part of both flanking exons (i.e. included the intron) that used pBS-OEP15-1 as the template had to be discarded, as the intended protein product would then be truncated because of the internal stop codon in the intron. One clone, namely OEP15-1NT:GFP, from Yeung's work was retained because the cDNA was amplified from only the first exon of pBS-OEP15-1 and thus did not include the unwanted STOP codon. This means that translation of the OEP15-1NT:GFP fusion should proceed uninterrupted and produce the expected GFP fusion protein in cell expression systems.

ESTs were recorded from BLAST and Araport11 search results as part of analysis into the existence of an mRNA encoding OEP15-1 and can be found at the bottom of Figure 2.1.
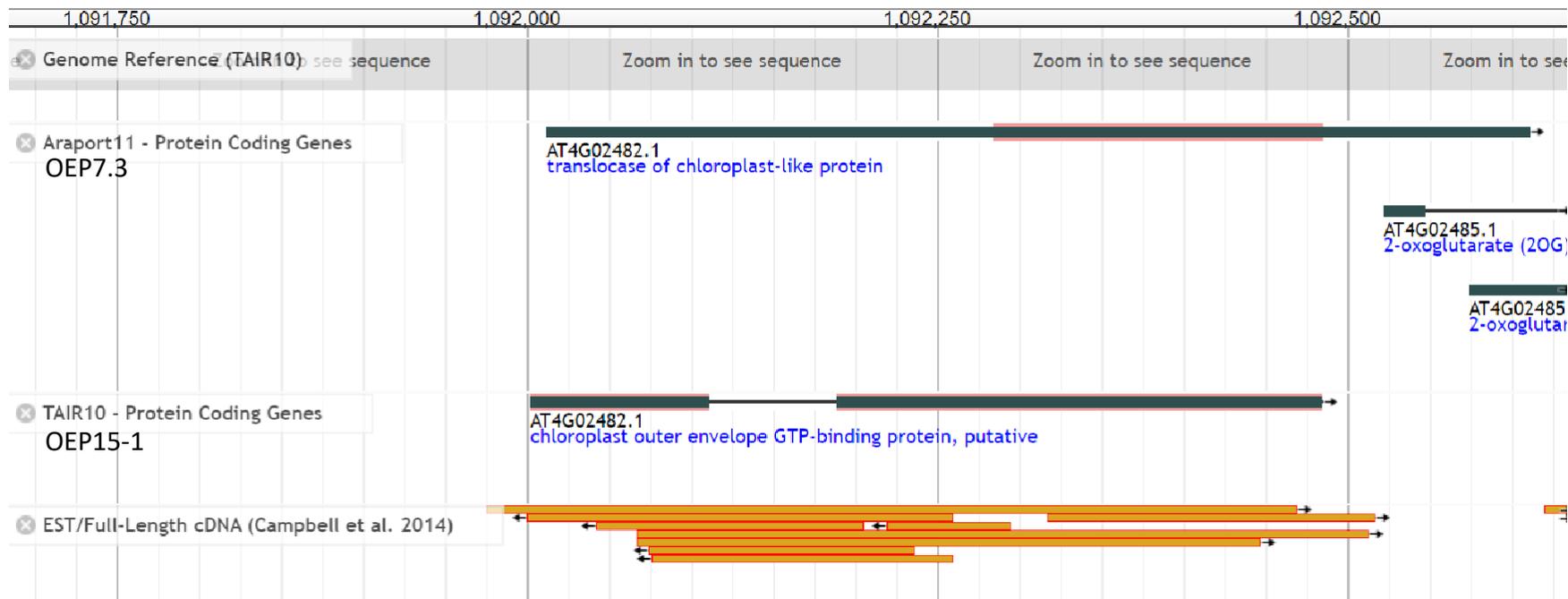
Figure 2.1

Please see legend on next page

Figure 2.1: Gene Browser Showing Two At4G02482 Annotations and ESTs. Readout from Jbrowse.arabidopsis.org search for At4G02482. The top light-grey bar denotes the numbered position on chromosome four of *A. thalliana.* The two annotations have been labelled OEP7.3 and OEP15-1 underneath the source tag. Black bars indicate that this region is predicted to be a part of the mature mRNA, red highlighting indicates coding sequence and the underlying thin black line indicates the gene location before transcription. The orange bars at the bottom indicate aligned ESTs from various experimental sources. (Cheng et al., 2017; Lamesch et al., 2012; Campbell et al., 2014b).

Total RNA was reverse transcribed using iScript supermix for RT-qPCR following the manufacturer's instructions to generate cDNA for later PCRs (Bio-Rad Laboratories Inc., Mississauga, ON, Canada).

Endogenous OEP15-1 and OEP7.3 transcripts were detected using PCR. First, the long isoform transcript was amplified. A product was detected in the negative control lane so a primer pair was designed to test for genomic DNA contamination. Lastly, 3'UTRs were amplified. All 3 PCRs were performed in 25 µL reaction volumes including 1 µL 20 ng/µL starting template, 1 µL KOD taq-polymerase, 2.5 µL 10xKOD Buffer, 3 µL 50 mM $MgCl_2$, 2 µL 5 mM dNTPs, 1 µL 10 ng/µL each of forward and reverse primers and 13.5 µL UV treated de-ionized water. Thermocycler conditions were 5 sec at 98 °C to denature, 10 sec at a variable temperature to anneal and 15 sec at 72 °C to extend. The variable temperature was decided based on the primer pair used. Primer names and corresponding temperatures for these three PCRs can be found in appendix Table A3 and primer positions in the context of genomic DNA can be viewed in appendix Figure A4.

2.5 Construct Cloning

Three separate methods were used to obtain the constructs used in this thesis.

Constructs OEP15-1:GFP, GFP:OEP15-1, OEP15-1 CT:GFP, GFP:OEP15-1 CT, OEP15-1 ΔCT:GFP, OEP7.3:GFP, GFP:OEP7.3 and Hfh:DsRed were ordered from BioBasic's Gene in Vector service (BioBasic, Markham, ON, Canada). Construct design is detailed in results section 3.4 and Figure 3.4 provides an illustrated summary of all GFP constructs.

The truncation constructs of OEP7.3 named OEP7.3β1-4:GFP, OEP7.3β3-4:GFP and OEP7.3 β3-4α:GFP were cloned by Dr. Simon Chuong. Because all of OEP7.3 is contained

within the second exon of OEP15-1, the ordered clone OEP15-1:GFP was used as the starting material for insert amplification in all three cases. Four primers were designed as follows: OEP7.3F1-ATGGGAGTGTCTCTA which amplifies a product beginning at the first amino acid of OEP7.3, OEP7.3F2-ATGCGACATCAAGTCTCT which begins amplification at the 19th amino acid of OEP7.3, OEP7.3R1-TGAGCTGTTTGTTCGGAC which will define the end of the product at the 51st amino acid and OEP7.3R2-CATGGCAAGGAGAAGCAG which will end the product at the predicted coding sequence (CDS) end, the 66th amino acid. Primer position in the context of genomic OEP7.3 can be viewed in Figure A5. For PCR amplification of the insert for construct OEP7.3β1-4:GFP, primers OEP7.3F1 and OEP7.3R1 were used. For OEP7.3β3-4:GFP, primers OEP7.3F2 and OEP7.3R1 were used and for OEP7.3β3-4α:GFP the primers used were OEP7.3F2 and OEP7.3R2. PCR was performed using 50 µL reactions composed of 2 µL of 20 ng/µL template plasmid, 5 µL each of the appropriate forward and reverse primers at a working concentration of 5 µM, 0.5 µL of 2 unit/µL Phusion High-Fidelity DNA polymerase (NEB Cat #M0530S), 10 µL of 5x Phusion HF Buffer (NEB), 1.5 µL of 100% DMSO, 1 µL of 50 mM $MgCl_2$, 2 µl of 5 mM dNTPs and 23 µL UV sterilized and autoclaved milli-Q filtered H_2O. The cycle settings used were a 3 min initial denaturation at 98 °C followed by 35 cycles of amplification [5 sec at 98 °C to denature, 10 sec at 53 °C to anneal, 15 sec at 72 °C to extend] and ending with a final extension at 72 °C for 10min. The finished reactions were then stored at 4 °C or -20 °C until they could be run on a 1.5 % agarose gel and gel purified. Gel purification was necessary due to multiple band sizes being present on an initial screening gel. Gel purification of expected band sizes (F1R1 = 153 bp, F2R1= 102 bp and F2R2= 147 bp) was performed using an EZ-10 Spin Column DNA Gel Extraction kit following manufacturer's instructions (BioBasic). Blunt end ligation into pSAT6-N1 vector digested with SmaI was performed using the purified

PCR product. Due to low concentrations of purified product (10-12 ng/µL each), the 20 µL

ligation reactions contained the maximum volume of insert possible at 15 µL per reaction. The

rest of the volume was made up of 1 µL T4 Ligase, 2 µL 10x Ligase Buffer (NEB Catalogue

numbers M0202S and B0202A), 1.5 µL of 125 ng/uL pSAT6-N1-SmaI and 0.5 µL H$_2$O. The

reactions were incubated at 4 °C for 18hrs before the whole reaction was mixed into ice cold

DH5α competent E-coli cells which were prepared following the protocol in Sambrook et al.,

(1989). The cells were then heat shocked and plated on solid LB with 100 µg/mL Ampicillin and

incubated at 37 °C for 18 hrs. Colonies which were observed the next day were screened by PCR

using the same primers and reaction conditions as described for the previous PCR, but

substituting the Phusion polymerase and Buffer for a homemade KOD DNA polymerase and 10x

KOD Buffer [100 mM Tris-HCl pH 8, 100 mM KCl, 20mM MgSO$_4$, 0.1% Triton-100, 1 mg/mL

BSA] .

One construct, OEP15-1NT:GFP was cloned in the lab by previous honours thesis

undergraduate Kelly Yeung (Yeung and Chuong, 2016). For this construct, primers designed to

amplify the full-length cDNA were first used to create a pBluescript clone. cDNA was

synthesized from *A. thaliana* rosette leaf total RNA using the iScript cDNA synthesis kit

according to the manufacturer's instruction (Bio-Rad Laboratories, cat # 1708890). The purified

PCR product was then cloned into pBluescript and this was used as the starting template for

preparing the N-terminus insert of this clone. Primers ATGTPASEF3ASAL1 -

CGCGTCGACCATGGGGACAATGGTTCT and ATGTPASER5BAMH1 -

CGCGGATCCCCCGTGTGTTGCCAAAGAATGT were designed to amplify a product

containing only the first 40 codons of OEP15-1 with the addition of a SalI restriction site

upstream of the 1st codon and a BamHI restriction site downstream of the 40th codon. The PCR

product and pSAT6-N1 empty vector were both then digested with BamHI and SalI, gel purified, and then ligated together for transformation into XL10-gold *E. coli* by heat shock method. . Plasmid DNA was purified from colonies grown from glycerol stock and re-transformed into DH5α *E. coli* for consistency before use. Plasmids were isolated from these new cells and sent for sequencing before use in the data presented in this thesis to verify insert presence, length, sequence, and orientation.

All constructs were transformed into competent DH5α *E. coli* cells by the heat shock method and LB broth of overnight cultures were stored at -80 ºC in 50% v/v 80% w/v glycerol until use. Plasmid DNA isolation was performed using 150 mL broth cultures and the Geneaid Presto Midi Plasmid Kit following the manufacturer protocol (Geneaid cat # Pif025).

2.6 Biolistic Bombardment of Onion Epidermal Cells

Tungsten microcarriers were coated in plasmid DNA for each construct using a modified protocol adapted from Sanford et al., (1993) as follows. Briefly, 30 mg of tungsten particles were prepared by first adding 1 mL of 70% (v/v) ethanol, vortexing for 1 min, and allowing to soak for 15 min. The particles were collected by centrifugation at 14,000 rpm for 15 sec and the ethanol was removed. The particles were then washed three times by adding 1 mL sterile DI water, vortexed for 1 min, and pelleted by centrifugation as above before removing the wash. The washed tungsten was then pelleted by centrifugation as above, re-suspended in 500 µL of 50% (w/v) glycerol to bring the particle concentration to 60 mg/mL and stored at -20ºC.

For each bombardment, the prepared tungsten particles were coated with plasmid DNA by adding ~1-2 µg of plasmid DNA, 10 µL of 2.5 M $CaCl_2$ and 5 µL of 0.1 M spermidine to 0.5 mg of tungsten and the solution was vortexed for 2 min. The solution was allowed to settle for

another 2 min before pelleting by centrifugation at 14,000 rpm for 15 sec. The pellet was washed

once with 70% ethanol and once with 100% ethanol, re-suspended in 10 µL of 100% ethanol and

the entire suspension was loaded onto the center of a microcarrier disc (Bio-Rad). For each

bombardment, three sections of 1x2 cm of onion scales prepared from the third layer or deeper

layer of white onion bulb (*Allium cepa*) obtained from a local grocery store were placed in a petri

dish, adaxial epidermis up. The DNA-coated tungsten was bombarded onto the adaxial epidermis

of the onion bulb sections at 1350 psi and a distance of 10 cm using the Biolistic PDS-1000/He

particle-delivery system (Bio-Rad, Mississauga, ON, Canada) according to the manufacturer's

instructions. The bombarded sections were incubated on water-saturated filter paper in the petri

dish at room temperature in the dark for 20 hours. Wet mount slides of the adaxial epidermis

were prepared for imaging of cells expressing the fusion constructs using a Zeiss AxioImager D1

epifluorescence microscope equipped with an AxioCam MRm camera and AxioVision 4.7

software (Carl Zeiss Canada Inc., Toronto, ON, Canada). To detect GFP, MitoTracker Red, and

DsRed signals, fluorescence filters for EGFP (excitation 470 nm, emission 525 nm) and CY3

(excitation 550 nm, emission 570 nm) were used, respectively. Representative images of

expressing cells from at least three independent experiments were captured.

2.7 Protoplast Transient Expression

Protoplasts were used for the bulk of localization experiments in this work and the

protocols used will be separated into the major stages of their processing.

2.7.1 Protoplast Isolation from <u>Arabidopsis thaliana</u>

Protoplasts were isolated from 3–5-week-old WT *A. thaliana* rosette leaves using a

modified version of the TAPE method originally described by Wu and colleagues (2009).

Briefly, this was performed by peeling the abaxial epidermis off of ~30 of the rosette leaves and incubating the exposed mesophyll cells in enzyme solution for 1-2 h at RT in light with gentle shaking. Enzyme solution was prepared by incubating 10 mL of CS-mannitol buffer [0.4 M mannitol, 20 mM MES-KOK (pH 5.7) and 20 mM KCl] at 70 °C for 10 min before addition of 1% w/v cellulase and 0.25% macerozyme then incubated at 55 °C for 10 min to dissolve. The solution was then cooled to RT before addition of 0.1% (w/v) BSA and 10 mM $CaCl_2$. Protoplasts released into the solution during incubation were then transferred to 5 mL tubes and pelleted by centrifugation for 3 min at 100 X g. The protoplasts were resuspended in W5 buffer [2 mM MES-KOH (pH 5.7), 154 mM NaCl, 125 mM $CaCl_2$, 5 mM KCl] and pelleted again by centrifugation as above as a wash step. This wash was done only once to avoid excess damage to cell membrane integrity. After washing, the pelleted protoplasts were re-suspended in CS-sucrose buffer [0.4 M sucrose, 20 mM MES-KOH (pH 5.7), 20 mM KCl] and centrifuged for 3 min at 100 X g to separate ruptured cells from healthy, intact cells, which float in sucrose buffer (Lung et al., 2014). The pellet and internatant were discarded and the healthy cells were re-suspended in 1 mL of W5 so that the concentration of cells could be calculated on a haemocytometer. The cells were allowed to settle, the supernatant was discarded, and the protoplasts were re-suspended in enough Mg Mannitol buffer [0.4 M mannitol, 4 mM MES-KOH (pH 5.7), 15 mM $MgCl_2$] to make the concentration of cells 20,000 cells per 100 µL.

2.7.2 PEG-Mediated Transfection of A. thaliana Protoplasts

7.5 µg of plasmid DNA was used to transfect 20,000 protoplasts in the presence of 110 µL PEG solution [40% (w/v) PEG, 0.4 M mannitol, and 100 mM $CaCl_2$] in 15 mL tubes. The mixture was incubated at room temperature in the dark for 15 min after which the PEG was inactivated by the addition of 440 µL W5 buffer. The tubes were inverted gently 3 times and

centrifuged for 3 min at 100Xg to pellet the protoplasts. The supernatant was removed and the protoplasts were re-suspended in 1 mL W5 buffer per 20,000 protoplasts and incubated 16-20 h in a growth chamber (Percival Scientific, Perry, IA, USA) at 23°C with a photon flux density of approximately 30 µmol m^{-2} s^{-1}. When more than 20,000 protoplasts were required, the previous procedure was scaled up to 60,000-160,000 protoplast transfections, adding enough DNA, PEG solution and W5 buffer to maintain the original ratios. Western blots of total protein were performed using 100,000 protoplasts while Western blots of crude chloroplast and cytosol fractions were performed using 160,000 protoplasts. Number of protoplasts prepared for imaging on confocal microscope varied.

2.7.3 Screening and Imaging of Transfected Protoplasts

After incubation the protoplasts were allowed to settle to the bottom and most supernatant was removed so that only 1 mL of buffer remained, in which they were re-suspended by gently swirling the tube. 10 µl of concentrated protoplasts was then removed from the tubes and mounted on a depression slide made in-house for screening on a Zeiss AxioImager D1 epifluorescence microscope (Carl Zeiss Canada Inc., Toronto, ON, Canada). The in-house depression slide was made using a standard microscope slide with cover glass attached to define a rectangular shallow well into which the samples were loaded. Filter 38 HE GFP was used when viewing fluorescence of GFP which has an excitation and emission wavelength of 470 and 525 nm respectively. Filter 43 HE CY3 was used to view chlorophyll autofluorescence in the case of protoplasts, DsRed and mitotracker dye in the case of onion cells, which all appear red and have an excitation and emission wavelength of 550 and 570 nm respectively. Screening was performed on protoplasts before the protein collection described in the following sections.

When capturing images of protoplasts, 100 µL of the resuspended concentrated protoplasts was loaded into a Thermo Fisher Nunc Lab Tek II chamber slide and viewed using an inverted Olympus FluoView 1000 confocal laser-scanning microscope (Olympus Co., Japan) equipped with a multi-line argon laser, a helium-neon (He-Ne) laser at 543 nm, two diode (blue and red) lasers, and a mercury lamp using the 40X Plan Fluor objective (numerical aperture 0.6). A z-stack of optical sections (0.75-1.0 µm optical slice thickness, 10-12 slices) were acquired and processed using the Olympus Fluoview imaging software (FV10-ASW 1.7). For monitoring EGFP the 488 nm line of the Argon laser was used and the emitted fluorescence was collected with the BA505–525 filter. For detection of chlorophyll autofluorescence, the He-Ne laser was used and the emitted fluorescence signals were collected with the BA655-755 filter. The experiments were repeated at least three times.

2.7.4 Protoplast Cellular Fractionation and Protein Collection

If a transfection rate of 60% or higher was observed on the epifluorescence microscope, protoplasts were processed as follows to collect the total protein. The concentrated protoplasts remaining in the tubes after screening were allowed to settle, the remainder of the W5 buffer was discarded, and the protoplasts were re-suspended in 100 µL of solubilization buffer [100 mM Tris-HCl, 100 mM NaCl, 1% SDS and 1% Triton X-100] per 20,000 cells. The protoplasts in solubilization buffer were vortexed for 2 min to break the plasma membranes, centrifuged at 15,000 rpm for 15 min at RT, the supernatant containing the total soluble extract was transferred to a clean 1.5 mL microfuge tube and the pellet was discarded. Protein was precipitated from the supernatant by adding 4 volumes of acetone and incubating for 2 h at -20ºC followed by centrifugation for 15 min 14,000 rpm at 4ºC. The supernatant was discarded, and the pellet was allowed to air dry.

Once three total protein replicates were collected as above, transfection was scaled up to 160,000 to 200,000 protoplasts for lysis and fractionation into crude chloroplast and cytosol protein fractions. Only protoplast samples which had an estimated transfection rate of 75% or higher were processed for fractionation as follows. The protoplasts were first suspended in 600 µL lysis HS buffer [50 mM HEPES-KOH (pH 7.3), 330 mM sorbitol, 1 mM PMSF] and then forced through 10 µm nylon mesh attached to the end of a 3 mL syringe into a microfuge tube 300 µL at a time, replacing the mesh in between. The lysed sample was then centrifuged at 5000 x g for 5 minutes to pellet intact chloroplasts, cell membranes, and any insoluble proteins that may have been in the cells (crude chloroplast fraction). The supernatant (crude cytosol fraction) was transferred to a new tube and protein was precipitated by adding 4 volumes of acetone and incubating at -20°C for 2 hr followed by centrifugation at 14,000 rpm at 4°C for 20 min. The supernatant containing the acetone was discarded and the protein pellet was allowed to air dry before re-suspending in 20 µL 4x SDS loading dye [0.2 M Tris-HCl (pH 6.8), 40% (v/v) glycerol, 8% (w/v) SDS, 0.12% (w/v) bromophenol blue and 6% β-mercaptoethanol]. The chloroplast fraction was re-suspended in 10 µL solubilization buffer [50 mM Tris-HCL (pH 8), mM EDTA, 0.2% w/v SDS] and pipetted repeatedly to break chloroplast membranes. 15 µL 4X SDS sample buffer was added to this fraction as well before boiling.

Minimum transfection rates at screening were chosen to increase the likelihood that enough recombinant protein would be present in the processed protein fractions to be detected on the Western blots.

2.7.5 SDS-PAGE

Protein pellets were boiled at 95°C in SDS loading dye for 10 minutes before being loaded onto SDS-PAGE gels in a Mini-Protean Tetra Electrophoresis Cell (Bio-Rad,

Mississauga, ON, Canada). A 10% separating gel was prepared as follows; 2 mL DI water, 1.25 mL 1 M Tris pH 8.8, 1.7 mL 30% (w/v) acrylamide, 50 µL of 10% SDS, 50 µL of 10% APS and 5 µL TEMED. A 4.8% stacking gel was prepared as follows; 1.7 mL DI water, 313 µL 1 M Tris pH 6.8, 416 µL 30% (w/v) acrylamide, 25 µL of 10% SDS, 50 µL of 10% APS and 5 µL TEMED. Samples were run at 90 V until they entered the separating gel at which point the voltage was turned up to 140 V. The run was stopped when the loading dye began to run off the bottom of the gel. The stacking gel was trimmed away from the separating gel and the proteins were transferred from the separating gel to a PVDF membrane after soaking in transfer buffer as described below.

2.7.6 Western Blot of Transfected Protoplasts

Proteins separated on SDS-PAGE gel were transferred onto a PVDF membrane, which had previously been briefly wetted with methanol and soaked in transfer buffer [48 mM Tris, 39 mM glycine, 1.3 mM SDS and 2% v/v methanol], at 20 V for 30 min using a Trans-Blot SD Semi-Dry Electrophoretic Transfer Cell (Bio-Rad, Mississauga, ON, Canada).

To visualize total proteins transferred, the membrane was then incubated in Ponceau stain [0.2 % w/v ponceau S (Sigma-Aldrich P-35040), 10 % acetic acid] for 5 minutes on a shake plate at RT and rinsed thoroughly in DI water. At this stage a photograph was taken of the membrane and the transfer was evaluated to see if protein bands were visible and, in the case of fractionated protein, how much of the protein in the band found at the molecular weight of the large subunit of Rubisco was in the chloroplast pellet lane. The membrane was trimmed down to the desired lanes and then gently shaken in blocking buffer of 5% (w/v) skim milk powder in TBS-T buffer [25 mM Tris-HCl (pH 7.4), 137 mM NaCl, 2.7 mM KCl, and 0.05% (v/v) Tween-20] for 1 h at RT. Blocking buffer was discarded and the membrane was incubated in primary antibody

(Rabbit α-GFP) diluted 1:5,000 in TBST containing 5% skim milk for 2 h at RT or O/N at 4℃

with gentle shaking. The membrane was washed three times in TBST buffer for 5 min each with

gentle shaking at RT. The membrane was then incubated in secondary antibody (Horse radish

peroxidase tagged α rabbit) diluted 1:250,000 in TBST containing 5% skim milk at the same

temperature and time conditions as were used for the primary antibody incubation. Three rinses

in TBST were then performed again as above.

To visualize immuno-detected bands, 300 µL of peroxide solution and 300 µL of

Luminol from Clarity Western ECL Substrate (BioRad cat # 070-5061) were mixed in a 1.5 mL

tube. The rinsed PVDF membrane was laid out on plastic wrap and the A+B imaging solution

was applied to it using a pipette to evenly distribute the solution across the surface of the

membrane. The membrane was then covered with cling wrap and kept in the dark for as short a

time as possible while being transferred into the BioRad ChemiDoc MP Imaging System.

Imaging was performed first using the "chemi-blot Hi sensitivity" settings then the

"colorimetric" settings to capture the ladder. The two captured images were then trimmed and

edited together to create a single image allowing the viewer to compare band sizes to the ladder.

2.8 Bioinformatic Structure Predictions

DNA sequences for both isoforms were submitted to a number of online tools for

structure prediction and amino acid analysis. The long isoform OEP15-1 was always submitted

as only the two exons, with the annotated intron removed. PSIPred and JPred4 were used to

predict secondary structure (Drozdetskiy et al., 2015; http://bioinf.cs.ucl.ac.uk/psipred/;

https://www.compbio.dundee.ac.uk/jpred/). I-TASSER online server (https://zhanggroup.org/I-

TASSER/) for protein modelling using profile – profile threading was used to predict tertiary

structure for OEP15-1 (Yang et al., 2015). OEP7.3 tertiary structure was predicted by a new

database called AlphaFold (https://www.alphafold.ebi.ac.uk/) and this model will be taken into account as well (Jumper et al., 2021).

Chapter 3: Results

3.1 RT-PCR Analyses to Investigate Isoform Presence in *A. thaliana* Tissue.

cDNA reverse transcribed from total RNA was used as starting material for PCRs designed to test the presence of mRNA encoding OEP15-1 and OEP7.3 in 3-5 week old *A. thaliana* rosette leaves. This tissue was chosen due to ease of tissue harvesting. RNA was treated with DNase before reverse transcription.

3.1.1 OEP15-1 Amplification from cDNA

There are two historical annotations of At4G02482 corresponding to mRNA transcripts which would encode a 134 AA protein and a 66 AA protein, where the 66 AA protein is identical in sequence to the final 66 AA of the longer protein. The larger of the two, referred to as OEP15-1, was the first protein predicted to be encoded at this locus, but this has recently been replaced with an annotation suggesting that the locus encodes the smaller protein, referred to here as OEP7.3 (Lamesch et al., 2012; Cheng et al., 2017). The rationale behind this annotation change is not recorded explicitly in either the database or the literature. Given the uncertainty about the true nature of the protein corresponding to At4g02482, it became necessary to verify whether the larger transcript, OEP15-1, does naturally occur in *A. thaliana,* or otherwise proceed with only OEP7.3 as the focus of the current study.

When the primers originally designed by Yeung and Chuong (2016) were used to amplify the full-length transcript of OEP15-1 were used in PCR with cDNA as starting material, an approximately 500-bp product was consistently produced (Figure 3.1 A). If the original

44

annotation available in databases for OEP15-1 is accurate, the processed, mature mRNA should be 405 bp in length. If instead the mRNA still contained the predicted intron (see Figure 2.1), or if the RNA was contaminated with genomic DNA, the length would be 483 bp.

The 500-bp fragment that is visible in Figure 3.1 A was also present in the negative control sample without RNA starting material, so the purity of the RNA needed to be verified. A forward primer was designed to compliment DNA beginning 436 bp upstream of the OEP7.3 start site. This location was chosen because it is predicted to be part of an intron by netgene2, meaning that even if the 5' UTR were long or another gene was encoded in that region, this site should not be present in total RNA (Hebsgaard et al., 1996). This +436bp primer was paired with a reverse primer which was inside the CDS of both isoforms for PCR using cDNA (from total RNA) and total DNA as starting material (Figure 3.1 B). Only when total DNA was used as the starting material was a band of ~500 bp produced, indicating that the RNA preparation did not contain DNA complementary to the +436 bp primer. Therefore, it was concluded that there were no DNA contaminants in the RNA used to reverse transcribe the cDNA. The band in the negative control in Figure 3.1 A could indicate contamination in another PCR component.
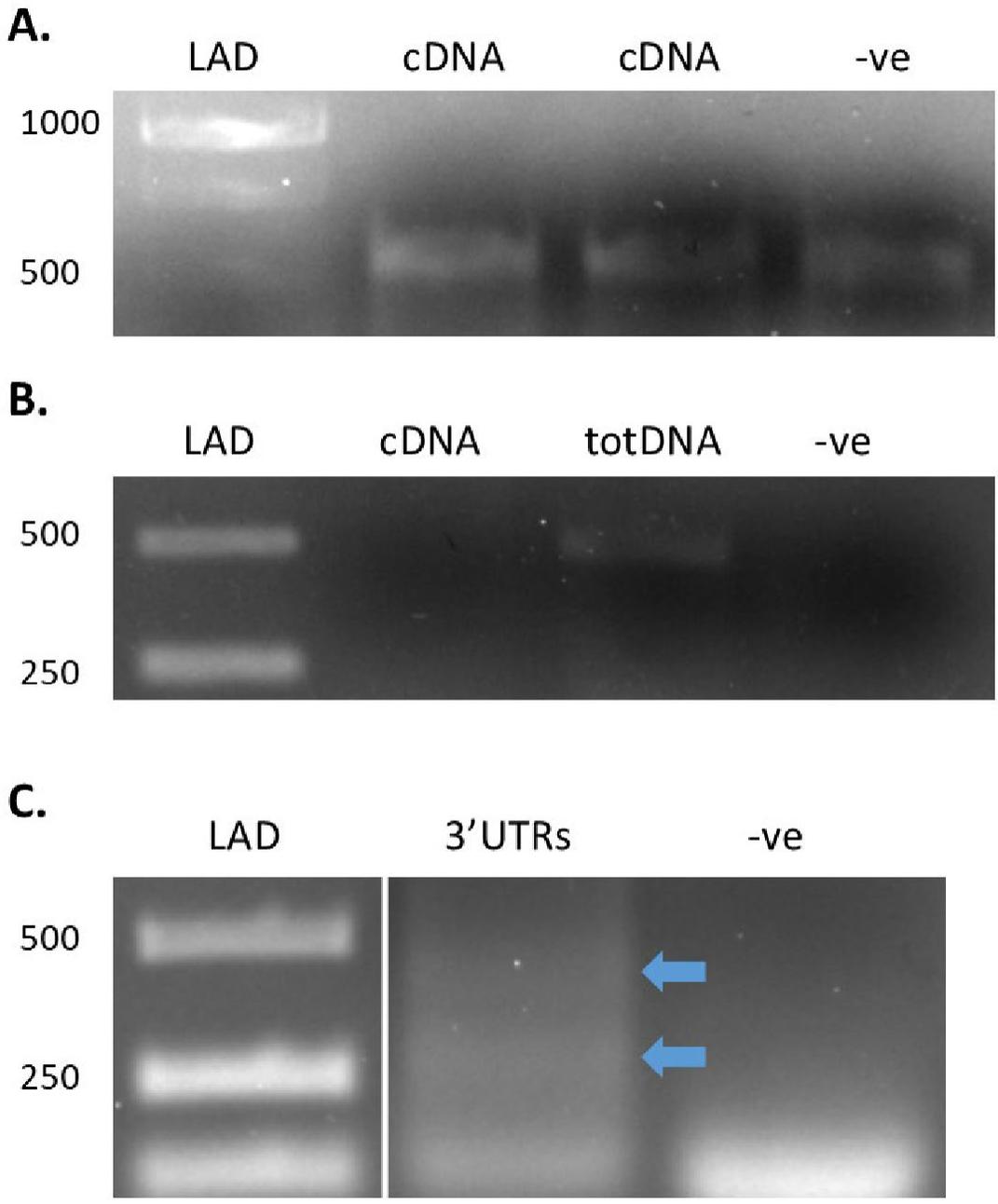
Figure 3.1

Please see legend on next page

Figure 3.1: PCR Verification of mRNA Presence. Three gels visualizing PCR reaction products generated to verify the presence of both isoforms mRNA in *A. thaliana* rosette leaves. Reactions run in lanes marked cDNA or 3'UTRs were performed using cDNA reverse transcribed from total RNA treated with DNAse as the template. Lanes marked -ve are the same reaction without this starting material. All reactions were run on 1.5% agarose gel next to a 1Kb plus DNA ladder (NEB cat #N3200L). A. Gel showing amplified products when PCR is performed using primers designed to amplify full length OEP15-1. B. A primer with a binding site 436 bp upstream of the OEP7.3 start site was paired with a reverse primer in OEP7.3 CDS (R4BAMH1) to test RNA purity. The reaction run on the lane marked totDNA was performed using a total DNA extraction from rosette leaves as the template instead of the cDNA. C. Gel showing products generated when a forward primer inside the CDS of OEP7.3 (OEP7.3F2) and randomly anchored oligo DTs were used to PCR amplify off of cDNA. Blue arrows indicate faintly detected bands. Primer positions can be viewed in Figure A4.

3.1.2 Two 3' UTRs Can be Amplified Downstream of OEP7.3.

To determine if one or both of the two predicted mRNA transcripts are produced in Arabidopsis the 3' UTRs were amplified. When a PCR was performed using a forward primer 134 bp upstream of the 3' end of OEP15-1 in combination with anchored oligo dTs, two fragments of unique lengths (one slightly larger than 250 bp and one roughly 450 bp) were amplified (Figure 3.1 C). This was intended as a first step towards designing a PCR strategy for determining the two transcript's tissue-specific expression patterns which will be described in section 4.5 Future Directions as it was not completed as part of the current study.

3.2 Some Previous Work on OEP15-1 by Yeung (2016) was Discarded.

When Kelly Yeung prepared full-length and truncated clones of OEP15-1, she began by cloning the product amplified from cDNA using a primer set designed to amplify the full-length cDNA into pBluescript (OEP15-1pBs). This clone was then used as the starting material to make a series of OEP15-1 GFP fusion constructs under the assumption that the amplified product contained the 405 bp CDS annotated in TAIR10 in 2016. Some of these clones were transiently expressed in onion epidermal cells and in protoplasts (Yeung and Chuong, 2016).

The initial goal of the current project was to repeat the transient expression in protoplasts for the purpose of confirming Yeung's observations using confocal microscopy so that Western blots probed for GFP could then be performed to verify fusion protein localization. When expression in protoplasts was attempted at the beginning of the current thesis project, most clones did not produce GFP in an amount detectable by fluorescence microscopy. Two clones which did result in consistent expression of the GFP fusion proteins did not appear to be localized to the cytosol exclusively. One localized in a pattern which suggested plastid

localization (OEP15-1:GFP-KY) and the other (OEP15-1 NT:GFP) had an expression pattern consistent with protein aggregation. The lack of expression of all other clones led to a re-examination of the clones generated by Yeung.

As an initial trouble shooting effort, all of Yeung's clones were re-sequenced at their junction with GFP to ensure that they were in frame. While all results came back showing in-frame constructs, the sequencing readout for OEP15-1:GFP-KY showed that this clone did not contain OEP15-1 as the insert but another C-terminal TP candidate OEP which was being studied in the lab at the time by another member of the group, Tianlun Zhou. Work on OEP15-1:GFP-KY was discontinued as transient expression had already been characterized thoroughly for this protein in unpublished work by Tianlun Zhou. The lack of expression of the other constructs in protoplasts could not be explained based on the sequencing, as the sequences were as expected for OEP15-1 and were each in frame with the coding sequence for EGFP.

With the knowledge that veracity of the work performed by Yeung (2016) could not be assumed, the cDNA insert of OEP15-1pBs, the clone used as starting material for pSAT6 constructs, was sequenced. The sequence results are shown in Figure 3.2 aligned with OEP15-1 exons as annotated in TAIR10. The alignment reveals that the cDNA insert contains additional sequence which corresponds to the genomic intron sequence. In Figure 3.2, the predicted intron is highlighted and the internal stop codon it contains is framed by a red rectangle. Because the intron was included in the clone used as starting material for GFP fusion constructs, any construct in which EGFP was fused to the C-terminus downstream of the intron would not express EGFP effectively, as translation would be terminated at the stop codon that is located in the intron. The expression of OEP15-1 NT:GFP was possible because the amplified segment of OEP15-1 was upstream of the intron but all other clones had to be discarded.
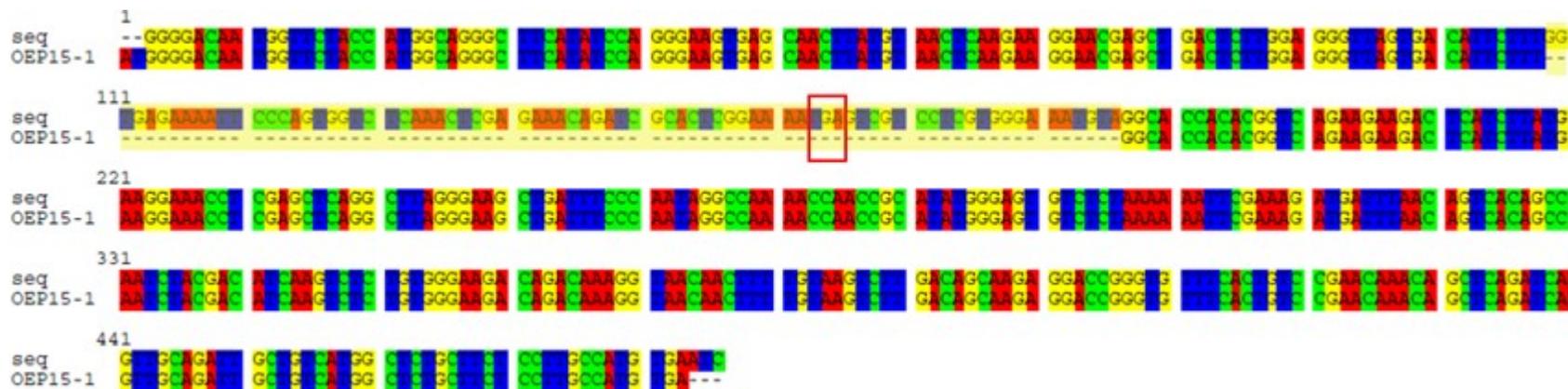
Figure 3.2: Sequence results for OEP15-1pBs compared to OEP15-1 annotated sequence. An alignment between the sequence results (top row, seq) and the annotation for OEP15-1. Predicted intron which is present in the sequence results is highlighted in yellow. The in-frame Stop codon is indicated by the red box. Image generated in seaview.

3.3 Predicted Secondary Structures of OEP15-1 and OEP7.3.

The secondary structure of OEP15-1 was predicted by JPred4 (Figure 3.3). OEP15-1 is predicted to contain many β-strands and 3-4 alpha helices. The helix at the C-terminus was hypothesized to contain COM targeting information and is unlikely to function as a TMD due to its length of only 12 AA while TMD α-helices are usually ~20 AA in length (Teresinski et al., 2019). OEP7.3 is roughly half the length of OEP15-1, identical in sequence to the C-terminus of OEP15-1 and is indicated in Figure 3.3 by a yellow line. The alpha helix at the N-terminus of OEP7.3 is not predicted when OEP7.3 sequence is submitted alone to PSIpred structure prediction server (Figure 3.5), but the two proteins share the C-terminal alpha helix and the four upstream β-strands.

3.4 Construct Design

All OEP15-1 fusion constructs were designed as though the gene is spliced to remove the intron between Exon 1 and Exon 2, as was originally predicted, and results in the production of a 134 AA length protein product. Figure 3.4 provides a graphical overview of constructs designed to produce EGFP fusion proteins in plant cells. OEP15-1 insert sequences and target vectors for those that were ordered are listed in Appendix Table A1. Protein coding sequences for all OEP15-1 inserts excluding OEP15-1NT and the full length OEP7.3 were submitted to BioBasic for synthesis and the sequences were codon-optimized for expression in *A. thaliana* by BioBasic. Codon optimization does not alter the amino acid sequence of the protein product. OEP15-1 constructs were designed to include or exclude the C-terminus containing the predicted transit peptide or the N-terminus which would contain the transit peptide if the pre-protein were stroma-bound. The TP lengths at both termini as predicted by ChloroP were 44 AA at the C-terminus and 40 AA at the N-terminus (Grimberg, 2016). The regions included in OEP7.3 constructs were

based around the secondary structures of interest as predicted by PSIpred (Figure 3.5) and their names reflect which predicted secondary structures were included in each (Buchan and Jones, 2019). Notably, the PSIpred workbench also includes a membrane association prediction tool, Memstat, which predicts that the α-helix is membrane associated (Buchan and Jones, 2019).

Another construct ordered from BioBasic, hFH:DsRed was used as a mitochondrial marker. This construct was designed so that the leader sequence of human fumarate hydratase (hFH), a known mitochondria targeting protein, could direct the localization of DsRed to mitochondria in plant cells. The sequence for the hFH leader sequence was taken from Lee et al., (2020).

3.5 Transient Expression in Onion Epidermal Cells

In onion epidermal cells, a few different organelles and structures will appear as punctate structures when marked with fluorescent proteins. Co-expression with proteins that have known single organelle localization avoids confusion as the overlapping signals then indicate that the protein of interest is localizing to the same organelle as the co-expressed protein. For this reason, each pSAT6 construct (Figure 3.4) was transiently co-expressed in onion epidermal cells with a DsRed construct tagged with the ferredoxin transit peptide (FdTP-DsRed), a known plastid localized protein (Taiz and Zeiger, 2006). Fluorescent signals were captured in gray scale and then pseudo coloured using the Zeiss AxioVision imaging software so that GFP appears green, DsRed appears red and co-localization will appear yellow. Expression of proteins encoded on the pSAT6 vector are controlled by a 35S promoter and GFP expression was too low to detect with the epifluorescent microscope camera until about 20 hrs incubation; therefore, 20 hrs incubation was allowed before photographing for all constructs.
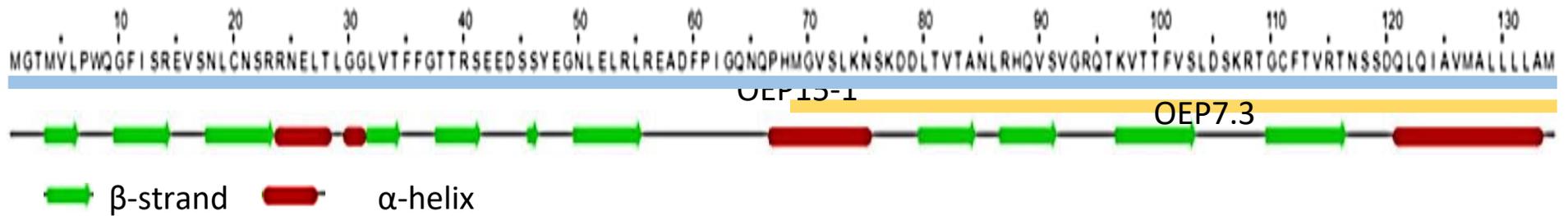
Figure 3.3: Secondary Structure Prediction for OEP15-1 and OEP7.3. Secondary structure as predicted by JPred4 (Drozdetskiy et al., 2015). Green arrows represent β-strands and red cylinders represent α-helices. Amino acid sequence and position numbers are shown at the top of the image. A blue bar represents parts of the sequence included in OEP15-1 and a yellow bar represents parts of the sequence included in OEP7.2.
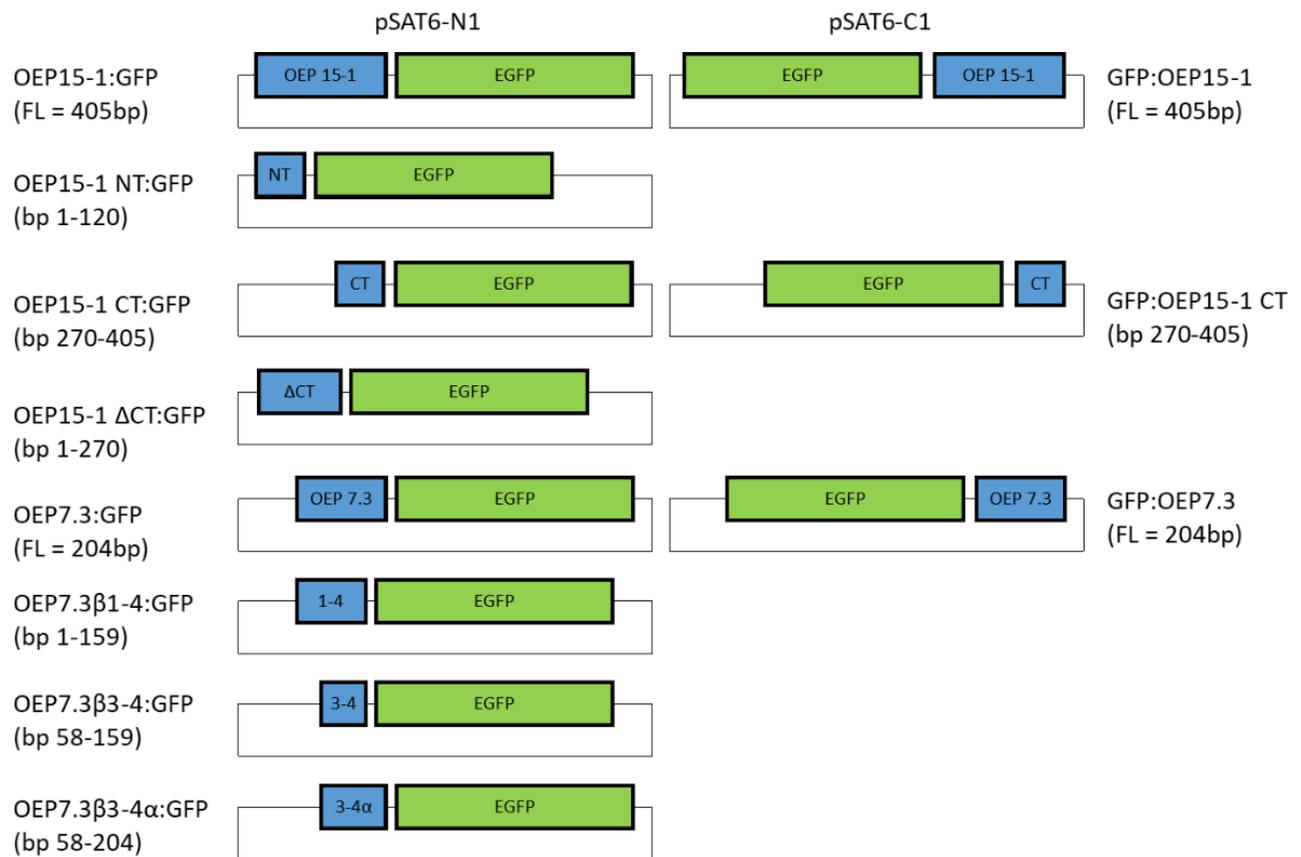
Figure 3.4: Schematic of GFP-Tagged Constructs. Diagram of 11 constructs used to view protein localization in onion epidermal cells and *A. thaliana* protoplasts. Each construct is pictured next to a name which will be used to refer to it throughout the document and the base pairs from the full length which it contains. Constructs in the left column were cloned into the vector pSAT6-N1 so that the GFP was fused to the C-terminus of the OEP protein fragment. Any stop codons in the RNA were removed from these constructs. The constructs in the right column were cloned into the vector pSAT6-C1 so that the GFP was fused to the N-terminus of the OEP protein fragment and STOP codons at the C-terminus were left present.

Figure 3.5: OEP7.3 Truncation Design with Predicted Secondary Structures. Amino Acid sequence of OEP7.3 and 3 truncation constructs listed on left. Secondary structures as predicted by PSIpred. Are shown by coloured blocks around letter code where β-strands are indicated by yellow and α-helices are indicated by pink. Remaining grey blocks represent random coils. (Buchan and Jones, 2019).

In onion epidermal cells, the empty pSAT6-N1 vector encoding EGFP alone has an expression pattern which shows faint diffuse signal throughout the cytosol which becomes more intense towards the edges of the cell (Figure 3.6). There is also brighter signal at the nucleus and in the transvacuolar strands. None of the control signal overlaps with the punctate DsRed signal. All recombinant proteins expressed in onion epidermal cells showed both punctate structures and cytosolic signal, though the number of punctate signals and fluorescence intensity of both punctate and cytosolic signal varied.

While all of the constructs had some punctate expression pattern when expressed in onion epidermal cells, not all of these signals overlapped with the DsRed signal. The non-overlapping punctate signals are not likely to correspond to plastids (Figure 3.6 and 3.7). Localization of GFP to plastids in onion epidermal cells generally appear as small pinpricks of signal referred to as punctate structures which overlap with FdTP-DsRed signal, but the EGFP can also localize to small projections that extend from these uniform punctate structures, which may represent stromules (see Figure 3.6 construct OEP15-1ΔCT:GFP for stromule example) (Lung et al., 2014). Those cells that had punctates overlaping with the DsRed signal but which had cloudy cytosolic signal could be only appearing to have plastid shaped signals due to over expression of the fusion protein throughout the cell.

To determine if the punctate structures that did not co-localize with FdTP-DsRed could result from localization to mitochondria instead of plastids, a leader sequence from the human mitochondrial fumarate hydratase (hFH) protein was used to direct DsRed (hFH-DsRed) to the mitochondria and co-expressed with constructs which showed the highest number of punctate structures. As seen in the insets of images 3.6 and 3.7, there was not better co-localization with hFH-DsRed, than with FdTP-DsRed, indicating the GFP punctate signals may represent an

organelle or aggregates of protein caused by misfolding or by overexpression of the fusion construct causing a decline in health of the cell (Lung et al., 2014).

In the second and third rows of Figure 3.6, the full length OEP15-1 with GFP fused to the N-terminus (GFP:OEP15-1) and C-terminus (OEP15-1:GFP) show punctate GFP expression patterns which overlaps with the DsRed decorated plastids well. A few GFP punctates appear next to but not immediately on top of a DsRed signal which could be due to movement of plastids during the time it takes to change the filters on the epifluourescent microscope and capture the complementary image. OEP15-1 without the CT with GFP fused to the C-terminus (OEP15-1 ΔCT:GFP) shows a GFP expression pattern which includes faint punctate signals which do align with plastids as well as the substantial cytosolic signal. Due to the amount of cytosolic signal present in all cells observed for this construct it was not clear if localization to plastids was truly being observed. The C-terminus alone of OEP15-1 with GFP fused to its C-terminus (OEP15-1 CT:GFP) showed a GFP expression pattern which included irregular punctates all of which aligned with the DsRed decorated plastids. Such irregularity could be caused by both fluorescent fusion constructs localizing to the stromules as well, which would also provide evidence that this construct is localizing to plastids. The C-terminus of OEP15-1 with GFP fused to its N-terminus instead (GFP:OEP15-1 CT) shows many GFP punctate signals but none overlap with the DsRed signal, suggesting that the punctate structures formed by this fusion protein correspond to structures other than plastids. Likewise, the N-terminal region of OEP15-1 with GFP fused to its C-terminus (OEP15-1 NT:GFP) shows very little cytosolic GFP signal and many punctates, none of which overlapped with the DsRed signals, thereby ruling out plastid localization for this fusion protein as well. The two constructs with the most punctate signals (OEP15-1NT:GFP and GFP:OEP15-1CT) were also co-expressed with mitochondrial

protein fumarate hydratase's leader sequence but no better overlap was observed, suggesting those punctate GFP signals which do not overlap with plastids are aggregates or another organelle rather than mitochondria localization.

The three OEP7.3 truncation constructs cloned by Dr. Simon Chuong showed an abnormally high number of irregularly sized EGFP signal punctate structures when protoplasts expressing these constructs were screened at 20 hrs, but EGFP expression was difficult to visualize at all if cells were screened earlier. Cells with higher levels of EGFP fusion proteins after 20 hrs of expression also contained more punctates than would appear if a single organelle type were being localized to. At the time of clone development, insert size and orientation was confirmed using PCR screens by Dr. Simon Chuong. Before DNA isolation, transformed *E. coli* colonies were tested again for insert presence and sequenced. Sequence results showed that the restriction site chosen to make quick blunt end cloning possible, SmaI, caused the addition of 21 upstream amino acids. This was due to the restriction site being in frame with both a methionine near the beginning of the MCS of the vector (pSAT6-N1) as well as GFP (Figure A1).

When the amino acid sequence of the recombinant protein including the 21 amino acids is entered into the online server AGGRESCAN, the first 14 amino acids are predicted to be an aggregate prone "hot spot" (Conchillo-Solé et al., 2007) (Figure A1). Importantly, when the 21 AA peptide is entered into the same server alone, the server does not predict any hotspots. The punctate structures in these highly expressing cells could therefore be aggregates formed due to dying cells or due to the additional peptide at the N-terminus of each of these fusion proteins. For this reason 20 hrs was chosen for the length of expression after transformation, but only cells with relatively low or moderate expression levels were photographed; cells with stronger EGFP signal were not taken into account when selecting a representative cell due to the higher apparent

58

protein aggregation observed in such cells. The cells which had a lower expression level and fewer punctate signals also had very low cytosolic expression making them harder to visualize both under the microscope and in captured images (see Figure 3.7). Knowing that this additional peptide was encoded by the MCS of pSAT6-N1, all other constructs in this vector were re-examined. None of the constructs ordered from BioBasics included the 21 AA but the OEP15-1NT:GFP construct designed by Yeung and Chuong (2016) , which also forms aggregates when expressed, did (Figure 3.6).

When expressed in onion cells, OEP7.3 full length protein with GFP fused to the C-terminus (OEP7.3:GFP) shows an excess of cytosolic GFP signal (Figure 3.7). OEP7.3 with GFP fused to the N-terminus instead (GFP:OEP7.3) shows high levels of cytosolic GFP expression as well as defined punctate structures which overlap with the FdTP-DsRed signals suggesting localization to plastids (Figure 3.7). The truncations of OEP7.3 which include all 4 predicted β-strands, the 3rd and 4th β-strand only and the 3rd and 4th β-strand plus the predicted C-terminal alpha helix (see Figure 3.5) each have GFP fused to their C-termini and are named OEP7.3β1-4:GFP, OEP7.3β3-4:GFP and OEP7.3β3-4α, respectively. All three show GFP expression in the form of many punctate structures of varying sizes, only some of which overlap with FdTP-DsRed punctate signals in each case. None of the three constructs appear to have GFP signal localizing to more than half of the DsRed-marked plastids; however, OEP7.3β3-4:GFP has many more overlapping signals than OEP7.3β1-4:GFP and many of the GFP signals in the image of OEP7.3β3-4α have a DsRed signal slightly offset as though the signals would have overlapped if the plastids had not moved slightly between photographs. When these OEP7.3 truncation constructs were co-expressed with DsRed targeted to the mitochondria by human fumarate hydratase leader sequence instead (see Figure 3.7, merge mito column), few if any punctate

signals overlap (Lee et al., 2020). Collectively, the onion cell expression data suggests that all three recombinant OEP7.3 truncation proteins target to plastids, albeit inefficiently, with OEP7.3β3-4:GFP having the least efficient targeting of the three.
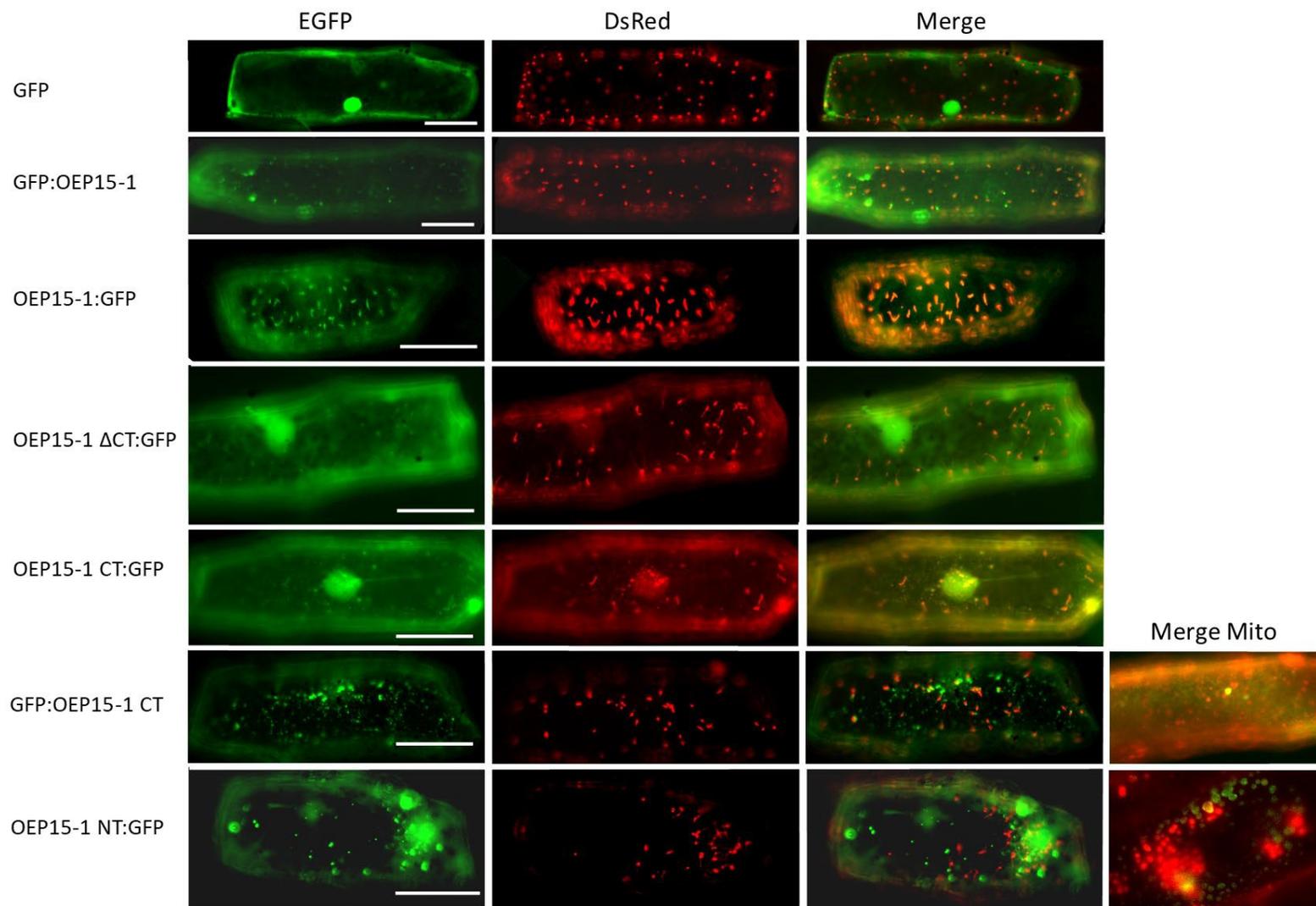
Figure 3.6

Please see legend on next page

Figure 3.6: Onion Epidermal Cells Expressing OEP15-1 Recombinant Proteins fused to EGFP.
Onion epidermal cells were transfected with constructs (indicated on the left) and with Ferredoxin-DsRed and the resulting fluorescence signals were observed using an epifluorescence microscope. EGFP fluorescence was false coloured green (column EGFP). Ferredoxin targets DsRed to plastids and the fluorescence signal was false coloured red (column DsRed). Overlapping fluorescence signals from EGFP and DsRed shown in column labelled Merge appear yellow. Column Merge Mito shows the merge image only of a portion of an onion cell co-expressing a mitochondria leader sequence of human fumarate hydratase fused to DsRed false coloured red, and the EGFP construct in the corresponding row false coloured green again. Scale bar = 30 um.
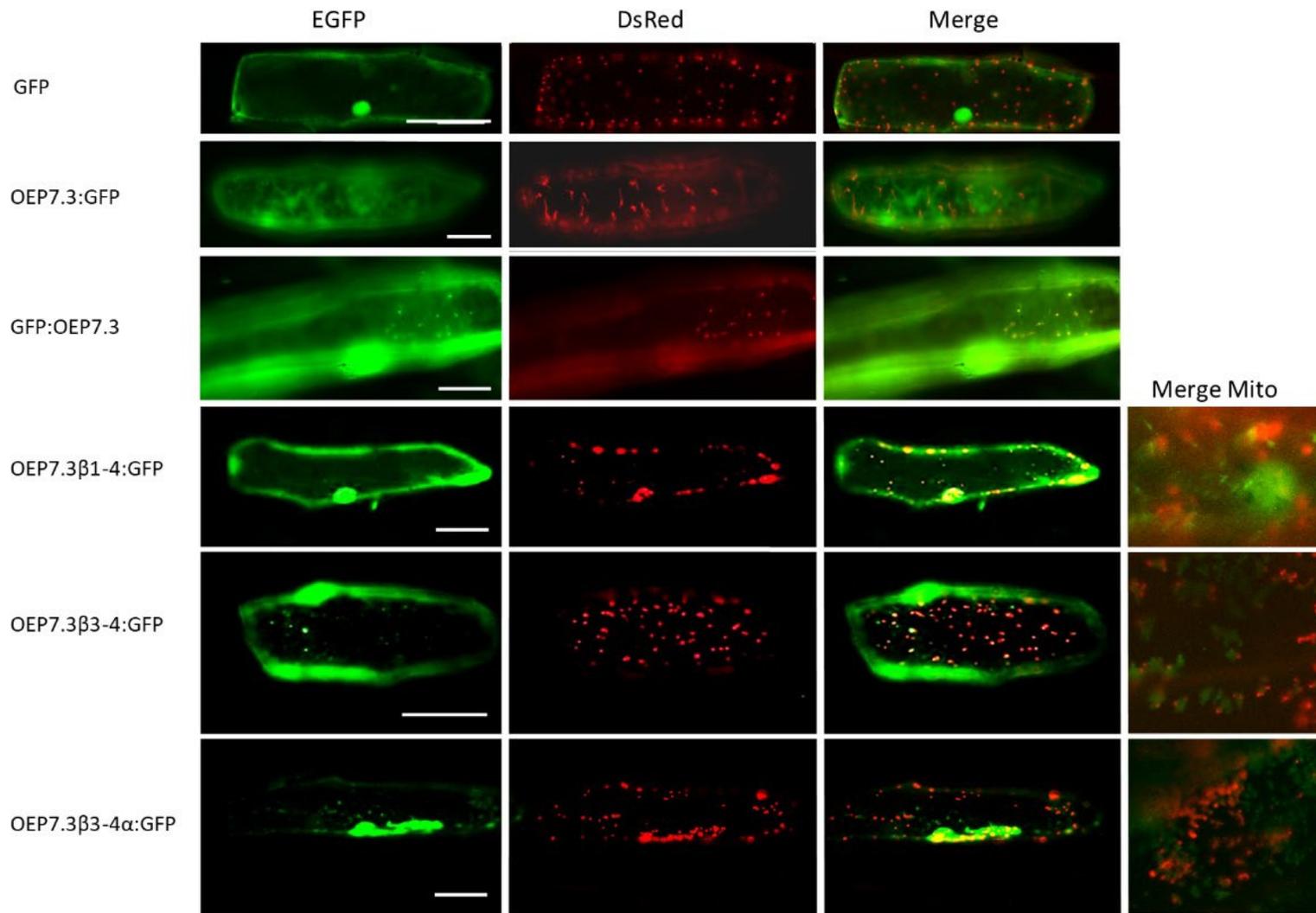
Figure 3.7

Please see legend on next page

Figure 3.7: Onion Epidermal Cells Expressing Recombinant OEP7.3 Proteins fused to EGFP.
Onion epidermal cells were co-transfected with constructs (indicated on the left) and with Ferredoxin-DsRed and the resulting fluorescence signals were observed using an epifluorescence microscope. EGFP fluorescence was false coloured green (column EGFP). Ferredoxin transit peptide targets DsRed to plastids and the fluorescence signal was false coloured red (column DsRed). Overlapping fluorescence signals from EGFP and DsRed shown in column labelled Merge appear yellow. Column Merge Mito shows the merge image only of a portion of an onion cell co-expressing a mitochondria leader sequence of human fumarate hydratase fused to DsRed false coloured red, and the EGFP construct in the corresponding row false coloured green again. Scale bar = 30 um.

3.6 Transient Expression in *A. thaliana* Protoplasts

The expression pattern of the EGFP control in protoplasts appears as a hazy signal in all spaces not taken up by the auto-fluorescent chloroplasts, which is representative of its cytosolic localization (Figure 3.8). In contrast, full length OEP15-1 with GFP fused to either terminus (EGFP:OEP15-1 and OEP15-1:EGFP) display more concentrated GFP signal that closely surrounds each chloroplast. Furthermore, spaces in between the GFP signals can be observed that contain no fluorescence at all. This indicates that the recombinant proteins OEP15-1:EGFP and EGFP:OEP15-1 are localizing to the outer surface of the chloroplasts, either associating closely with the chloroplast or a part of one of the chloroplast envelopes and are not accumulating in the cytosol. Some punctate signals that do not appear to correspond to chloroplasts can also be seen in these cells (Figure 3.8, EGFP:OEP15-1 and OEP15-1:EGFP). When the short isoform (OEP7.3) constructs are expressed in protoplasts (Figure 3.8, EGFP-OEP7.3 and OEP7.3-EGFP), only one orientation of GFP produces a pattern like that of the OEP15-1 full length constructs. EGFP:OEP7.3 shows localization to the outside of the chloroplasts while OEP7.3:EGFP shows only cytosolic signal (Figure 3.8).

The same images for GFP alone are included for reference to the control/ cytosolic expression pattern when truncations are presented (Figure 3.9, 3.10). The C-terminus (defined in Figure 3.4) of OEP15-1 with GFP fused to the N-terminus (EGFP:OEP15-1CT) shows a hazy signal representative of cytosolic localization, while the same C-terminal domain of OEP15-1 with GFP fused to its C-terminus (OEP15-1CT:EGFP) shows localization more closely associated with the outside of the plastids which can be seen clearly by noting the blank spaces without any signal in between some plastids (see top of cell, Figure 3.9). The construct containing all of OEP15-1 except for the extreme C-terminus with GFP fused to the C-terminus

(OEP15-1 ΔCT:EGFP) also shows localization to the outside of the chloroplast, but some signal also appears to overlap with the autofluorescence signal of the plastids suggesting the fusion protein may be localizing to the stroma as well. The N-terminus alone of OEP15-1 with EGFP fused to the C-terminus (EGFP:OEP15-1NT) shows cytosolic signal as well as numerable punctate signals of inconsistent size (Figure 3.9).

The four β-strands of OEP7.3, as well as the last two β-strands alone, each with GFP tagged to their C-terminus (OEP7.3β1-4:EGFP and OEP7.3β3-4:EGFP respectively) show the ring-like pattern of signal around the chloroplasts which suggests localization to the outer membrane as well as punctate structures of various sizes (Figure 3.10). When the C-terminal α-helix is added to β-strands 3 and 4 and the resulting segment is tagged with GFP on the C-terminus (OEP7.3β3-4α:EGFP) the fluorescent pattern appears cytosolic with the addition of many string-like structures crossing the cells at all angles and a faint but consistently present signal inside of the chloroplasts (Figure 3.10). The cells expressing construct OEP7.3β3-4α:EGFP had less signal overall than the other two constructs in this image, so it is likely that the stromal-looking signal is bleed-over from the autofluorescence of the chloroplasts themselves, especially because the borders of the chloroplasts are not sharply defined as would be expected in stroma targeted expression (Lung and Chuong, 2012). OEP7.3β3-4α:EGFP expressing cells also have a suggestion of localization to the outer envelope of only some chloroplasts (see bottom of cell), and the edges of the signals around the chloroplasts are not as sharply defined as in other construct expression patterns. This could be an indication of low targeting efficiency.

	EGFP	Autofluorescence	Merge
EGFP			
EGFP:OEP15-1			
OEP15-1:EGFP			
EGFP:OEP7.3			
OEP7.3:EGFP			
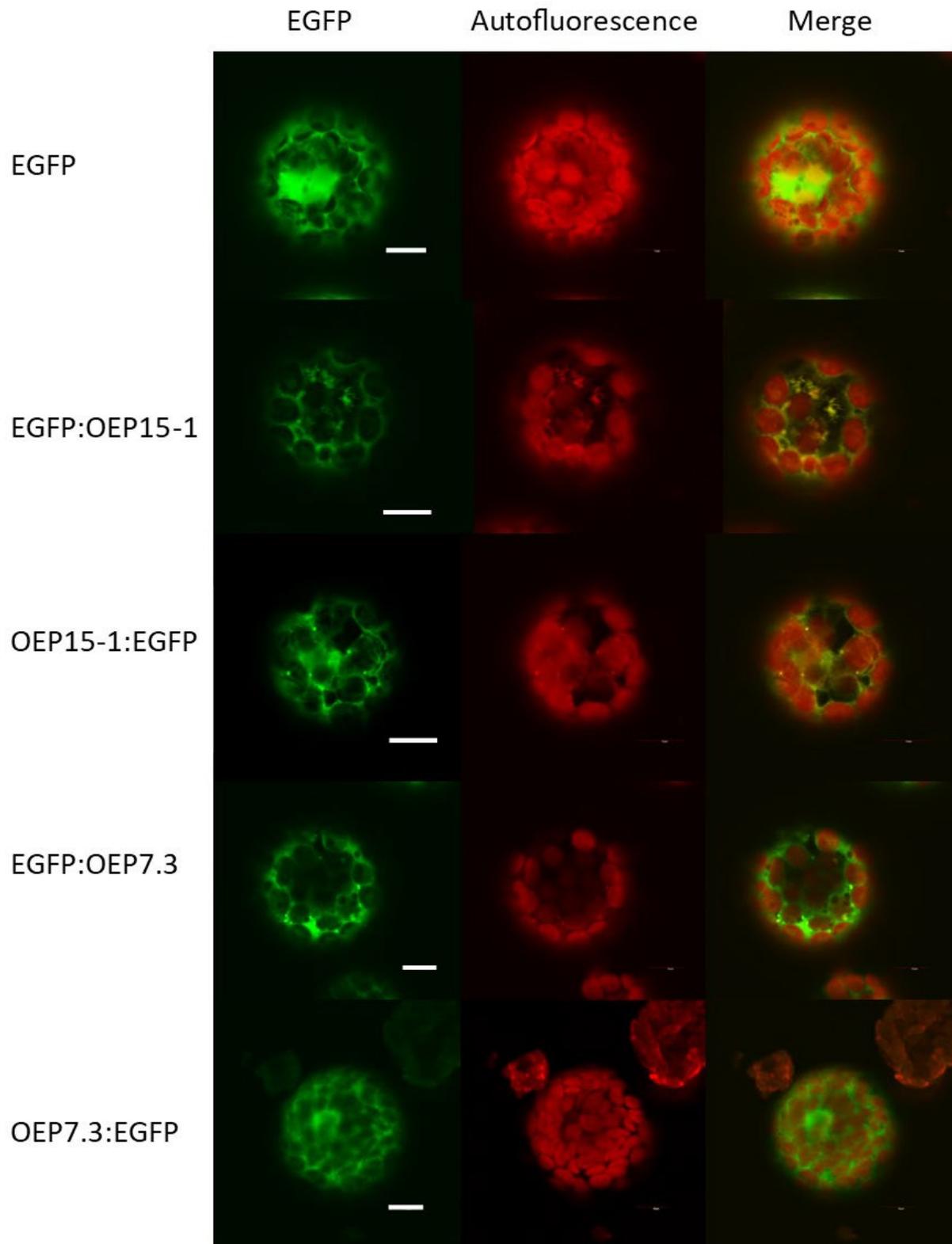
Figure 3.8

Please see legend on next page

Figure 3.8: Protoplasts Transiently Expressing Full Length OEP15-1 and OEP7.3 Fusion Constructs. *A. thaliana* protoplasts transiently expressing recombinant full-length proteins OEP15-1 and OEP7.3 with GFP fused to both termini as indicated by the construct names on the left of the image. GFP signal (EGFP column) allows visualization of recombinant protein localization within the cell and is false coloured green. The natural autofluorescence of the chloroplasts shown in the second column was false coloured red. A yellow signal in the Merge column is caused by overlapping false coloured signals and can indicate recombinant protein localization to the chloroplasts. Scale bars = 10 μm
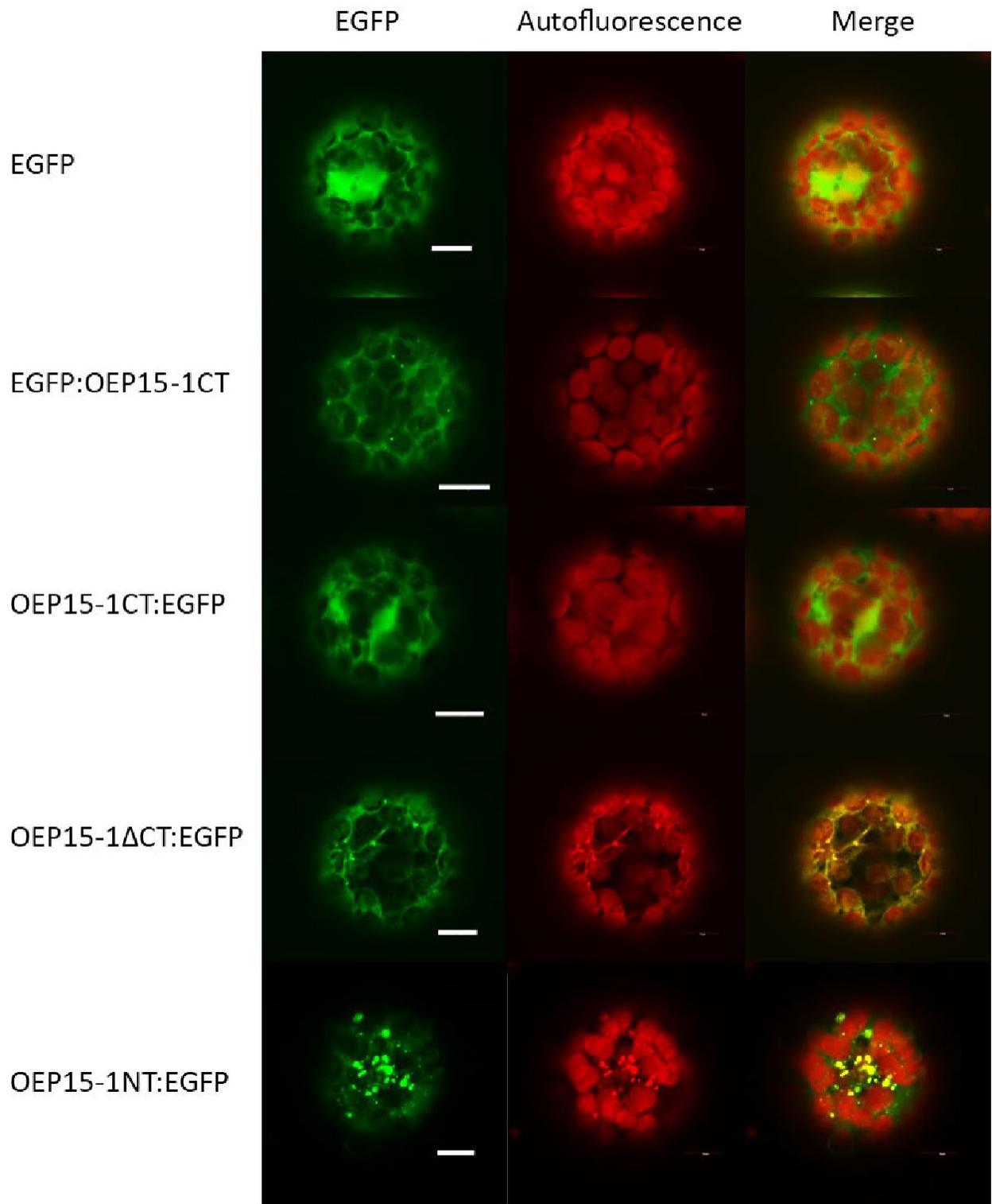
Figure 3.9

Please see legend on next page

Figure 3.9: Protoplasts Transiently Expressing Truncated OEP15-1 Fusion Constructs. *A. thaliana* protoplasts transiently expressing recombinant proteins of OEP15-1 truncations and GFP as indicated by the construct names on the left of the image. GFP signal (EGFP column) allows visualization of recombinant protein localization within the cell and is false coloured green. The natural autofluorescence of the chloroplasts shown in the second column was false coloured red. A yellow signal in the Merge column is caused by overlapping false coloured signal and can indicate recombinant protein localization to the chloroplasts. Scale bars = 10 µm
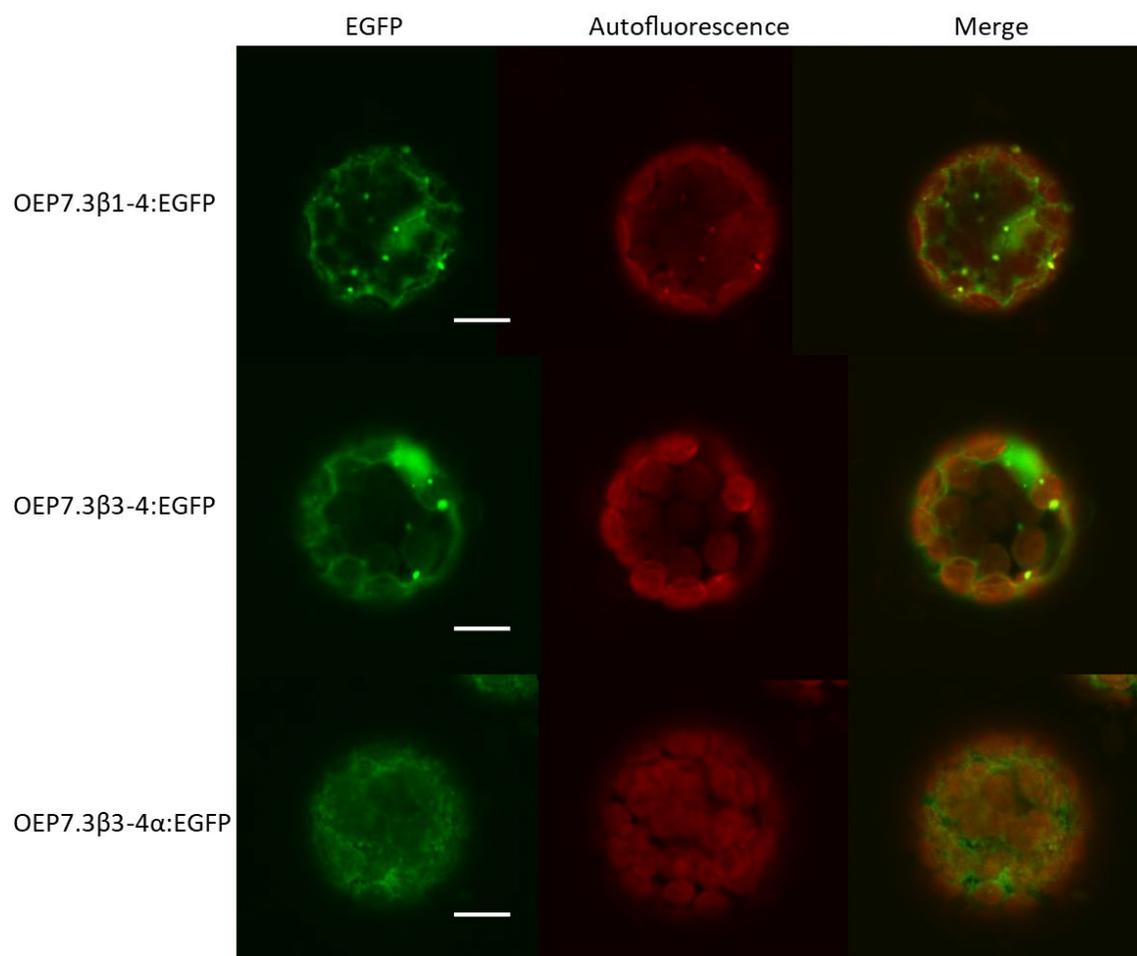
Figure 3.10: Protoplasts Transiently Expressing Truncated OEP7.3 Fusion Constructs. *A. thaliana* protoplasts transiently expressing recombinant proteins of OEP7.3 truncations and GFP as indicated by the construct names on the left of the image. GFP signal (EGFP column) allows visualization of recombinant protein localization within the cell and is false coloured green. The natural autofluorescence of the chloroplasts shown in the second column was false coloured red. A yellow signal in the Merge column is caused by overlapping false coloured signal and can indicate recombinant protein localization to the chloroplasts. Scale bars = 10 μ

3.7 Western Blots

Two bands were detected in total protein extracts from protoplasts expressing all constructs tested when analyzed using Western blots probed with anti-GFP antibody (Figure 3.11). Of those two bands one appears consistently at the expected size for the GFP fusion protein and is the larger of the two bands. The smaller of the two bands appears closer to the size of GFP or slightly larger (~27-30 kDa). Expected product sizes were calculated using Expasy Compute pI/Mw tool using sequence inputs of the protein of interest, any linking residues and EGFP. The expected product sizes for each construct are as follows; OEP15-1:GFP and GFP:OEP15-1 are expected to be 42027.7 Da and 42090.7 Da respectively, OEP15-1CT:GFP and GFP:OEP15-1CT should be 32211.75 Da and 32143.58 Da respectively, OEP15-1ΔCT:GFP should be 37344.18 Da, OEP7.3:GFP and GFP:OEP7.3 should be 34489.32 Da and 34552.34 Da respectively. OEP15-1NT:GFP should be 34260.92 Da and is the only construct for which no degraded product was detected in the total fraction.

The crude cytosolic fraction isolated from protoplasts is likely to also contain the stroma of any accidentally lysed chloroplasts (as well as the soluble contents of other lysed compartments). The crude chloroplast fraction contains intact plastids as well as the outer (and inner) membranes and thylakoids of any broken chloroplasts, plus other whole organelles, membranes and debris. Based on the isolation method the crude chloroplast fraction is also where aggregates of insoluble protein would most likely be found. Ponceau stains of whole blots can be found in appendix Figure A3 and give an idea of how many chloroplasts were lysed by comparing the bands at the molecular weight of the stromal protein RbcL, where the more protein at this size in the cytosolic lane corresponds to more chloroplasts lysed. In some cases most chloroplasts appear to have been lysed but the chloroplast membranes would still be found mostly in the

chloroplast fraction, allowing separation of membrane embedded protein from cytosolic proteins despite breakage.

Some proteins ran lower than expected and if this was the case it was true of all lanes in a single Western blot. This was due to samples being mixed with old SDS loading dye which did not contain enough β-mercaptoethanol.

For most constructs which appear to localize in protoplasts, the chloroplast vs cytosolic Western blots (Figure 3.12, constructs GFP:OEP15-1, OEP15-1CT:GFP and OEP15-1ΔCT:GFP) show the same two band sizes as in the total protein fraction (Figure 3.11) but with the smaller band in the cytosolic fraction and the larger band in the chloroplast fraction. This is also the pattern observed for OEP7.3:GFP (Figure 3.12) though this recombinant protein did not appear to localize to chloroplasts when observed using confocal microscopy. The Western blot for OEP7.3β3-4α:GFP shows a band in the chloroplast fraction which is the expected size of the recombinant protein, and no cytosolic band is detected (Figure 3.12). Two bands were detected in each protein fraction isolated from cells expressing GFP:OEP15-1CT. One band of the expected recombinant protein size was detected in both fractions. A band larger than GFP alone but smaller than the recombinant protein was detected in the chloroplast fraction and a band roughly the size of GFP was very faintly detected in the cytosolic fraction. OEP15-1NT:GFP has only one size of band around the expected recombinant protein size and the majority of the protein was detected in the chloroplast fraction.
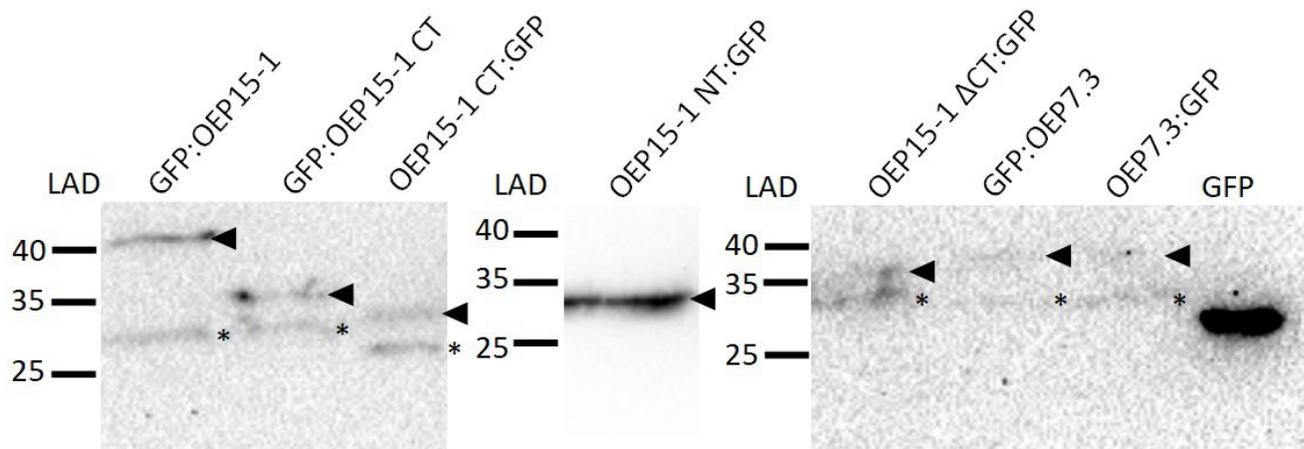
Figure 3.11: Detection of GFP Fusion Proteins in Total Protein Fractions from Transfected Protoplasts by Western Blot. Total protein isolated from protoplasts expressing each construct bound by an antibody against GFP as the primary antibody and a secondary antibody conjugated to peroxidase which was then detected by exposure to peroxide and luminol. Arrow heads point to faintly-detected bands. Purified GFP was included as a control. The image is a combination of two gels run with 10-250 kDa wide range blue/red two colour pre-stained protein ladder (BioBasic cat # BZ0010R). Red-dyed ladder bands did not appear clearly but were at the position marked 25. Molecular weights on ladder are in kDa. n=2 for each construct. Bands at expected molecular weights and therefore representing the full length recombinant proteins are marked by solid black arrow heads. Bands representing degraded products are marked with asterisks.

GFP:OEP15-1

OEP15-1CT:GFP

GFP:OEP15-1CT

OEP15-1ΔCT:GFP

OEP15-1NT:GFP

OEP7.3β3-4α:GFP

OEP7.3:GFP

GFP:OEP7.3
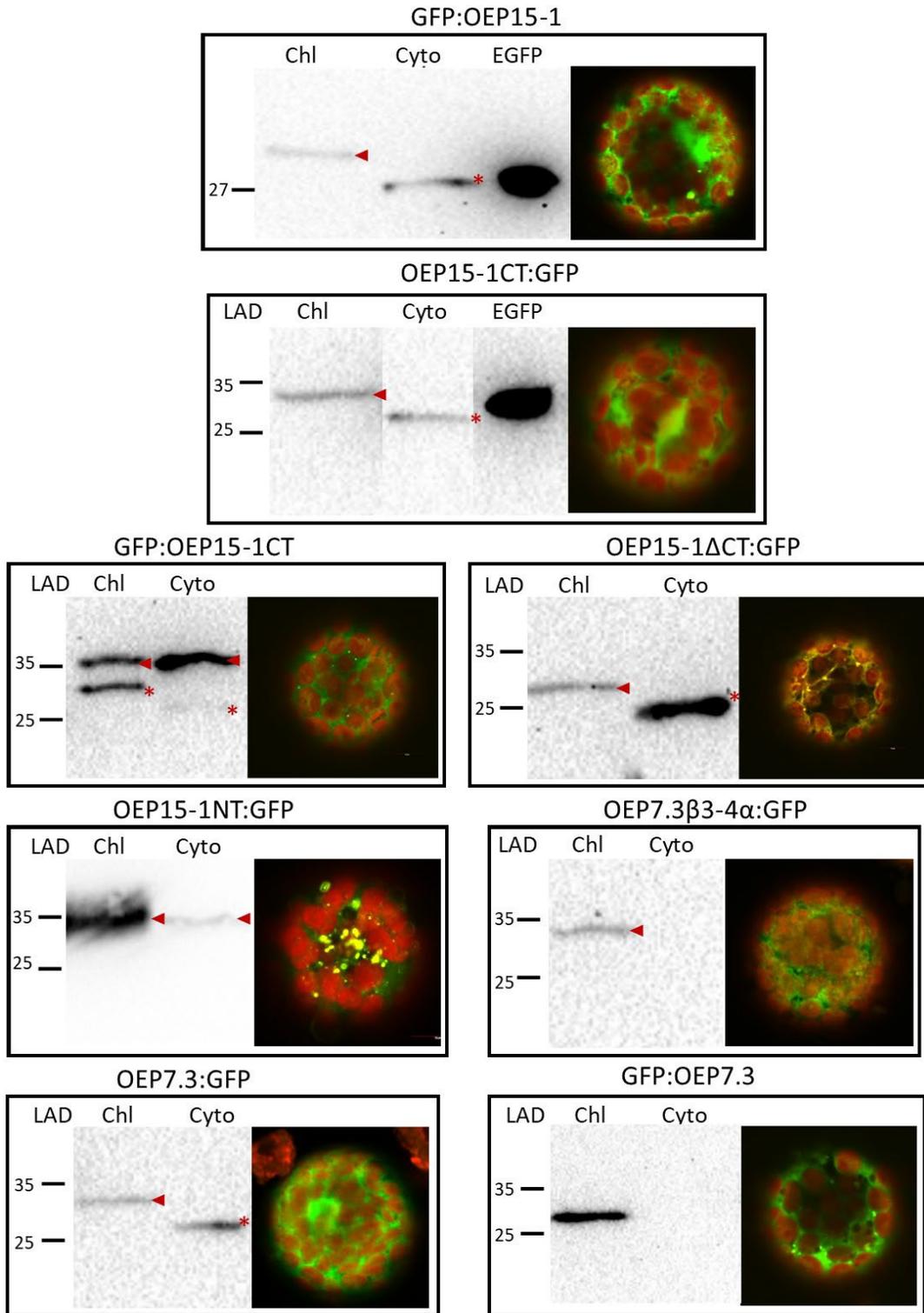
Figure 3.12

Please see legend on next page

Figure 3.12: Detection of GFP Fusion Proteins in Crude Cytosolic and Chloroplast Protein Fractions by Western Blot Next to Merge Confocal Images. Crude cytosolic and crude chloroplast protein fractions isolated from protoplasts expressing each construct bound by an antibody against GFP as the primary antibody and a secondary antibody conjugated to peroxidase which was then detected by exposure to peroxide and luminol. Purified GFP was included as a control in some constructs, however the bright signal it produced interfered with detections of fainter signals so not all blots contain this control lane. The ladder used in all blots is the 10-250 kDa wide range blue/red two colour pertained protein ladder (BioBasic cat # BZ0010R). Molecular weights on ladder are in kDa. Each blot is paired with the merge image of a protoplast transiently expressing the same fusion protein directly to its right. The gel for GFP:OEP15-1 which corresponds to the top blot in this image ran at a lower molecular weight than expected when compared to the ladder and so should be read using the EGFP control as the molecular weight marker (27 kDa). The lane order for construct OEP15-1CT:GFP was changed for ease of reading. n≥2 for all constructs other than OEP7.3β3-4α for which n=1. Bands at expected molecular weights and therefore representing the full length recombinant proteins are marked by red arrow heads. Bands representing degraded products are marked by red asterisks.

3.8 Modeling OEP15-1 After Toc159 Family Proteins.

When the current project was initiated, OEP15-1 was annotated in the TAIR database as a "putative GTPase" because Toc159, a known GTPase, is the protein it resembles the most. GTPases have numerous conserved functional GTP-binding domains called G-domains which Toc159 has three (Smith et al, 2002a). The G1 domain of Toc159 shares the most identity with the corresponding domains in other known GTPases, G2 and G3 are present but the G4 domain found in many GTPases is absent in Toc159 (Smith et al, 2002a). An alignment of both versions of OEP15-1 (OEP15-1 and OEP7.3) with atToc159 revealed that the identity of OEP15-1 with the Toc GTPases did not correspond to the known GTPase domains. Rather, the sequence alignment occurs almost exclusively with the membrane-domain of Toc159 (Figure 3.13) so OEP15-1 and OEP7.3 are likely not GTPases, and that categorizing them as "putative GTPases" would be inaccurate.

With GTPase ruled out as a possible function, BLAST and phylogenetic methods were used to seek clues about the possible function of OEP15-1. These analyses yielded few leads. There were a few "hits" to uncharacterized or putative proteins from many different plant species, but none were informative about a possible function of OEP15-1, as BLAST searches with these uncharacterized proteins as queries revealed that each appeared to be Toc159 homologs due to very high homology with Toc159. Working under the assumption that any functional domains may reside in the small region of OEP15-1 which does not align well to the membrane domain of Toc159, the BLAST search was repeated without the region of the protein which aligns best with Toc159 M-domain. The resulting aligned sequences were ~8 amino acids long and with high E-values and were therefore likely not relevant as E-values measure the likelihood that sequences with the given homology score would be found by chance. Use of the

BLAST algorithm was set aside after this on the assumption that the most insight that could be gained from it was already accomplished with the result that OEP15-1 has high homology to the M-domain of Toc159.

The BLAST results included Toc159 homologues or putative Toc159 proteins from other plant species with higher identity than atToc159 in a few cases. This led to a new question: 'Which Toc159 family member is most similar to OEP15-1?' There are four Toc159 family members in Arabidopsis (atToc159, atToc132, atToc120, atToc90), which are substituted modularly into the TOC complex in Toc159's place to gain specificity for subsets of chloroplast pre-proteins which were each included in a phylogenetic analysis (Ivanova et al., 2004; Richardson et al., 2009; Wise and Hoober, 2006). The species chosen were the species from which the putative Toc159 proteins with highest identity came from, but only the true sequences of the Toc159 family of proteins were compared. Each of the homologs of the Toc159 family proteins from the plant proteomes of *Arabidopsis thaliana* (At), *Arabidopsis lyrata* (Al), *Capsella rubella* (Cr), *Eutrema salsugineum* (Es) and *Brassica napus* (Bn) were included as well to see if there would be an even closer match to proteins from another species, as it was possible some of the proteins annotated as putative Toc159 in less studied species may be more homologous. These protein sequences were used to generate a phylogenetic tree (see Figure 3.14).

The SeaView alignment from which the tree was generated was a G-blocks set which left a length of 116 AAs of aligned proteins, all of which are located towards the C-terminus/ M-domain of the Toc proteins. The G-blocks spanned the vast majority of OEP15-1, a 134 AA protein, which contributed to the difficulty in eliminating the membrane domain during BLAST searches.

```
                                  G1                          G2                      G3
     841
Toc 159   AVESEAEGNE ELIFSLNILV LGKAGVGKSA TINSILGNQI ASIDAFGLST TSVREISGTV NGVKITFIDT
OEP15-1   ---------- ---------- ---------- ---------- ---------- ---------- ----------
OEP7.3    ---------- ---------- ---------- ---------- ---------- ---------- ----------

     911
Toc 159   PGLKSAAMDQ STNAKMLSSV KKVMKKCPPD IVLYVDRLDT QTRDLNNLPL LRTITASLGT SIWKNAIVTL
OEP15-1   ---------- ---------- ---------- ---------- ---------- ---------- ----------
OEP7.3    ---------- ---------- ---------- ---------- ---------- ---------- ----------

                                              ...

     1261
Toc 159   VLDTHGWDHD CGYDGVNAEH SLALASRFPA TATVQVTKDK KEFNIHLDSS VSAKHGENGS TMAGFDIQNV
OEP15-1   ---------- ---------- ---------- ---------- ---------- ---------- -MGTM----V
OEP7.3    ---------- ---------- ---------- ---------- ---------- ---------- ----------

     1331
Toc 159   GKQLAYVVRG ETKFKNLRKN KTTVGGSVTF LGENIATGVK LEDQIALGKR LVLVGSTGTM RSQGDSAYGA
OEP15-1   LPWQGFISRE VSNLCNSRRN ELTLGGLVTF F--------- ---------- -------GTT RSEEDSSYEG
OEP7.3    ---------- ---------- ---------- ---------- ---------- ---------- ----------

     1401
Toc 159   NLEVRLREAD FPIGQDQSSF GLSLVKWRGD LALGANLQSQ VSVGRNSKIA LRAGLNNKMS GQITVRTSSS
OEP15-1   NLELRLREAD FPIGQNQPHM GVSLKNSKDD LTVTANLRHQ VSVGRQTKVT TFVSLDSKRT GCFTVRTNSS
OEP7.3    ---------- ---------M GVSLKNSKDD LTVTANLRHQ VSVGRQTKVT TFVSLDSKRT GCFTVRTNSS

     1471
Toc 159   DQLQIALTAI LPIAMSIYKS IRPEATNDKY SMY
OEP15-1   DQLQIAVMAL LLLAM----- ---------- ---
OEP7.3    DQLQIAVMAL LLLAM----- ---------- ---
```

Figure 3.13: OEP15-1 Does Not Align to the G-Domain of Toc159. A multiple sequence alignment between two isoforms of OEP15-1 (OEP15-1 and OEP7.3) and Toc159 with G domains highlighted on Toc159. Alignment generated in Seaview using ClustalO.
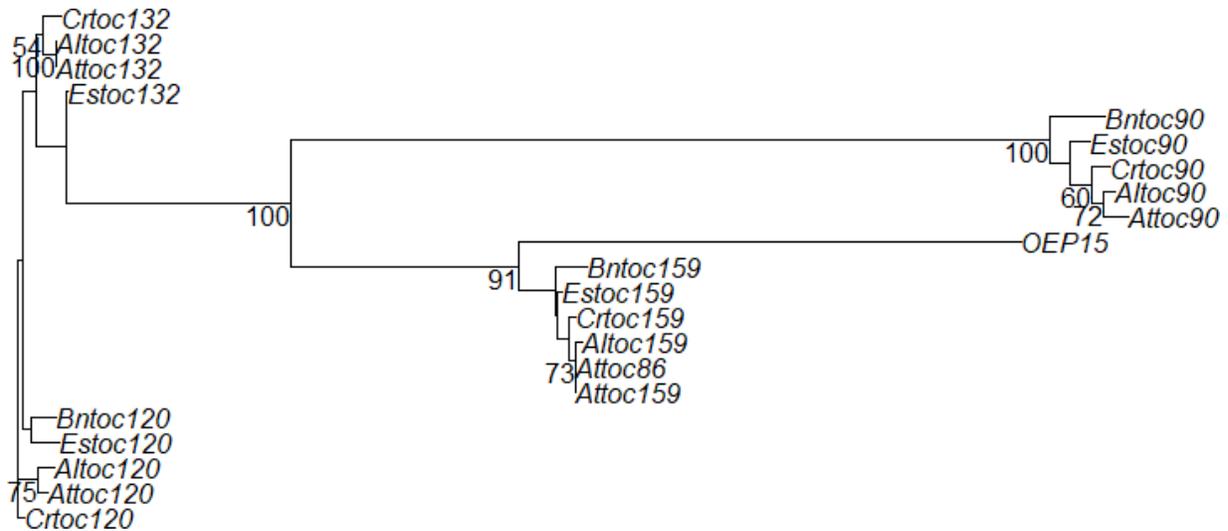
Figure 3.14: Phylogenetic Analysis of AtOEP15-1. Relatedness to Toc159 Family Proteins from Five Plant Species. Tree generated in R package Phangorn using model LG which was chosen using a fit command. Numbers appearing under nodes represent bootstrap values out of 100. Branch lengths un-adjusted. Identities of proteins recorded as genus and species initials followed by translocase type. Species used were *Arabidopsis thaliana (At), Arabidopsis lyrata (Al), Capsella rubella (Cr), Eutrema salsugineum (Es) and Brassica napus (Bn)*. The accession numbers for each gene can be found in appendix Table A2. AtOEP15-1 sequence used and shortened to OEP15 for Figure simplicity.

In Figure 3.14, AtOEP15-1 groups with Toc159 proteins closer than the other translocases. The protein which is most similar to AtOEP15-1 can be found by counting the number of nodes in between it and another protein (Freeman et al., 2004). The protein whose path includes the fewest nodes from the tree in Figure 3.14 is surprisingly the *Brassica napus* homologue of Toc159. The Toc159 and Toc86 proteins from *Arabidopsis thaliana* grouped closest to another *Arabidopsis* species, *A. lyrata,* followed by the *Capsella rubella* and *Eutrema salsugineum* homologues of Toc159. Toc86 is not a unique protein and its inclusion in this tree merely re-enforces that the homology is not in the A-domain of Toc159, which Toc86 is lacking.

As a follow up OEP15-1 was Blasted against the genomes of the closest matching Toc159 homologues (*Brassica napus* and *Eutrema salsugineum*) in hopes of finding a true homolog, but none were found. There is no homolog in the genomes of extant plants which represent ancestral lineages as determined by a Blast search of OEP15-1 against the specific genomes of Amborellaceae (taxid:22097), Nymphea (taxid:4418) and other basal angiosperms with no results other than a Toc159 homologue found in *Amborella trichopoda.* Narrowing the search to only those species which diverged most recently in evolutionary time may increase the odds of there being a homolog in those species. There were no hits in the *E. salsugineum* or *B. napus* genomes which were well aligned in the non-membrane associated region of OEP15-1.

3.9 I-TASSER and AlphaFold Protein Modelling

Both OEP15-1 and OEP7.3 were submitted to online server I-TASSER in using the sequences predicted in their respective genome assemblies (Yang et al., 2015). I-TASSER combines predicted secondary structure and sequence data together for threading through 8 different fold recognition programs which use profile-profile alignments to proteins with known fold structure and each of these programs has a little variation on how it chooses its best match.

From there they build models by keeping only the aligned regions of the templates which they can modularly mix and match and then any regions which don't have a good alignment are modelled from scratch using ab-initio modelling (Yang et al., 2015). The program then generates C-scores for the models which serve as a score of likelihood that the model is accurate and are generated by factoring similarity to models, cluster size and similarity across clusters. The C-scores range from -5 to 2, where higher scores represent more likely models (Yang et al., 2015). When OEP7.3 was submitted in 2020, all models had C-scores less than -4 and the threading readout showed very short sequences having any alignment with the protein templates. For these reasons it was concluded that the server was not equipped to make meaningful models for OEP7.3 and the readout was discarded. OEP15-1 was first submitted in 2019 and the top 5 models chosen by cluster size at this time for OEP15-1 had C-scores ranging from -2.98 to -5. The model which scored the highest C-score and the second highest cluster size is shown in Figure 3.15 for interest due to its likelihood as well as because it resembles a pore (Yang et al., 2015). The I-TASSER server does not store results long-term, and the files require niche software to open on a desktop, so when reviewing this data for inclusion in the thesis, OEP15-1 was re-submitted in 2021 in order to measure the pore size. Because I-TASSER uses all of the PDB as models to thread against, the results generated were different having gained 2 years PDB submissions to compare to. In this new model set, the highest cluster size model had a C-score of -3.11 and was again pore-like visually (Yang et al., 2015). The pore opening diameter was measured in the I-TASSER server view (Figure 3.15 B.2).

Another part of the readout given by I-TASSER is gene ontology (GO) predictions which are generated using the GO terms associated with the threaded proteins used to generate the models. Table 3.1 lists the GO terms associated with the most homologous template proteins,

which were outer membrane protein OmpW in 2019 and an NMR relaxed fold of OmpX in 2021, as well as those that appear in the consensus of the top 10 templates. OmpW is a bacterial protein 356 AA in length which belongs to the porin superfamily of proteins and resides on the plasma membrane (Hua et al., 2012; Lu et al., 2020). OmpX is a homo-hexamer of β-barrel proteins which belong to a family of outer membrane proteins involved in host cell adhesion between *E-coli* and mammalian cells (Vogt and Shultz, 1999). The NMR experiment from which the PDB hit was submitted used a technique in which the protein was re-folded in micelles before the structure was measured, so conformation may differ from the native structure (Fernandez et al., 2001).

OEP7.3 could not be modeled in I-TASSER but is in the current annotation of the *Arabidopsis* genome which was used to generate models in AlphaFold (Jumper et al., 2021). Figure 3.16 is the generated model, and the colours used in the structure represent the program's confidence in the prediction at each amino acid (Jumper et al., 2021). The β-strands are predicted to form a β-sheet with relatively high confidence but are too few to form a β-barrel as a monomer. The program shows less confident predictions at helical residues and at the boundaries of each β-strand (Jumper et al., 2021). No function predictions are included in the AlphaFold entry.
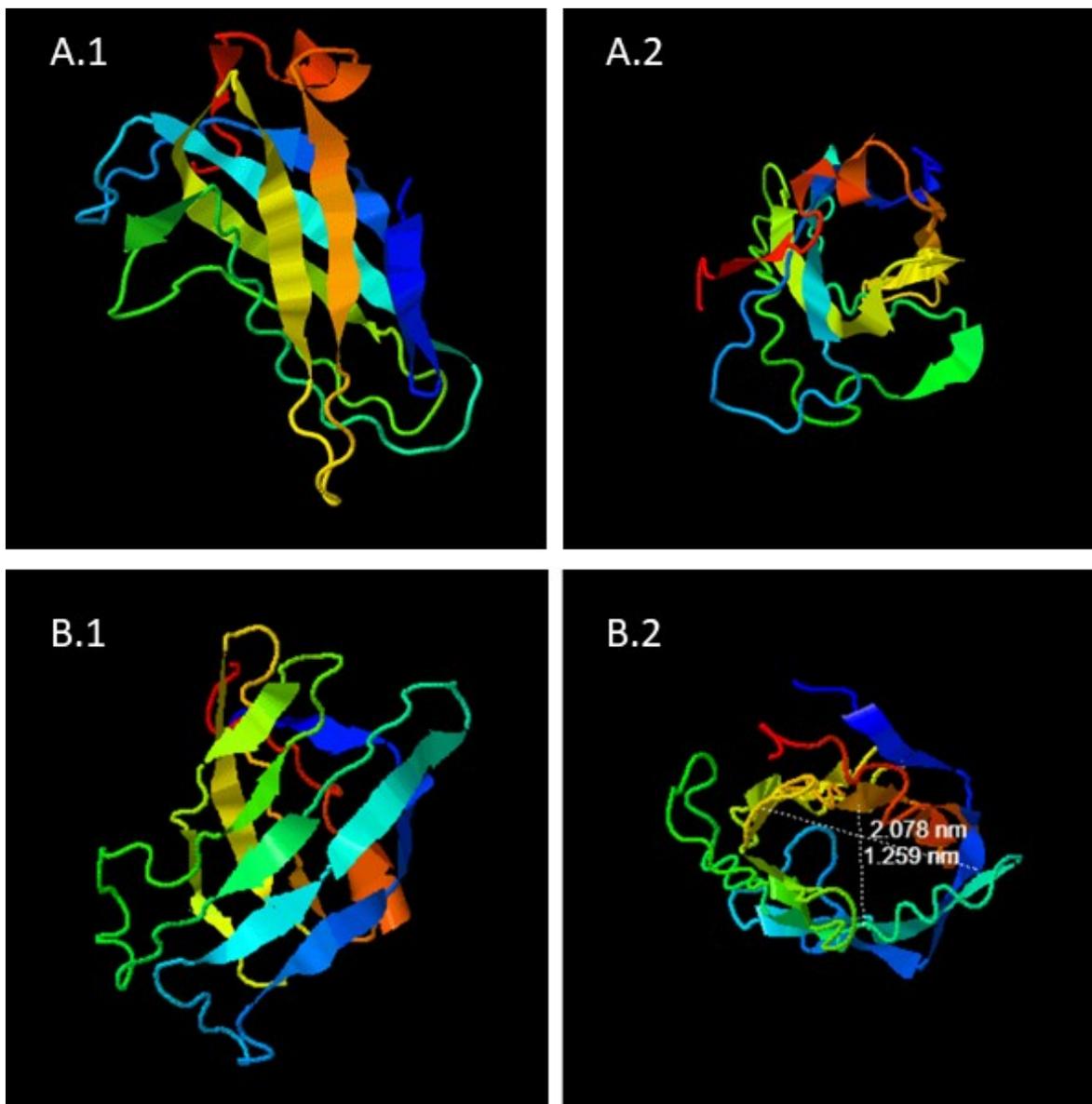
Figure 3.15: I-TASSER Pore-like Models for OEP15-1. Two views each of the two models generated by I-TASSER for input of OEP15-1 full length amino acid sequence. A panels show the model from 2019, A.1 showing β-strands and A.2 showing pore space. This model had the second highest cluster size and the highest C-score of -2.89 in the model set generated that year. B panels show model from 2021, B.1 showing β-strands and B.2 showing pore space with measurements across the asymmetric pore at a middle depth of the model. This model had the highest cluster size and second highest C-score of -3.11 in the model set generated that year (Yang et al., 2015).

Table 3.1: GO Terms for OEP15-1 as Predicted by I-TASSER

GO#	TERM	TOP TEMPLATE?[1]	CONSENCUS?[1]
GO:0016020	Membrane	2019, 2021	
GO:0009279	Cell outer membrane	2019, 2021	2019
GO:0005886	Plasma membrane	2019, 2021	
GO:0016021	Integral component of membrane	2019, 2021	2019
GO:0005515	Protein binding	2019	2019
GO:0019867	Outer membrane	2019	2021
GO:0016853	Isomerase activity		2019, 2021
GO:0006810	Transport		2021
GO:0030312	External encapsulating structure		2021
GO:0030313	Cell envelope		2021
GO:0031224	Intrinsic to membrane		2021

[1]Presence of GO term in the top template protein associations list or in the consensus of the top 10 template proteins. Recorded such that the years in the cells indicate presence of that term in that year and a blank cell indicates a lack of presence.
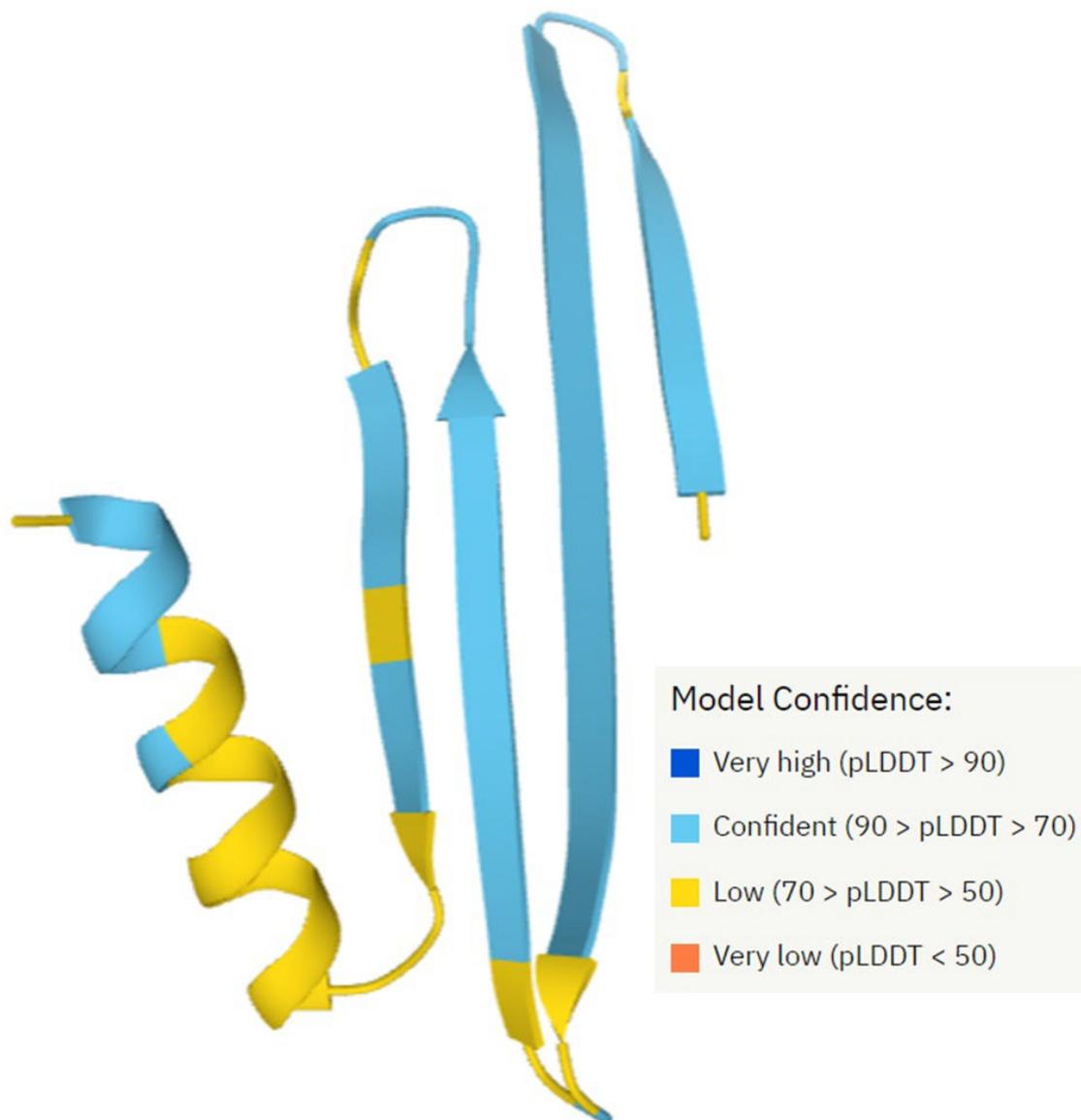
Figure 3.16: AlphaFold prediction of OEP7.3 Tertiary Structure. Taken from AlphaFold database entry for At4G02482, which is the genomic location of the short form annotation which corresponds to the protein named OEP7.3. Legend included from entry showing colour coded confidence rating. (Jumper et al., 2021).

Chapter 4: Discussion

4.1 OEP15-1 and OEP7.3 Encoding mRNAs Appear to be Naturally Present in *Arabidopsis thaliana* Rosette Leaves.

OEP7.3 is currently represented in ARAPORT11, the most recent *Arabidopsis thaliana* genome annotation which used RNA seq libraries to predict protein coding genes and loci which are transcribed to produce other RNA products, at the At4G02482 gene (Cheng et al., 2017). The presence of OEP15-1 was called into question by the annotation change (Lamesch et al., 2012). It is important then to verify OEP15-1 presence, but protein detection would have required generation of an OEP15-1 specific antibody, so detection of mRNA transcripts corresponding to the protein isoforms was focused on instead.

The first strategy used to verify OEP15-1 mRNA presence was to amplify the CDS of OEP15-1 from cDNA reverse transcribed from total RNA or *A. thaliana* rosette leaves. When primers designed to flank the CDS of OEP15-1 are used to PCR amplify a fragment from cDNA, a ~500 bp fragment is amplified (Figure 3.1 A). This product size suggests that the total RNA contains a sequence which compliments the primer set and in which the binding sites for each primer are near enough to each other to form a product. The distance between these primers' binding sites at genomic location At4G02482 would be ~500 bp, so it is likely they are binding as designed rather than off target.

One limitation of the current study was that this ~500 bp fragment was not cloned for sequencing as was originally planned. The low copy number which seems to be present of the mRNA from rosette leaves made amplifying enough product to isolate and clone difficult. In

future work concerning OEP15-1 it would be valuable sequence this product so that the intron present in Kelly's amplified product could be verified.

To lend confidence to the identity of the fragment the primer sequences (excluding the restriction enzyme sites added to the ends) used in this reaction were instead submitted to NCBI's Primer BLAST tool against the *A. thaliana* genome to further test specificity. The only product which is 100% complementary to the primer sequences is the 483 bp sequence found at At4G02482. Other possible products which could be produced by this primer pair and are 500 +/- 50 bp have at least 3 mismatched bp per primer. When the primer pair are submitted to the same tool but against the RNA data set for *A. thaliana,* only matches with at least 3 mismatched bp are generated at any product size. No RNA molecule which corresponds to At4G02482 has been added to this RNA collection yet. It was assumed the PCR conditions were adequately stringent due to the amplification of only one product, so it is unlikely that alternative products resulting from annealing of mismatched template and primers are what is being amplified.

The ~500 bp product from At4G02482, however, contains a predicted intron which contains a stop codon that is in-frame with the exons (Figure 3.2) (Lamesch et al., 2012). If this sequence was present on a mature mRNA, translation would cease before the second exon from the OEP15-1 TAIR10 annotation and the 135 AA long protein could not exist. It is possible that PCRs using these primers amplify a cDNA reverse transcribed from a pre-mRNA prior to splicing (Syed et al., 2012). If this were the case, the mature form should appear in the reaction as well as a smaller band on the gel, which was not seen. If the annotation in TAIR10 were correct, the amplified region of the mature mRNA made up of only the two predicted exons would be 405 bp in length, a size which is different enough from the 500 bp product to appear as

a distinct band on a 1% w/v agarose gel as was used in the current study, but such a band is not seen (Figure 3.1) (Lamesch et al., 2012).

Another strategy used to detect the presence of OEP15-1 mRNA was to amplify the 3' UTR region(s) downstream of At4G02482. A forward primer which complemented a sequence inside the shared C-terminus of both OEP15-1 and OEP7.3 was paired with randomly anchored oligo dTs and two products of distinct sizes were amplified (Figure 3.1 C) suggesting the presence of two mRNA molecules with the forward primer sequence in common. This is evidence in favour of both isoforms being encoded on separate mRNA molecules because while the sequences of the CDS would be the same, the UTRs would be unique in length due to alternative polyadenylation (Srivastava et al., 2018). To form both mRNAs from the same site on the DNA, alternate splicing of the pre-mRNA molecules would have to occur (Syed et al., 2012).

Alternative splicing occurs in an estimated 61% of all intron-containing genes in *A. thaliana* and can occur in untranslated regions of mRNAs (Syed et al., 2012). It is not then statistically unlikely that this could be occurring in the case of OEP15-1, nor does the presence of only two exons exclude the gene from this possibility. Alternative splicing can take many forms, the most relevant to what we see in Figure 3.1A and 3.2 being intron retention, in which an intron is simply skipped over during splicing and left in the mRNA (Syed et al., 2012). An intron retention event would explain the presence of a stop codon, termed a premature termination codon, in the resulting code. Introduction of a premature stop codon often results in translation of a microprotein or the resulting mRNA is targeted for nonsense mediated degradation (Syed et al., 2012). It is unlikely that this alternate splice product is being targeted for degradation efficiently, because the ~500 bp product which matches the expected length of OEP15-1 mRNA CDS including the intron is possible to amplify. Perhaps the mRNA which

includes only the exons of OEP15-1 is spliced 'correctly' less often than the intron retention event takes place, making this mature mRNA molecule very infrequently transcribed and difficult to detect via PCR.

Collectively, the data shown in Figure 3.1 support the theory that two mRNA molecules are transcribed from the At4G02482 gene but does not directly support the translation of OEP15-1 from either of those transcripts due to the presence of a stop codon midway through the CDS. Research was continued with OEP15-1, however, due to its previous inclusion in outer envelope proteome compilations and because the data presented do not exclude the possibility of its presence in *A. thaliana* definitively (Inoue et al., 2015).

In Figure 2.1 a collection of ESTs are included in orange. ESTs do not necessarily capture the full length of an mRNA molecule but can be used to predict where CDSs reside in the genome (Slater et al., 2008). At the gene At4g02482, the ESTs span the full length of both isoforms' predicted mRNAs (Cheng et al., 2017; Lamesch et al., 2012; Campbell et al., 2014b). The ESTs do not span as far upstream as would be expected if mRNA for OEP15-1 is made as it would then need a 5' UTR. 5' UTRs average 155 bp in *Arabidopsis,* but can be much shorter (Srivastava et al., 2018). The ESTs often include the predicted intron, which means the RNA molecule they were amplified from contains the stop codon present in this intron and therefore cannot be translated to form a full length OEP15-1 protein. One EST spans the complete CDS of OEP7.3. Another EST begins 28 bp upstream of OEP15-1 start site and ends 16 bp upstream of the predicted stop site (Alexandrov et al., 2006). The annotation for the predicted OEP7.3 mRNA 5'UTR begins downstream of the start site of OEP15-1, so this EST could be of OEP15-1 mRNA including only a short portion of the 5' UTR.

That the EST ends prematurely does not mean that an mRNA which spans all of OEP15-1 does not exist. The predicted intron is included in this EST, however, and no EST exists which has a gap in alignment corresponding to the intron. There are very few ESTs included (10) in the alignment generated by BLAST when OEP15-1 is submitted, however, so it is possible the transcription of mRNA from this gene is low and the mature transcript which would encode OEP15-1 has not yet been detected. mRNA transcripts for this gene may also be rapidly targeted for degradation again causing their abundance to be low in the cells and not detectable during creation of EST collections. The slow production of protein seen experimentally when transiently expressing recombinant versions of each of these OEPs in this work could also be explained by a low rate of transcription. If this were the case, the reason for this low rate would then have to be intrinsic to the CDS itself as the recombinant proteins are under control of a 35S promoter and do not have their native UTRs. The CDS could be a target of micro RNAs, causing decreased translation. As mentioned above, it is possible the spliceosome complex proteins which process pre-mRNAs do not have a high affinity for OEP15-1 mRNA, causing the intron removal event to be a rarer occurrence than the already low transcription of the whole mRNA (Syed et al., 2012).

4.2 At4G02482 May be a Product of Gene Duplication

Although included in Inoue's list of COM proteins (2015), OEP15-1 has not been included in more recent proteome compilations for the outer envelope of chloroplasts (Bouchnak et al., 2019). Homologues to *A. thaliana* OEP15-1 or OEP7.3 could not be found in other species. Blast protein search results included many more putative translocase-like genes which show homology to the C-terminus of TOC159. These were not considered homologues to the proteins of interest in this work as their length always far exceeded the length of OEP15-1. The

existence of so many similar genes, both uncharacterized and the other members of the protein family Toc120, Toc132 and Toc90, suggest Toc159 may have been a frequent target for gene duplication during evolution (Panchy et al., 2016; Richardson et al., 2014).

Gene duplication is a driving force for evolution in plants and contributes to their quick differentiation relative to other eukaryotic taxa (Panchy et al., 2016). A low estimate for duplicate gene content in *Arabidopsis* is 47% of the genome (Panchy et al., 2016). Most of these duplicate genes originate from whole genome duplication events, but local tandem gene duplication events occur as well causing paralogous genes on the same chromosome with few genes in between (Panchy et al., 2016). OEP7.3 and OEP15-1 share the gene location At4G02482 and the gene with the highest homology in *A. thaliana* (Figure 3.14), Toc159 is located at At4G02510 with few intervening genes in the Araport11 genome annotation (Cheng et al., 2017). It is reasonable then to postulate that At4G02482 could have arisen from an event in which an unequal crossing over event happened to include only the end of the Toc159 CDS causing tandem duplication of this short segment. This origin as a gene duplication artifact does not exclude OEP7.3 and 15-1 from performing a function in the cell, as gene duplication often begins evolutionary divergence of genes to perform separate functions. It may, however, offer an explanation for OEP7.3's small size and lack of apparent functional domains if it were an artifact of gene duplication.

The timing of this gene duplication event is difficult to determine with the data included in this work. On one hand, BLAST results show no homologous genes in other plant species which were not considered Toc159 family proteins as all were much longer than OEP15-1. This would suggest the duplication even was recent in the evolution of *A. thaliana* as no other related lineages seem to include the short gene. *A. thaliana* is one of the most thoroughly studied plant

genomes, however, so it is possible that homologues in other species have yet to be predicted or discovered experimentally. On the other hand, the higher similarity between AtOEP15-1 and BnToc159 (Figure 3.14) might suggest that the duplication event took place before these lineages diverged and then AtOEP15-1 differentiated further through evolution from the duplicated ancestral Toc159 M-domain segment than BnToc159 did.

If At4G02482 arose from a relatively recent gene duplication event it is unlikely to remain in the genome for a long time unless it does code for a non-redundant or functionally unique protein, as it is believed that duplicate genes are deleted preferentially with a bias towards deleting paralogs with reduced expression level (Panchy et al., 2016). Yeung's initial amplification of the full length CDS of OEP15-1 from cDNA included the annotated intron with a stop codon. The presence of the stop codon in an RNA product suggests At4G02482 may be the location of a pseudogenic duplicate (Figure 3.2). Some pseudogenes function as truncated proteins or as the RNA molecule itself, which may be increasing the life span of At4G02482 as a duplicate gene, or it may currently be in the process of being deleted (Panchy et al., 2016).

4.3 I-TASSER Predicts that OEP15-1 is a β-barrel Outer Envelope Protein

Further research inspired by bioinformatics results has led to a model for the function of OEP15-1 which fits with and lends context to the localization patterns of the GFP fusion proteins tested in this thesis.

Both submissions of OEP15-1 to the I-TASSER server led to tertiary structure predictions which at first glance resemble small β-barrel proteins (Figure 3.15) (Yang et al., 2015). The GO terms in Table 3.1 center on membrane association, though the membranes it predicts OEP15-1 is associated with are the plasma and cell outer membranes rather than the

chloroplast outer membrane (Yang et al., 2015). The cell outer membrane refers to the outer membrane of gram-negative bacteria which are the evolutionary ancestors of chloroplasts. The prediction of gram-negative cell outer membrane localization instead of chloroplast outer membrane localization is understandable as the lipid composition and therefore the associating protein TMDs may be similar for both outer membranes (Gould et al., 2008).

OmpW was selected as the most similar PDB entry to use for threading when I-TASSER was predicting the structure of OEP15-1 the first time it was submitted to the server in 2019 (Yang et al., 2015). OmpW belongs to a family of porins in the outer membrane of gram-negative bacteria and functions as a transporter of small hydrophobic molecules (ions) (Hong et al., 2006). Porins are a large family of outer membrane proteins in gram negative bacteria which often function as transport channels for small hydrophilic molecules and range in size averaging 16 β-strands (Lu et al., 2020). OmpW, however, only forms an 8 β-strand pore (Hong et al., 2006). Transmembrane β-strands range in length from 5-15 AAs (Uniprot, 2021 KW-1134). The secondary structure of OEP15-1 as predicted by PSIpred and JPred (Figure 3.3) contains at least 7 β-strands which are 5 AA or longer and has a region in which many β-strands of shorter length are predicted. Each secondary structure prediction tool generates secondary structures which differ slightly in length and placement (Figure 3.3, 3.5), so it is not unreasonable to assume that the region predicted here to contain many short β-strands may instead contain an 8th strand of 5 AA or longer . In fact, the secondary structure prediction in the I-TASSER output featured 8 β-strands of appropriate length upstream of the α-helix and one shorter β-strand at the C-terminus (Figure A2).

The smallest, well studied, porin-like β-barrel channel in the outer envelope of chloroplasts is OEP21, an ATP dependant, 8 β-strand anion channel (Hemmler et al., 2006;

Pottosin and Dobrovinskaya, 2015). The smaller amino acid channel OEP16 is comprised of 4 membrane spanning α-helices instead (Pudelski et al., 2012) and is not considered here. The evolutionary origin of chloroplast membranes explains the presence of any β-barrel proteins in the outer envelope so the ubiquity of small porins in gram negative bacteria suggests that a variety of small porins may be present in the chloroplast outer membrane as well (Lu et al., 2020).

When OEP15-1 was analyzed in 2021 by I-TASSER again, the program had more models to choose from in the PDB, but interestingly the top protein template was a subunit of OmpX, the NMR data for which was submitted in 2003 (Fernandez et al., 2003; Yang et al., 2015). Presumably, with the larger data set of updated PDB submissions to thread against, the program should generate more accurate results for the second submission, but it is interesting that protein entries added in the years between submissions to I-TASSER did not show up in the output. OmpX, however, is still predicted to be a β-barrel protein similar to those found in gram negative bacteria outer membranes, but in its mature conformation a single β-barrel would represent one of six subunits of the homo-hexamer (Vogt and Shultz, 1999). The barrel of each subunit is elliptical measuring 1 nm and 1.6 nm across at its shortest and longest distance respectively and the barrel forming portion of the protein is again comprised of only 8 β-strands (Vogt and Shultz, 1999). Measurements across the channel of the 2021 pore-like I-TASSER model for OEP15-1 made across the widest and narrowest angles for comparison and measured 2.078 nm and 1.259 nm, respectively (Figure 3.15).

If no assumption is made as to which model set from I-TASSER is more accurate, significance can at least be attributed to the fact that the highest scoring GO term template protein in both outputs was a small porin-like β-barrel protein in the outer envelope of gram-

negative bacteria (Yang et al., 2015). Additionally, the existence of OMP proteins with only 8 channel forming β-strands and internal diameters as low as 1.6 nm lends support to the idea that OEP15-1 could be a porin even at its particularly small size (Vogt and Shultz, 1999). Because OEP15-1 is predicted to contain enough β-strands which are of sufficient length to span the membrane and was modeled after a porin or β-barrel channel subunit by I-TASSER, I propose that OEP15-1 may function as a β-barrel protein involved in cross membrane transport on the outer envelope of chloroplasts.

OEP7.3 on the other hand may not perform a function on its own at half the size of OEP15-1. In Figure 3.16, AlphaFold predicts the formation of a β-sheet-like structure with confidence other than in determining the boundaries of the individual β-strands (Jumper et al., 2021). The placement of the β-strands in both OEP15-1 and OEP7.3 differ by a couple of amino acids depending on which secondary structure prediction server is used as well (Figure 3.3 vs 3.5).

I-TASSER was chosen as a tertiary structure prediction tool for its user friendly interface and comprehensive output as well as because by modularly threading against all PPDB entries it can be expected to remain accurate and flexible as information is added to our knowledge of protein folding. When I-TASSER appeared to have difficulty making predictions for the shorter isoform OEP7.3, AlphaFold was chosen because it was recently made available for research purposes (2021) and OEP7.3 was already present in the dataset however there is much less information given by this tool towards how the predicted structure was reached in that homologous proteins are not given to draw functional predictions from. There are trimeric autotransporters in bacteria which form 12 β-stranded pores to which each of the three subunits contribute 4 β-strands (Ulrich et al., 2014). One such protein, Yersinia Adhesin A (YadA), was

shown to be assembled by and inserted into the membrane of mitochondria when expressed in eukaryotic cells, indicating that 4 β-strands is sufficient for recognition by mitochondrial assembly machinery (Ulrich et al., 2014). The many emerging similarities between plastidic and mitochondrial assembly machinery make it likely that if similarly trimeric β-stranded pores exist in plant cells that plastids could assemble them as well (Jores et al., 2016; Gross et al., 2021). OEP7.3 is predicted to contain 4 β-strands and reside in the COM, so it is possible that it forms a multimeric protein channel.

4.4 OEP15-1 and OEP7.3 Utilize β-Strands to Target the Outer Envelope of Chloroplasts

Transient expression analysis of proteins tagged with a fluorescent domain allows visualization of protein localization in living cells (Sharma et al., 2018). In the current study two living cell systems were used to express GFP fusion proteins including a variety of truncation mutants to determine which domains contribute to localization. Onion epidermal cells were chosen as one of the cell systems due to ease of transfection by tungsten biolistic bombardment and the robust viability of the transfected cells afforded by incubating the epidermal layer while it is still in contact with the rest of the onion. The onion epidermal peels in this work did not behave as expected causing a variety of expression patterns to be seen for any given construct. Cells showed very little expression for the first 16-18 hours and quickly reached overexpression following the accumulation of enough protein to detect with the fluorescent microscope. This was especially true in those constructs with the upstream peptide addition which was an artifact of the cloning technique used (peptide described in section 3.5 and Figure A1). These recombinant proteins were aggregate-prone when left to express for too long while also being slow to begin protein production. This made the optimal screening window very small and inconsistent, likely varying slightly for each peel, or possibly, each cell.

After 20 hrs all constructs had detectable fluorescence, so this period of incubation was used for all onion epidermal expression assays. This is longer than has been allowed typically in past research of other OEPs (Grimberg, 2016; Lung and Chuong, 2012). All constructs were cloned into vectors controlled by a 35S promoter. Both the promoter and incubation time increase the likelihood of over-expression which could result in artifacts such as protein aggregation, which was seen in onion epidermal cells to some extent for all constructs (e.g. Figures 3.6 and 3.7). The inconsistency of expression described above as well as propensity for over expression contribute to the importance of observing expression patterns in another cell system before drawing any conclusions.

The second cell system used was *A. thaliana* mesophyll protoplasts. These cells were chosen because they contain chloroplasts, the plastid best studied in terms of OEPs, and it is possible the proteome differs between the COM and leucoplast outer membranes due to their being in different stages of differentiation (Taiz and Zeiger, 2006). Protoplast isolation methods for *A. thaliana* are established and the isolated cells are easy to chemically transfect due to their lack of cell wall (Yoo et al., 2007; Wu et al., 2009). The removal of the cell wall during the process of making protoplasts alters the cytoskeleton but because the cells are otherwise intact, they still offer a system which should be very close to the cellular environment of cells which are still a part of a plant tissue (Gross et al., 2021). Unlike the onion epidermal cells, protoplasts used in this work showed consistent expression patterns both across replicates and per transfected batch of cells.

The following subsections will summarize which truncation constructs localized to chloroplasts and what that reveals about possible targeting pathways used by OEP7.3 and OEP15-1.

4.4.1 OEP15-1 and OEP7.3 are Localized to the Outer Envelope Membrane of Chloroplasts

Both OEP15-1 and OEP7.3 were cloned into vectors which would produce fusion proteins with GFP on either terminus. It is important to use both vectors because GFP is very large in comparison to either protein of interest and could therefore theoretically block interactions between the terminus nearest GFP and cellular components which may be necessary for targeting (Lung and Chuong, 2012).

OEP15-1 is shown to successfully localize to chloroplasts of *A. thaliana* protoplasts when GFP is attached at either terminus (Figure 3.8). One of two GFP tagged OEP7.3 proteins successfully targets chloroplasts, although the Western blot and confocal imaging data do not agree on which terminus GFP must be fused to in order to allow targeting. The Western blots (Figure 3.12) suggest that OEP7.3:GFP was able to localize due to the full-length fusion protein being detected in the crude chloroplast lane only, while the microscopy data (Figure 3.8) suggests GFP:OEP7.3 was localized at chloroplasts. In the case of both OEP15-1 fusion proteins and GFP:OEP7.3, localization to the COM appears as a pattern resembling a halo of signal around each chloroplast which is the expected pattern when visualizing OEPs as seen by Lung and Chuong (2012). While the placement of GFP making the difference to targeting effectiveness in OEP7.3 is a detail which needs further investigation, these results alone suggest that both endogenous OEP15-1 and OEP7.3 localize to the chloroplast membrane.

Figure 4.1 is included in this section both as a summary of which recombinant proteins appear to localize to the outer membrane based on fluorescent microscopy (Figures 3.8-3.10), as well as to place the included peptides in the context of the secondary structures predicted for OEP7.3. Combinations of these secondary structures appear to determine localization so it is important to note which β-strands and α-helices are included in each construct. Outer membrane

designation was given to recombinant proteins which appeared as a halo of signal around the chloroplast in protoplasts transient expression and stroma designation was given to those which had signal inside the chloroplasts. These designations were made for ease of understanding though it is not certain from visual data alone that these are the sub compartments to which the recombinant proteins are being localized. A halo of fluorescence may appear if the inner membrane were targeted and stroma and thylakoid localization would be difficult to differentiate. A gradient of green to red was used when the localizations suggested by the protoplast microscopy did not agree with either the Western blot or the onion epidermal cell microscopy.

4.4.2 The C-terminus of OEP15-1 and OEP7.3 is Not Necessary for Targeting Plastids

Collectively, the data indicated that it is unlikely that OEP15-1 and OEP7.3 use the C-terminal Transit Peptide targeting pathway used by Toc159 described by Lung and Chuong (2012). Importantly, the construct which includes all parts of OEP15-1 except the C-terminus is shown to successfully localize to the outer membrane, though it may also localize to the stroma (Figure 3.9). This construct is missing the alpha helix at the C-terminus as well as the two β-strands directly upstream of it. This missing region is comprised of the 45 AA corresponding to the predicted TP length (44 AA) predicted by Chloro-P in Grimberg's work (2016). Constructs can be seen in context with the secondary structures of OEP7.3 in summary Figure 4.1 A. When
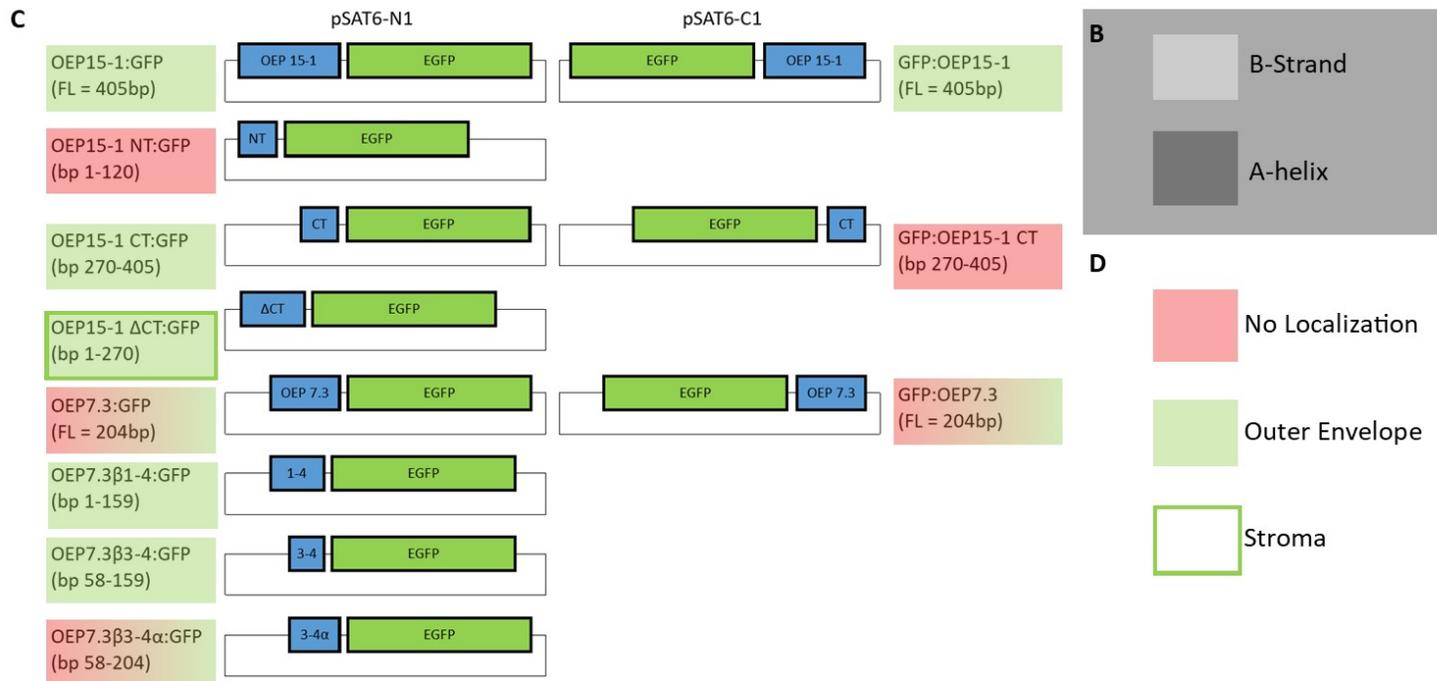
Figure 4.1

Please see legend on next page

Figure 4.1: Localization Summary Figure. A. All protein truncations are aligned, and the four β-strands and single α-helix predicted for OEP7.3 by PSIpred are highlighted in shades of grey (Legend B). C. Recombinant proteins depicted again, colours behind and surrounding the clone names correspond to expression pattern seen in protoplasts transient expression (Legend in D.)

the pair of constructs that included only the C-terminal 45 AA fused to GFP on either terminus were expressed in protoplasts, only the construct with GFP attached to the C-terminus was able to localize to chloroplasts (Figures 3.9 and 3.12). If the final 44 AA comprised a TP-like signal I would expect the exposure of the C-terminus allowed by fusion of GFP to the N-terminus to be more favourable to localizing than a C-terminus potentially interfered with by GFP. In Lung and Chuong's work on Toc159 from which the cTP model originated, it was shown that a construct's ability to localize was dependent on the C-terminal TP orientation compared to the GFP it was fused to so that if the extreme C-terminal end of the protein was not furthest from GFP, the fusion protein could no longer localize (Lung and Chuong, 2012). This was consistent with their findings when they reversed the orientation of N-terminal TPs from various stroma targeting proteins causing inability to target the chloroplasts as well (Lung and Chuong, 2012). The inability then for the C-terminus of OEP15-1 to direct localization of GFP when the α-helix is on the C-terminal end of the fusion protein suggests that the C-terminus does not function as a TP. That this protein does not localize does not exclude a TP as a possibility on its own because there is always the possibility that unexpected localization is a by-product of fusion protein creation itself.

In the OEP15-1CT:GFP construct, the exposed structures are two β-strands at the N-terminus of this peptide. That this construct localizes to the COM suggests that the β-strands have a role in targeting instead (Figure 3.12). Additionally, in Figure 3.10 the addition of the alpha helix to the final two and a half β-strands inhibits the localization of the construct compared to the final β-strands alone rather than increasing the targeting efficiency as would be expected if the α-helix were involved in targeting.

4.4.3 A Combination of Secondary Structure Motifs Upstream of the C-terminus Provide

Targeting Information

In the context of TPs as mentioned in the introduction, Lee and colleagues (2015) hypothesize that each TOC protein and OEP which interact with pre-proteins prior to or as a part of their translocation recognizes more than one motif. In the current study, it appears that more than one secondary structural element or motif is involved with the targeting of OEP15-1 and OEP7.3 as the many constructs which appear to target the outer envelope do not have an overlapping predicted secondary structure motif included in the complete group (Figures 3.8-3.10, 3.12 and 4.1).

The possibility that OEP15-1 is a small β-barrel protein (Figure 4.2) yields some insight into how the multiple targeting motifs may work together to promote targeting to the outer envelope of the chloroplast. Recently, some β-barrel OEPs were shown to translocate across the OM via the TOC complex then re-insert into the OM from the IMS with the aid of TOC75-V/OEP80 (Gross et al., 2021). Gross and colleagues found that in the case of OEP37 and OEP24, which are 18 and 14 β-stranded proteins, respectively, the first 6-8 β-strands are sufficient to target the fusion protein to chloroplasts (and leucoplasts), but both proteins aggregate in the IMS if the most C-terminal β-strands are absent (2021). OEP15-1 has fewer β-strands than OEP37 and OEP24 in total, but the two most C-terminal β-strands seem to prevent expression in parts of the chloroplast other than the outer envelope as see with OEP15-1ΔCT:GFP (Figure 3.9). As summarized in Figure 4.1 it appears that generally, constructs containing the first two β-strands of OEP7.3 (corresponding to the 6th and 7th strands of OEP15-1) are targeted to chloroplasts but are expressed inside the chloroplast as well if the final two β-strands are not present. The final two β-strands of OEP15-1 and OEP7.3 however appear to be sufficient for targeting to the

chloroplast outer envelope on their own as demonstrated by clone OEP7.2β3-4:GFP (Figure 3.10). Gross et al (2021) did not test whether the final β-hairpin (pair of β-strands) of either OEP24 or OEP37 could direct a fusion tag to the chloroplast on its own; the most C-terminal fragment tested included β-strands 9-14 of OEP24, which localized more efficiently compared to strands 7-14. Amino acid changes in the terminal β strand were also shown to inhibit interaction with TOC75-V/OEP80, which Gross and colleagues propose is responsible for β-barrel protein insertion into the outer membrane from the IMS (2021). It seems likely from Gross and colleagues' work that the final two β-strands of some β-barrel OEPs could be the site recognized by OEP80 and therefore the fragment necessary for insertion into the outer membrane, while some combination of upstream β-strands in the barrel improve targeting efficiency to the chloroplast and may be recognized by Toc159 for translocation. The localization patterns in Figures 3.8-3.10 support this theory as constructs containing all 4 β-strands localize neatly around the outside of the chloroplasts. Meanwhile the only construct which contained the first two β-strands of OEP7.3 but not the final two, OEP15-1ΔCT:GFP appears to localize throughout the chloroplast in irregular shapes which could represent the aggregates Gross and colleagues found in the IMS or could suggest that if the β-barrel in the IMS is not recognized for insertion by OEP80 it may continue translocation and end up in the stroma (2021). Because OEP7.3 also targets the outer envelope of chloroplasts, the β-strands which are responsible for the initial targeting to the outer membrane and are possibly recognized by Toc159 must be contained in OEP7.3 as well and be the first two β-strands of this small protein. OEP7.3 is then likely also recognized for insertion by its final two β-strands. Importantly, Gross and colleagues also propose that an aromatic amino acid followed by a negatively charged amino acid is a necessary

motif for TOC75-V recognition which must be present in the final β-strand. OEP7.3 does not meet this requirement.

In mitochondria, a β-hairpin with many hydrophobic residues on one of the two comprising β-strands serves as a targeting signal for β-barrel proteins (Jores et al., 2016). There has been experimental evidence that plastid targeted β-barrel proteins may contain a similar signal and that their specificity to the chloroplast rather than the mitochondria is dictated by a hydrophilic N-terminus of the second last β-strand and a very hydrophobic final hairpin turn (Jores et al., 2016; Klinger et al., 2019). The second last β-strand in OEP15-1 and OEP7.3 is very short as predicted by PSIpred, but does have the AA series RQTK (R, Q and K being hydrophilic residues) in the second last turn leading into the N-terminus of the second last strand, and VSL (V and L being hydrophobic residues) leading from that second last strand into the final turn (Pommié et al., 2004). The final hairpin turn itself however is mainly comprised of hydrophilic residues and the second to last β-strand contains mostly neutral residues in its centre. While the residues which do not fit the pattern established by Klinger and colleagues (2019) outnumber the residues which do, it is worth mentioning again that each secondary structure prediction yields a slightly different output and the β-strands in these proteins are short, so small differences in the prediction can result in noticeably better (or worse) agreement with the pattern identified by Klinger et al (2019). Three secondary prediction tools were used (JPred, PSIpred, and I-TASSER's secondary structure output) and while the β-strand predictions differed enough that the hydrophobic stretch described sometimes fell in the turn rather than the C-terminus of the strand itself, in all three cases a hydrophilic sequence of DSKR also occurs in the final β-hairpin turn, so this turn does not meet the proposed requirement of being very hydrophobic.

The β-strand targeting and insertion signal model does not fully explain the data presented in the current study, however, as the last two β-strands when expressed in combination with the C-terminal α-helix appears to interfere with targeting to the outer membrane of chloroplasts. Even if the C-terminal α-helix were a TP, when in this orientation next to GFP it would not be expected to direct targeting, so an explanation is only needed for its interference (Lung and Chuong, 2012). In Figure 3.10, OEP7.3β3-4α:GFP appears mainly cytosolic or perhaps localized to the ER or cytoskeleton due to the web-like appearance of the majority of the signal (Razzak et al., 2019). Organ-specific markers for peroxisomes or mitochondria which would appear cyan or yellow and could then be co-expressed with this construct would be an effective way to verify if this recombinant protein is being miss-targeted to a specific organelle. It is difficult to determine if there are low levels of OEP7.3β3-4α:GFP targeting to the outer envelope using microscopy as there is some signal clinging to the outside of a few chloroplasts. This is a good reason to use the data collected in onion cells (Figure 3.10 vs 3.7) as it also showed colocalization of OEP7.3β3-4α:GFP to only some of the DsRed decorated plastids. In combination it would appear there is some inefficient targeting to the outer membrane of chloroplasts in this construct. Although the consistently faint signal in the chloroplasts observed using confocal microscopy with this fusion protein suggest the possibility that some GFP is present inside the chloroplasts, the fuzzy boundaries of these signals suggest this signal is more likely bleed-over from the autofluorescence of the plastids than stromal targeting.

Why the α-helix sometimes appears to interfere with OEP15-1 targeting to the outer envelope of chloroplasts is not obvious, but one pattern can be seen which may provide some insight. The α-helix at the C-terminus of OEP15-1 only seems to interfere with targeting in cases in which it is tagged with GFP on its C-terminal end and the total protein of interest is 66 AA or

fewer. The excess of aggregates and inconsistent expression patterns observed for OEP7.3β3-4α:GFP suggest an error in folding. It could be that the proximity of the C-terminal helix and the 238 AA EGFP adopt a conformation which causes misfolding of the fused peptide, and that particularly small peptides are then sometimes inaccessible for recognition by localization machinery in the cell. In Figure 3.9, the CT alone of OEP15-1 with GFP adjacent to the α-helix targets the fusion to chloroplasts. This exception could be because the second to last strand which appears to be important to localization is directly terminal as opposed to in the very similar construct OEP7.3β3-4α:GFP where the strand is a few residues in from the end. It appears however that full length OEP15-1 was able to fold correctly and direct localization of GFP on either end even with the adjacent α-helix (Figure 3.8). The α-helix does not interfere with targeting when GFP is attached at the N-terminus of whichever protein fragment has been included of the OEPs, so it may still perform some yet to define function when allowed to fold correctly.

The Western blot for OEP7.3:GFP (Figure 3.12) shows a banding pattern similar to those constructs which localized to chloroplasts consisting of a full length fusion sized band in the chloroplast fraction and a single lower molecular weight (degraded) product in the cytosolic fraction. This suggests that there is poor enough localization that it was hard to visualize by microscopy but that some protein does localize to the chloroplasts, allowing it to be protected from degradation. By contrast, when GFP:OEP7.3 is viewed under microscopy in either cell system it appears to localize however the Western blot shows only one product at an unexpectedly low molecular weight. One limitation of fluorescence microscopy is that it is hard to visually differentiate between localizing and background signals. In a Western blot the data is also limited, this time by the possibility of proteins migrating at unexpected sizes (and by the

effectiveness of the fractionation). For that reason, it is not certain weather the single band seen in the Western blot of fractionated protoplasts for GFP:OEP7.3 (Figure 3.12) is the recombinant product or the degraded product seen in the total protein fraction (Figure 3.11). If OEP7.3 fusion proteins are excluded due to inconsistencies between the microscopy and Western blot data from the pattern described above, misfolding may only influence OEP7.3 fragments that are less than 66AA in length.

4.4.4. Cytosolic Peptidases may Interfere with OEP15-1 and OEP7.3 Localization, Decreasing Targeting Efficiency

In Figures 3.11 and 3.12 proteins of two different sizes were detected in Western Blots performed on extracts from protoplasts expressing most recombinant GFP proteins which appeared to localize to the outer membrane of chloroplasts (GFP:OEP15-1, OEP15-1CT:GFP and OEP15-1ΔCT:GFP). The smaller of the two bands was consistently the size of GFP or slightly larger, and this band was not present in the chloroplast fraction which consistently contained the product of the expected recombinant size. This provides independent evidence that those fusion proteins are located at the chloroplast, and also suggests that the full fusion proteins are being degraded or cleaved in the cytosol but are protected once they reach the chloroplast. That there is only one size of degraded product would be consistent with the degradation by peptidases specifically affecting the OEP15-1/OEP7.3 portion of the fusion protein, rather than being the target of general proteosome degradation. The preserved part of the recombinant proteins must correspond to GFP because the bands are detectable. It is likely this cleaved GFP is responsible for the cytosolic signal seen during microscopy (Figures 3.6 to 3.10) as this smaller band is the only band present in these cytosolic fraction lanes (Figure 3.12).

The specific site of peptidase cleavage is difficult to predict based on the size of apparent GFP fragment(s) on the Western blots because it appears that only a small fragment of the OEP protein remains attached in each case. In Figure 3.12, when GFP is fused to the C-terminus, the smaller band size appears to be almost exactly 27kDa, suggesting the linking peptide in between all the OEP fragments and GFP could be what is being cleaved. In Figure 3.11 however, there does not appear to be the same pattern of size and GFP terminus. There are fewer constructs with GFP fused to the N-terminus, but these more consistently appear to retain some of the OEP protein fragment in this degraded product (have an apparent weight clearly higher than GFP alone). That these degraded products are of a similar size would mean that the cleavage site could not be the same sequence in each, as all 3 constructs with GFP on the N-terminus include a different residue of OEP15-1 as the beginning of the peptide of interest. I suspect due to the single degraded product in each case that a peptidase may be cleaving the proteins a set distance away from GFP itself but determining what is cleaving the recombinant proteins and the specific site of cleavage does not seem critical for the exploration of OEP15-1 and OEP7.3 targeting at this point because the full-length recombinant proteins appear to be able to localize some of the time before being degraded (Figure 3.12).

In fractionated cytosolic vs chloroplast fractions, sometimes one of the two bands detected in the corresponding total protein fraction are not present. It is possible that protein degradation in the cytosol happens more often when the cell system is saturated. If the smaller of the two products is the missing band such as in OEP7.3β3-4α:GFP this may be because the protoplasts had not over-expressed the fusion protein yet, so localization was more complete/ efficient due to the chloroplast envelopes not being fully saturated.

For GFP:OEP15-1CT, the total protein fraction in Figure 3.11 displays two band sized like the other constructs do, however in Figure 3.12 the recombinant protein-sized band appears in both fractions and the possibly degraded product band of the smaller size appears only in the chloroplast fraction with a smaller yet band which was not seen on the total protein blot appearing in the cytosolic fraction. These were crude protein fractions and as stated in Results Section 3.7, aggregated proteins and other organelles are expected to be collected as part of the crude chloroplast pellet. Looking at the corresponding expression pattern in protoplasts (Figure 3.9), GFP signal is observed as a diffuse signal throughout the cytosol and in bright focussed points of expression which could represent localization to a second organelle or aggregated protein, so full-sized bands being present in both soluble and pellet fractions on the Western blot is consistent with the confocal images. It is interesting though that this GFP with partially digested OEP truncation band is only present in the chloroplast fraction as this means it is likely what is being visualized as punctate signals using confocal microscopy. It is possible the remaining protein after partial digestion is even more aggregate prone than the full recombinant protein in this case, causing none to remain soluble in the cytosol. Overexpression of this particular recombinant protein may be so great in these cells that a further degraded product is detectable which is not detectable in cells with less total expression. In addition, the Western blot for GFP:OEP7.3 suggests that cleaved product is so plentiful that it likely forms aggregates before a detectable amount can reach chloroplast (Figure 3.12).

Only one band is detectable on Westerns of factions from OEP15-1NT:GFP cells, which corresponds to the expected recombinant protein size. This could be again because aggregation (or possibly mis-localization) protects the construct from degradation, or because it is small enough to avoid degradation by proteases. The latter is more likely because when overly-high

expression occurs, a faint band can be detected in the cytosolic fraction as well, but it is not at a lower molecular weight suggesting it is also the intact (full-length) recombinant protein (Figure 3.12).

To truly test localization to the membranes of the chloroplast there are protocols to isolate chloroplast stroma, envelope membranes and thylakoids using density gradient centrifugation (Smith et al., 2002b). Importantly, the crude chloroplast fraction would first need to be further purified to remove small organelles and aggregate proteins, using a density gradient. The lysis of protoplasts in this work did not spare the chloroplasts from lysis in most cases despite optimization attempts on the syringe apparatus itself. When fractionating the cells further, it would be important to first optimize the lysis step so that the stroma could be collected separately from the cytosol. It would be technically challenging but rewarding to isolate and keep each organelle as separate fractions to run on Western blots to clarify if the irregular sized signals correspond to any organelle. Designing fluorescent fusion constructs to target each organelle would be a less precise but more efficient way to draw the same conclusions.

4.4.5 Chloroplast Outer Membrane Targeting motifs differ from those necessary for Targeting to Etioplasts.

Working in onion epidermal cells as well as *A. thaliana* mesophyll cells allows evaluation of targeting in two different genus, plastid types and cell types. Any differences in expression pattern could be due to differing proteomes between plastid types or differing cytosolic machinery in epidermal vs mesophyll cells. If the proteome differs slightly in either case, this could explain small differences in recombinant proteins' abilities to localize as whichever proteins recognize OEP15-1 and OEP7.3 for localization may then also differ.

It could not be determined visually if the overlapping signals seen when working in onion epidermal cells signify targeting to the etioplast membranes or stroma due to the etioplasts small size and microscopes available. The discrepancies in which recombinant proteins did or did not target plastids as a whole in either cell system however, is worth noting (Figures 3.6 and 3.7 compared to Figures 3.8-3.10). The majority of recombinant proteins which localized in one cell expression system localized in the other as well. Exceptions included OEP7.3:GFP which only appeared to localize in onions and OEP15-1ΔCT:GFP, which consistently localized to the chloroplasts in protoplasts but had poor localization to etioplasts in onion epidermal cells.

4.5 Future Directions

One of the questions this research fails to address fully is the existence of both OEP15-1 and OEP7.3 in the natural *A. thaliana* plant, and by extension where each occurs in the plant. Initially, an experiment to test differential expression on a tissue scale was planned in which the 3'UTR PCR shown in Figure 3.1 and discussed briefly in section 4.1 would be optimized to the point that the amplified UTRs could be sequenced. These 3' UTRs would likely be unique at least in length but could also have been subject to alternative splicing (Srivastava et al., 2018). With that sequence data, primers specific to each could be designed either to compliment unique sequences if there were any sequences not included in the shorter of the two, or to compliment the extreme 3'end and near the stop site of the CDS. These could be paired with the forward primer which compliments the start site of OEP15-1. The primer or primer(s) which successfully amplified a product as expected would show which UTR belonged to the OEP15-1 transcript. From this point, the size of the product each UTR primer amplified when paired with the forward primer at the start site of OEP7.3 would be recorded, and the band sizes would be used to determine which transcript was being amplified from cDNA reverse transcribed from the total

RNA of various tissues. If only the length differed, determining both mRNAs were present would require band intensity comparison or qPCR, with a higher intensity or faster replication of the short product band suggesting presence of both while an equal intensity or speed would suggest that only the mRNA which contains the longer 3'UTR was present in the tissue. If the UTRs contained unique sequences, the intensity of the bands could also be compared to give a rough estimate of amounts of each transcript present in each tissue. This experiment was not completed due to failure to obtain robust and consistent bands in the initial PCR in a timely enough manner and when weighed against the priority of pursuing the localization study, it was set aside. It would still be an interesting experiment to carry out and would become valuable if the function of this protein is ever studied.

Another follow up that could be very informative is a gel shift assay on proteins which complex with OEP15-1 and OEP7.3 when each are used as bait proteins in a pull-down column. This would contribute to answering if OEP15-1 is part of a complex as it passes through the outer membrane and in the IMS as was found for the β-barrel OEPs used in Gross and colleagues' work (2021). Gross and colleagues (2021) found that β-barrel OEPs would form complexes with Toc159, which is a protein the Chuong lab has already raised an antibody against (Gross et al., 2021; Lung and Chuong, 2012). The same assay could be used to determine if the localized protein is a monomer or a subunit of a larger β-barrel though an antibody to OEP15-1 would have to be raised.

A polyclonal antibody to OEP15-1 could also be used to answer definitively how many proteins are coded by At4G02482. The antibody could potentially bind to OEP15-1, OEP7.3 as well as any microprotein created by alternate splicing intron retention or if the gene is a pseudogene as long as some of the recognized antigen peptides are encoded by the first exon of

OEP15-1. If a microprotein does exist, this would lend support to the idea that the 500 bp

product in Figure 3.1A is not an intended mRNA but a product of alternate splicing. Reading into

data generated by Western blot should be done with caution however as band sizes at lower

molecular weights than expected could be degraded protein products again instead of a

microprotein.

In Gross et al, (2021), the localization assays were performed using a bait and prey

version of GFP in which the recombinant protein of interest only carries the final β-strand of

GFP and is co-expressed with the rest of the GFP molecule (GFP_{1-10}) fused to signaling

information from single localization proteins which reside in the desired organelle (e.g. IMS

localized protein MGD1 or on its own for cytosolic localization.

The localization experiments in this thesis could be improved upon by using this strategy

for visualization in a few different ways. Firstly, having the bulk of GFP localize separately to

the chloroplast IMS would make it unnecessary to consider whether the size difference between

GFP and OEP7.3 is causing mis-localization or if the nearest terminal is being made inaccessible

to localization machinery by GFP's proximity. Additionally, fluorescence patterns would be

made more informative because in order to fluoresce, the recombinant protein must have entered

the space which the remainder of the GFP was localized to. In the case of the IMS localized

GFP_{1-10}, a ring of signal around the chloroplasts would be less likely to represent close

association with the membrane and would instead suggest that the recombinant protein of

interest passed through the IMS on its way to either chloroplast membrane or may still be present

in the IMS. The concern about GFP interfering with OEP localization is made credible by the

higher localization rates seen for OEP24 and OEP37 truncations co-expressed with GFP_{1-10}

localized to the IMS than when co-expressed with cytosolic GFP$_{1-10}$ as the later would have to pick up the bulky part of GFP prior to translocation across the outer membrane.

Utilization of this bait and prey GFP system for fluorescence opens up many other possibilities as well because GFP$_{1-10}$ could theoretically be directed to any organelle. Some of the stranger protoplast localization patterns such as for construct OEP7.3β3-4α:GFP (Figure 3.10) could be looked at again using this system and colocalizing GFP$_{1-10}$ to the IMS, ER lumen, cytoskeleton etc. to determine where the truncation localizes. This would be more effective than co-expressing the truncated protein fused to full length GFP with a protein known to target each organelle fused to DsRed because two signal co-localizations rely on the human eye to decide if the signals are near enough to signify co-localization while the strategy in Gross et al., (2021) is dependent on the co-expressed proteins meeting in the same compartment to fluoresce at all.

Due to the agreement between the predicted tertiary structure and function of OEP15-1 by I-TASSER and the truncation construct regions which appear to be contributing to localization, OEP15-1 is hypothesized here to be a β-barrel protein involved in transport across the COM. It is important to note that there is no wet lab data to support this hypothesis at this time, and more experimentation is needed to support this conclusion. Ideally, protein crystallography would be performed to confirm tertiary structure. Because OEP15-1 has no homology in other plants or known function it may be of lower priority than better characterized proteins. A strategy for confirming tertiary structure I can propose here which would be more feasible is circular dichroism (CD) of full-length protein to compare to OMP protein CD results and possible CD of truncations designed in the context of predicted secondary structures to confirm the location of β-strands as a first step towards legitimizing the structure predictions by I-TASSER.

4.6 Concluding Remarks

OEP7.3 and OEP15-1 mRNAs appear to exist in *A. thaliana* rosette leaf tissue but the existence of an mRNA which includes the two exons of OEP15-1 but not the intron is still unclear. This means the existence of OEP15-1 protein is also uncertain. If both proteins are coded for in *A. thaliana* their mRNA transcripts must be the result of alternative splicing.

The gene At4G02482 which is predicted to encode these two OEPs is likely a product of gene duplication from an ancestral Toc159 homolog, and if this duplication was recent in evolutionary time it is possible the gene is a pseudogene and the encoded protein(s) do not perform unique functions.

If OEP15-1 does perform a unique function, tertiary structure prediction tells us it may be involved in transmembrane transport as a β-barrel protein at the COM, which would lend context to the truncation fusion protein localization patterns observed in the current study.

Localization patterns for fusion proteins suggest that the C-terminus of OEPs 7.3 and 15-1 does not contain a transit peptide like signal as in Toc159, but rather uses a combination of β-strands to target and insert into the COM. Similar to the localization pathway of β-barrel OEPs 37 and 24 proposed in Gross et al (2021), OEPs 15-1 and 7.3 appear to rely on the second to last β-strand to limit localization to the COM rather than allow translocation into the stroma, and upstream β-strands to target to the chloroplast efficiently. The hydrophobicity of the final β-hairpin of OEP7.3 is consistent with targeting signals observed in mitochondria (Jores et al., 2016) but the final hairpin turn does not match the hydrophobicity pattern of β-barrel OEPs analyzed in Klinger et al (2019).

OEP15-1 is hypothesized here to be a β-barrel protein which uses a β-hairpin signal to localize to the COM but more work is necessary to confirm this working hypothesis and to provide insight into the possible function of OEP7.3. The localization patterns observed in this work provide a possibly unique targeting strategy similar to that observed in mitochondria OMPs and a few larger β-barrel OEPs in which a combination of a C-terminal β-hairpin signal and immediately upstream β-strands confer targeting specificity to the COM

References

Alexandrov, N. N., Troukhan, M. E., Brover, V. V., Tatarinova, T., Flavell, R. B., & Feldmann, K. A. (2006). Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. Plant molecular biology, 60(1), 69–85. https://doi.org/10.1007/s11103-005-2564-9

Bauer, J., Chen, K., Hiltbunner, A., Wehrli, E., Eugster, M., Schnell, D., & Kessler, F. (2000). The major protein import receptor of plastids is essential for chloroplast biogenesis. Nature, 403(6766), 203–207. https://doi.org/10.1038/35003214

Biel, K., & Fomina, I. (2015). Benson-Bassham-Calvin cycle contribution to the organic life on our planet(Report). 53(2), 161–167.

Bölter, B., & Soll, J. (2016). Once upon a time–chloroplast protein import research from infancy to future challenges. Molecular plant, 9(6), 798-812.

Bruce, B. (2000). Review of Chloroplast transit peptides: structure, function and evolution. Trends in Cell Biology, 10(10), 440–447.

Bruce, B. (2001). The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. Biochimica et Biophysica Acta, 1541(1-2), 2–21.

Bouchnak, I., Brugière, S., Moyet, L., Le Gall, S., Salvi, D., Kuntz, M., Tardif, M., & Rolland, N. (2019). Unraveling hidden components of the chloroplast envelope proteome: opportunities and limits of better MS sensitivity[S]. Molecular & Cellular Proteomics, 18(7), 1285–1306. https://doi.org/10.1074/mcp.RA118.000988

Buchan, D.W.A., & Jones, D.T. (2019). The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Research. https://doi.org/10.1093/nar/gkz297

Calvin, M., & Bassham, J. (1955). The photosynthetic cycle. Retrieved from https://escholarship.org/uc/item/0nd0h0nd

Campbell, J., Hoang, T., Jelokhani-Niaraki, M., & Smith, M. (2014a). Folding and self-association of atTic20 in lipid membranes: implications for understanding protein transport across the inner envelope membrane of chloroplasts. BMC Biochemistry, 15(1), 29.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S. H., Childs, K. L., Sun, Y., Jiang, N., & Yandell, M. (2014b). MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant physiology, 164(2), 513–524. https://doi.org/10.1104/pp.113.230144

Chen, K., Chen, X., & Schnell, D. (2000). Initial binding of preproteins involving the Toc159 receptor can be bypassed during protein import into chloroplasts. Plant Physiology (Bethesda), 122(3), 813–822. https://doi.org/10.1104/pp.122.3.813

Chen, X., Smith, M. D., Fitzpatrick, L., & Schnell, D. J. (2002). In vivo analysis of the role of atTic20 in protein import into chloroplasts. The Plant Cell, 14(3), 641–654. https://doi.org/10.1105/tpc.010336

Cheng, C., Krishnakumar, V., Chan, A. P., Thibaud-Nissen, F., Schobel, S., & Town, C. D. (2017). Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. The Plant Journal: for Cell and Molecular Biology, 89(4), 789–804.

Chomczynski, P., & Sacchi, N. (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Analytical Biochemistry, 162(1), 156–159. https://doi.org/10.1016/0003-2697(87)90021-2

Conchillo-Solé, O., de Groot, N. S., Avilés, F. X., Vendrell, J., Daura, X., & Ventura, S. (2007). AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics, 8(1), 65–65. https://doi.org/10.1186/1471-2105-8-65

Dhanoa, P. K., Richardson, L. G., Smith, M. D., Gidda, S. K., Henderson, M. P., Andrews, D. W., & Mullen, R. T. (2010). Distinct pathways mediate the sorting of tail-anchored proteins to the plastid outer envelope. PLoS One, 5(4), e10098.

Drozdetskiy, A., Cole, C., Procter, J., & Barton, G. J. (2015). JPred4: a protein secondary structure prediction server. Nucleic acids research, 43(W1), W389-W394.

Dutta, S., Teresinski, H., & Smith, M. (2014). A split-ubiquitin yeast two-hybrid screen to examine the substrate specificity of atToc159 and atToc132, Two *Arabidopsis* chloroplast preprotein import receptors.(Report). PLoS ONE, 9(4), e95026.

Edelmann, H. G. (2018). Graviperception in maize plants: is amyloplast sedimentation a red herring?. Protoplasma, 255(6), 1877-1881.

Fernandez, C., Adeishvili, K., & Wuthrich, K. (2001). Transverse relaxation-optimized NMR spectroscopy with the outer membrane protein OmpX in dihexanoyl phosphatidylcholine micelles. Proceedings of the National Academy of Sciences - PNAS, 98(5), 2358–2363. https://doi.org/10.1073/pnas.051629298

Freeman, S., Herron, J. C., & Payton, M. (2004). Evolutionary analysis (3rd ed.). Pearson/Prentice Hall.

Ganesan, I., & Theg, S. (2019). Structural considerations of folded protein import through the chloroplast TOC/TIC translocons. FEBS Letters, 593(6), 565–572.

Garg, S., & Gould, S. (2016). The role of charge in protein targeting evolution. Trends in Cell Biology, 26(12), 894–905.

Gould, S., Garg, S., & Martin, W. (2016). Bacterial vesicle secretion and the evolutionary origin of the eukaryotic endomembrane system. Trends in Microbiology, 24(7), 525–534.

Gould, S. B., Waller, R. F., & McFadden, G. I. (2008). Plastid evolution. Annu. Rev. Plant Biol., 59, 491-517.

Grimberg, N. (2016). Characterizing an alternative chloroplast outer membrane targeting signal in *Arabidopsis thaliana* (Unpublished master's thesis). Wilfrid Laurier University, Waterloo, ON, Canada

Gross, L. E., Spies, N., Simm, S., & Schleiff, E. (2020). Toc75-V/OEP80 is processed during translocation into chloroplasts, and the membrane-embedded form exposes its POTRA domain to the intermembrane space. FEBS Open Bio, 10(3), 444–454. https://doi.org/10.1002/2211-5463.12791

Gross, L. E., Klinger, A., Spies, N., Ernst, T., Flinner, N., Simm, S., Ladig, R., Bodensohn, U., & Schleiff, E. (2021). Insertion of plastidic β-barrel proteins into the outer envelopes of plastids involves an intermembrane space intermediate formed with Toc75-V/OEP80. The Plant Cell, 33(5), 1657–1681. https://doi.org/10.1093/plcell/koab052

Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouzé, P., & Brunak, S. (1996). Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. Nucleic Acids Research, 24(17), 3439–3452. https://doi.org/10.1093/nar/24.17.3439

Hemmler, R., Becker, T., Schleiff, E., Bölter, B., Stahl, T., Soll, J., Götze, T. A., Braams, S., & Wagner, R. (2006). Molecular properties of Oep21, an ATP-regulated anion-selective solute channel from the outer chloroplast membrane. The Journal of Biological Chemistry, 281(17), 12020–12029. https://doi.org/10.1074/jbc.M513586200

Hofmann, N. R., & Theg, S. M. (2005). Chloroplast outer membrane protein targeting and insertion. Trends in plant science, 10(9), 450-457.

Holbrook, K., Subramanian, C., Chotewutmontri, P., Reddick, L., Wright, S., Zhang, H., … Bruce, B. (2016). Functional analysis of semi-conserved transit peptide motifs and mechanistic implications in precursor targeting and recognition. Molecular Plant, 9(9), 1286–1301.

Hong, H., Patel, D. R., Tamm, L. K., & van den Berg, B. (2006). The outer membrane protein OmpW forms an eight-stranded β-Barrel with a hydrophobic channel. The Journal of Biological Chemistry, 281(11), 7568–7577. https://doi.org/10.1074/jbc.M512365200

Hua, X., Zhou, H., Jiang, Y., Feng,Y., Chen, Q., Ruan, Z., & Yu, Y. (2012). Genome sequences of two multidrug-resistant *Acinetobacter baumannii* strains isolated from a patient before and after treatment with tigecycline. J Bacteriol. 194(24):6979-80. doi: 10.1128/JB.01887-12. PMID: 23209232; PMCID: PMC3510637.

Huang, P., Chan, P., Su, P., Chen, L., & Li, H. (2016). Chloroplast Hsp93 directly binds to transit peptides at an early stage of the preprotein import process. Plant Physiology, 170(2), 857–866.

Huang, X., & Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. Advances in Applied Mathematics, 12(3), 337–357. https://doi.org/10.1016/0196-8858(91)90017-D

Hwang, I., & Robinson, D. G. (2009). Transport vesicle formation in plant cells. Current opinion in plant biology, 12(6), 660-669.

Inoue, K. (2015). Emerging knowledge of the organelle outer membranes–research snapshots and an updated list of the chloroplast outer envelope proteins. Front. Plant Sci., 6, 278.

Ivanova, Y., Smith, M. D., Chen, K., & Schnell, D. J. (2004). Members of the Toc159 import receptor family represent distinct pathways for protein targeting to plastids. Molecular biology of the cell, 15(7), 3379-3392.

Jarvis, P. (2008). Targeting of nucleus-encoded proteins to chloroplasts in plants. New phytologist, 179(2), 257-285.

Jores, T., Klinger, A., Groß, L. E., Kawano, S., Flinner, N., Duchardt-Ferner, E., ... & Rapaport, D. (2016). Characterization of the targeting signal in mitochondrial β-barrel proteins. *Nature communications*, *7*(1), 1-16.

Jumper, J., Evans, R., Pritzel, A., … & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. https://doi-org.proxy.lib.uwaterloo.ca/10.1038/s41586-021-03819-2

Kessler, F., Blobel, G., Patel, H.A., & Schnell, D.J. (1994). Identification of two GTP-binding proteins in the chloroplast protein import machinery. Science 266, 1035–1039

Kikuchi, S., Asakura, Y., Imai, M., Nakahira, Y., Kotani, Y., Hashiguchi, Y., … Nakai, M. (2018). A Ycf2-FtsHi heteromeric AAA-ATPase complex is required for chloroplast protein import. The Plant Cell, 30(11), 2677–2703.

Kim, D., Park, M., Gwon, G., Silkov, A., Xu, Z., Yang, E., … Hwang, I. (2014). An ankyrin repeat domain of AKR2 drives chloroplast targeting through coincident binding of two chloroplast lipids. Developmental Cell, 30(5), 598–609.

Kim, Jonghak, Yun Jeong Na, Soon Ju Park, So-Hyeon Baek, & Dae Heon Kim. (2019) Biogenesis of chloroplast outer envelope membrane proteins. Plant cell reports, 38, 783–792.

Klinger, A., Gosch, V., Bodensohn, U., Ladig, R., & Schleiff, E. (2019). The signal distinguishing between targeting of outer membrane β-barrel protein to plastids and mitochondria in plants. BBA - Molecular Cell Research, 1866(4), 663–672.

Kouranov, A., Chen, X., Fuks, B., & Schnell, D (1998). Tic20 and Tic22 are new components of the protein import apparatus at the chloroplast inner envelope membrane. The Journal of Cell Biology, 143(4), 991–1002. https://doi.org/10.1083/jcb.143.4.991

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, *40*(D1), D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Lee, D., Lee, S., Lee, J., Woo, S., Razzak, M., Vitale, A., & Hwang, I. (2019). Molecular mechanism of the specificity of protein import into chloroplasts and mitochondria in plant cells. Molecular Plant, 12(7), 951–966.

Lee, D., Woo, S., Geem, K., & Hwang, I. (2015). Sequence motifs in transit peptides act as independent functional units and can be transferred to new sequence contexts. Plant Physiology, 169(1), 471–484.

Lee, D. W., Lee, J., & Hwang, I. (2017). Sorting of nuclear-encoded chloroplast membrane proteins. Current opinion in plant biology, 40, 1-7.

Lee, D. W., Lee, S., Min, C.-K., Park, C., Kim, J.-M., Hwang, C.-S., Park, S. K., Cho, N.-H., & Hwang, I. (2020). Cross-species functional conservation and possible origin of the N-terminal specificity domain of mitochondrial presequences. Frontiers in Plant Science, 11, 64–64. https://doi.org/10.3389/fpls.2020.00064

Lee, J., Kim, D. H., & Hwang, I. (2014). Specific targeting of proteins to outer envelope membranes of endosymbiotic organelles, chloroplasts, and mitochondria. Frontiers in plant science, 5, 173.

Lee, M. H., & Hwang, I. (2014). Adaptor proteins in protein trafficking between endomembrane compartments in plants. Journal of Plant Biology, 57(5), 265-273.

Li, H., Schnell, D., & Theg, S. (2020). Protein import motors in chloroplasts: on the role of chaperones. *The Plant Cell*, *32*(3), 536–542.

Lin, T., Chen, C., Wu, S., Chang, Y., Li, Y., Su, Y., Hsiao, C., & Chang, H. (2019). Structural analysis of chloroplast tail-anchored membrane protein recognition by ArsA1. The Plant Journal : for Cell and Molecular Biology, 99(1), 128–143. https://doi.org/10.1111/tpj.14316

Lopez-Juez, E., & Pyke, K. A. (2004). Plastids unleashed: their development and their integration in plant development. International Journal of Developmental Biology, 49(5- 6), 557-577.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, Thanki N, Yamashita RA, Yang M, Zhang D, Zheng C, Lanczycki CJ, Marchler-Bauer A. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020 Jan 8;48(D1):D265-D268. doi: 10.1093/nar/gkz991. PMID: 31777944; PMCID: PMC6943070.

Lung, S. C., & Chuong, S. D. (2012). A transit peptide–like sorting signal at the C terminus directs the *Bienertia sinuspersici* preprotein receptor Toc159 to the chloroplast outer membrane. The Plant Cell, tpc-112.

Lung, S. C., Smith, M. D., Weston, J. K., Gwynne, W., Secord, N., & Chuong, S. D. (2014). The C-terminus of *Bienertia sinuspersici* Toc159 contains essential elements for its targeting and anchorage to the chloroplast outer membrane. Front. Plant Sci., 5, 722.

Marfori, M. et al. (2011) Molecular basis for specificity of nuclear import and prediction of nuclear localization. BBA - Mol. Cell Res. 1813, 1562–1577

Natale, P., Brüser, T., & Driessen, A. (2008). Sec- and Tat-mediated protein secretion across the bacterial cytoplasmic membrane—Distinct translocases and mechanisms. BBA - Biomembranes, 1778(9), 1735–1756.

Nelson, N., & Ben-Shem, A. (2004). The complex architecture of oxygenic photosynthesis. Nature Reviews Molecular Cell Biology, 5(12), 971.

Pan, S., Carter, C. J., & Raikhel, N. V. (2005). Understanding protein trafficking in plant cells through proteomics. Expert review of proteomics, 2(5), 781-792.

Panchy, N., Lehti-Shiu, M., & Shiu, S.(2016). Evolution of gene duplication in plants. Plant Physiology (Bethesda), 171(4), 2294–2316. https://doi.org/10.1104/pp.16.00523

Park, M. H., Zhong, R., & Lamppa, G. (2018). Chloroplast stromal processing peptidase activity is modulated by transit peptide determinants that include inhibitory roles for its N-terminal domain and initial Met. Biochemical and biophysical research communications, 503(4), 3149-3154.

Pogson, B. J., & Albrecht-Borth, V. (2014). An overview of chloroplast biogenesis and development. In Plastid Biology(pp. 115-128). Springer, New York, NY.

Pommié, C., Levadoux, S., Sabatier, R., Lefranc, G., & Lefranc, M. P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of molecular recognition : JMR*, *17*(1), 17–32. https://doi.org/10.1002/jmr.647

Pottosin, I., & Dobrovinskaya, O. (2015). Ion channels in native chloroplast membranes: challenges and potential for direct patch-clamp studies. Frontiers in Physiology, 6, 396–396. https://doi.org/10.3389/fphys.2015.00396

Pudelski, B., Schock, A., Hoth, S., Radchuk, R., Weber, H., Hofmann, J., Sonnewald, U., Soll, J., & Philippar, K. (2012). The plastid outer envelope protein OEP16 affects metabolic fluxes during ABA-controlled seed development and germination. Journal of Experimental Botany, 63(5), 1919–1936. https://doi.org/10.1093/jxb/err375

Qbadou S, Tien R, Soll J, Schleiff E (2003) Membrane insertion of the chloroplast outer envelope protein, Toc34: constrains for insertion and topology. J Cell Sci 116: 837-846

Qbadou, S., Becker, T., Mirus, O., Tews, I., Soll, J., & Schleiff, E. (2006). The molecular chaperone Hsp90 delivers precursor proteins to the chloroplast import receptor Toc64. The EMBO journal, 25(9), 1836-1847.

Razzak, M. A., Lee, J., Lee, D. W., Kim, J. H., Yoon, H. S., & Hwang, I. (2019). Expression of seven carbonic anhydrases in red alga *Gracilariopsis chorda* and their subcellular localization in a heterologous system, *Arabidopsis thaliana*. Plant Cell Reports, 38(2), 147–159. https://doi.org/10.1007/s00299-018-2356-8

Richardson, L., Jelokhani-Niaraki, M., & Smith, M. (2009). The acidic domains of the Toc159 chloroplast preprotein receptor family are intrinsically disordered protein domains. BMC Biochemistry, 10, 35.

Richardson, L. G., Paila, Y. D., Siman, S. R., Chen, Y., Smith, M. D., & Schnell, D. J. (2014). Targeting and assembly of components of the TOC protein import complex at the chloroplast outer envelope membrane. Frontiers in plant science, 5, 269.

Sanford, J.C., Smith, F.D., & Russell, J.A. (1993). Optimizing the biolistic process for different biological applications. Methods Enzymol. 217: 482–509

Sambrook, J., Fritsch, E. F., & Maniatis, T. (1989) Molecular cloning: a laboratory manual. Second Edition. Plainview, NY: Cold Spring Harbor Laboratory Press, 1.75.

Schnell, D. J. (2019). The TOC GTPase receptors: regulators of the fidelity, specificity and substrate profiles of the general protein import machinery of chloroplasts. *The protein journal*, *38*(3), 343-350.

Sharma, M., Bennewitz, B., & Klösgen, R. B. (2018). Dual or not dual?-Comparative analysis of fluorescence microscopy-based approaches to study organelle targeting specificity of nuclear-encoded plant proteins. Frontiers in Plant Science, 9, 1350–1350. https://doi.org/10.3389/fpls.2018.01350

Shiota, T., Imai, K., Qiu, J., Hewitt, V., Tan, K., Shen, H., … Endo, T. (2015). Molecular architecture of the active mitochondrial protein gate. Science (New York, N.Y.), 349(6255), 1544–1548.

Sjuts, I., Soll, J., & Bölter, B. (2017). Import of soluble proteins into chloroplasts and potential regulatory mechanisms. Frontiers in plant science, 8, 168.

Slater, A., Scott, N. W., & Fowler, M. R. (2008) Plant biotechnology; the genetic manipulation of plants. (2ed). Oxford. Oxford, NY.

Smith, J.J. and Aitchison, J.D. (2013) Peroxisomes take shape. Nat. Rev. Mol. Cell. Biol. 14, 803–817

Smith, M. D., Hiltbrunner, A., Kessler, F., & Schnell, D. J. (2002a). The targeting of the atToc159 preprotein receptor to the chloroplast outer membrane is mediated by its GTPase domain and is regulated by GTP. J Cell Biol, 159(5), 833-843.

Smith, M. D., Fitzpatrick, L. M., Keegstra, K., & Schnell, D. J. (2002b) In vitro analysis of chloroplast protein import. In: Yamada KM (ed) Current Protocols in Cell Biology. John Wiley & Sons, New York, pp. 11.16.11-11.16.21

Smith, M. D., Rounds, C. M., Wang, F., Chen, K., Afitlhile, M., & Schnell, D. J. (2004). atToc159 is a selective transit peptide receptor for the import of nucleus-encoded chloroplast proteins. The Journal of cell biology, 165(3), 323–334. https://doi.org/10.1083/jcb.200311074

Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z., & Zhu, J.-K. (2018). UTR-dependent control of gene expression in plants. Trends in Plant Science, 23(3), 248–259. https://doi.org/10.1016/j.tplants.2017.11.003

Sun Q, Zybailov B, Majeran W, Friso G, Olinares PD, van Wijk KJ. (2008) PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res. Oct 2 [Epub]

Syed, N. H., Kalyna, M., Marquez, Y., Barta, A., & Brown, J. W. . (2012). Alternative splicing in plants – coming of age. Trends in Plant Science, 17(10), 616–623. https://doi.org/10.1016/j.tplants.2012.06.001

Taiz, L., & Zeiger, E. (2006). Plant physiology. Sinauer Associates. Inc., Sunderland, MA.

Thomas H, Huang L, Young M, Ougham H. (2009). Evolution of plant senescence. BMC Evol. Biol. 9, 163

Teresinski, H. J., Gidda, S. K., Nguyen, T. N., Howard, N. J. M., Porter, B. K., Grimberg, N., ... & Mullen, R. T. (2019). An RK/ST C-terminal motif is required for targeting of OEP7. 2 and a subset of other *Arabidopsis* tail-anchored proteins to the plastid outer envelope membrane. Plant and Cell Physiology.

Ulrich, T., Oberhettinger, P., Schütz, M., Holzer, K., Ramms, A. S., Linke, D., Autenrieth, I. B., & Rapaport, D. (2014). Evolutionary conservation in biogenesis of β-barrel proteins allows mitochondria to assemble a functional bacterial trimeric autotransporter protein. The Journal of Biological Chemistry, 289(43), 29457–29470. https://doi.org/10.1074/jbc.M114.565655

UniProt: the universal protein knowledgebase in 2021. (2021). Nucleic Acids Research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100

Vogt, J., & Schulz, G. E. (1999). The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. Structure (London), 7(10), 1301–1309. https://doi.org/10.1016/S0969-2126(00)80063-5

Wang, X., Cai, Y., Wang, H., Zeng, Y., Zhuang, X., Li, B., & Jiang, L. (2014). Trans-Golgi network-located AP1 gamma adaptins mediate dileucine motif-directed vacuolar targeting in *Arabidopsis.(*Report). 26(10), 4102–4118.

West-Eberhard, M., Smith, J., & Winter, K. (2011). Plant science. Photosynthesis, reorganized. Science (New York, N.Y.), 332(6027), 311–312.

Wiesemann, K., Simm, S., Mirus, O., Ladig, R., & Schleiff, E. (2019). Regulation of two GTPases Toc159 and Toc34 in the translocon of the outer envelope of chloroplasts. BBA - Proteins and Proteomics, 1867(6), 627–636.

Wise, R. R., & Hoober, J. K. (Eds.). (2006). The structure and function of plastids (Vol. 23). Springer Science & Business Media.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nature methods, 12(1), 7.

Yeung, K., & Chuong, S.D.X. (2016, June) Characterization and subcellular localization of *Arabidopsis* putative guanosine triphosphate binding protein. Poster session presented at the Canadian Society of Plant Biology meeting, Guelph, ON.

Yoo, S. D., Cho, Y. H., & Sheen, J. (2007). Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. Nature protocols, 2(7), 1565.

Zimorski, V., Ku, C., Martin, W. F., & Gould, S. B. (2014). Endosymbiotic theory for organelle origins. Current opinion in microbiology, 22, 38-48.

Appendix

Table A1: Construct Designs as Ordered from BioBasic

[1] Protein sequences represented in single letter code. Red Methionine (M) and red asterisks (*)

CONSTRUCT NAME	INSERT SEQUENCE TRANSLATION[1]	RESTRICTION ENZYMES	VECTOR
OEP15-1:GFP	MGTMVLPWQGFISREVSNLCNSRR NELTLGGLVTFFGTTRSEEDSSYEG NLELRLREADPIGQNQPHMGVSLK NSKDDLTVTANLRHQVSVGRQTK VTTFVSLDSKRTGCFTVRTNSDQL QIAVMALLLLAM	SalI/ApaI	pSAT6-N1
GFP:OEP15-1	MGTMVLPWQGFISREVSNLCNSRR NELTLGGLVTFFGTTRSEEDSSYEG NLELRLREADPIGQNQPHMGVSLK NSKDDLTVTANLRHQVSVGRQTK VTTFVSLDSKRTGCFTVRTNSDQL QIAVMALLLLAM*	BglII/BamHI	pSAT6-C1
OEP15-1CT:GFP	MSVGRQTKVTTFVSLDSKRTGCFT VRTNSDQLQIAVMALLLLAM	SalI/ApaI	pSAT6-N1
GFP:OEP15-1CT	SVGRQTKVTTFVSLDSKRTGCFTV RTNSDQLQIAVMALLLLAM*	BglII/SalI	pSAT6-C1
OEP15-1ΔCT:GFP	MGTMVLPWQGFISREVSNLCNSRR NELTLGGLVTFFGTTRSEEDSSYEG NLELRLREADPIGQNQPHMGVSLK NSKDDLTVTANLRHQV	SalI/ApaI	pSAT6-N1
OEP7.3:GFP	MGVSLKNSKDDLTVTANLRHQVS VGRQTKVTTFVSLDSKRTGCFTVR TNSDQLQIAVMALLLLAM	SalI/ApaI	pSAT6-N1
GFP:OEP7.3	MGVSLKNSKDDLTVTANLRHQVS VGRQTKVTTFVSLDSKRTGCFTVR TNSDQLQIAVMALLLLAM*	BglII/SalI	pSAT6-C1
HFH:DSRED	MYRALRLLARSRPLVRAPAAAL	SalI/SmaI	pSAT6-N1-DsRed[2]

[1] Protein sequences represented in single letter code. Red Methionine (M) and red asterisks (*) represent start and stop codons added to make vector code fusion protiens properly which were not in the original CDS or OEP15-1 or OEP7.3.

[2] A modified vector which replaces the EGFP in pSAT6-N1 with DsRed.

A

BsmI   SwaI

ccactatccttcgcaagacccttcctctatataaggaagttcatttcatttggagaggacgtcgagagttctcaacacaacatatacaaaacaaacgaatctcaagcaatcaagcattctacttctattgcagcaatttaaatcatttcttttaaagcaa
ggtgataggaagcgttctgggaaggagatatattccttcaagtaaagtaaacctctcctgcagctctcaagagttgtgttgtatatgttttgtttgcttagagttcgttagttcgtaagatgaagataacgtcgttaaatttagtaaagaaaatttcgtt

1280

SacII PspOMI
Eco53kI SacI   KpnI   XmaI
TliI   HindIII   TspMI
XhoI   Acc65I   ApaI
PaeR7I   BstBI   SalI   SmaI BamHI
StyI   EcoRI   PstI
NcoI   BglII

aagcaattttctgaaaattttcaccatttacgaacgatagcatggtccggactcagatctcgagctcaagcttcgaattctgcagtcgacggtaccgcgggccgggatcctgatggtgagcaagggcgaggagctgttcaccggggtggtgcccatcc
ttcgttaaaagactttttaaaagtggtaaatgcttgctatcggtaccaggcctgagtctagagctcgagttcgaagcttaagacgtcagctgccatggcgcccgggccctaggactaccactcgttcccgctcctcgacaagtggcccccaccacgggtagg

1440

M V R T Q I S S S S F E F C S R R R Y R G P G I L M V S K G E E L F T G V V P I →

B

#	AA	a4v	HSA	NHSA	a4vAHS
1	M	0.010	0.375	0.075	0.052
2	V	0.010	0.375	0.075	0.052
3	R	0.010	0.375	0.075	0.052
4	T	0.200	0.375	0.075	0.052
5	Q	0.028	0.375	0.075	0.052
6	I	-0.241	0.000	0.000	0.000
7	S	-0.106	0.000	0.000	0.000
8	S	0.167	1.213	0.173	0.153
9	S	0.141	1.213	0.173	0.153
10	S	0.131	1.213	0.173	0.153
11	F	0.260	1.213	0.173	0.153
12	E	0.260	1.213	0.173	0.153
13	F	0.125	1.213	0.173	0.153
14	C	-0.011	1.213	0.173	0.153
15	S	-0.096	0.000	0.000	0.000
16	R	-0.071	0.000	0.000	0.000
17	R	-0.398	0.000	0.000	0.000
18	Y	-0.532	0.000	0.000	0.000
19	R	-0.360	0.000	0.000	0.000
20	G	-0.259	0.000	0.000	0.000
21	P	0.146	0.000	0.000	0.000
22	M	-0.062	0.000	0.000	0.000
23	G	0.312	0.745	0.000	0.228
24	V	0.256	0.745	0.000	0.228
25	S	0.117	0.745	0.000	0.228
26	L	-0.055	0.000	0.000	0.000
27	K	-0.111	0.000	0.000	0.000
28	N	-0.601	0.000	0.000	0.000
29	S	-0.821	0.000	0.000	0.000
30	K	-0.821	0.000	0.000	0.000

Figure A1: pSAT6-N1 MCS with Highlighted 21 AA Additional in Frame Peptide and Submission of Peptide to AGGRESCAN. A. View from SnapGene showing the MCS of vector pSAT6-N1. The blue highlighting covers the 21 AA which are translated at the beginning of clones generated from blunt end cloning using RE SmaI due to their being in frame with the inserted protein and GFP. GFP begins "MVSKGEEL". Unique RE sites are shown above the nucleotide sequence. OEP15-1NT:GFP which also contains an additional upstream peptide was cloned using SalI and BamHI so it only includes the first 15 of the highlighted AAs. B. The first 30 AA of readout when OEP7.3β3-4α:GFP and additional 21 AA N-terminal peptide were submitted to AGGRESCAN. Red highlighting designates predicted hotspots. Again, these hotspots are not predicted when only the 21AA peptide is submitted to the same program.

129

Table A2: Phylogenetic Tree Sequence identities

Name on Tree	Translocase	Organism	Accession
OEP15	N/A	*Arabidopsis thaliana*	AT4G02482/ F4JHJ5-ARATH
Attoc159	Toc159	*Arabidopsis thaliana*	AT4G02510
Attoc120	Toc120	*Arabidopsis thaliana*	AT3G16620
Attoc90	Toc90	*Arabidopsis thaliana*	AT5G20300
Attoc132	Toc132	*Arabidopsis thaliana*	AT2G16640
Altoc132	Toc132	*Arabidopsis lyrata*	XP_020887996.1
Crtoc132	Toc132	*Capsella rubella*	XP_023641827.1
Estoc132	Toc132	*Eutrema salsugineum*	XP_024016148.1
Altoc90	Toc90	*Arabidopsis lyrata*	XP_020879152.1
Crtoc90	Toc90	*Capsella rubella*	XP_023635747.1
Estoc90	Toc90	*Eutrema salsugineum*	XP_006400598.1
Bntoc90	Toc90	*Brassica napus*	XP_013667313.1
Attoc86	Toc159 (alt splice)	*Arabidopsis thaliana*	AT4G02510
Altoc159	Toc159	*Arabidopsis lyrata*	ARALYDRAFT_911963
Crtoc159	Toc159	*Capsella rubella*	XP_023636426.1
Estoc159	Toc159	*Eutrema salsugineum*	XP_006396459.2
Bntoc159	Toc159	*Brassica napus*	XP_022563616.1
Altoc120	Toc120	*Arabidopsis lyrata*	XP_020889089.1
Crtoc120	Toc120	*Capsella rubella*	XP_006296873.1
Estoc120	Toc120	*Eutrema salsugineum*	XP_006406820.2
Bntoc120	Toc120	*Brassica napus*	XP_013749530.1

A

MGTMVLPWQGFISREVSNLCNSRRNELTLGGLVTFFGTTRSEEDSSYEGNLELRLREADFPIGQNQPHMGVSLKNSKDDLTVTANLRHQVSVGRQTKVTTFVSLDSKRTGCFTVRTNSSDQLQIAVMALLLLAM

β-strand   α-helix

B

Sequence   MGTMVLPWQGFISREVSNLCNSRRNELTLGGLVTFFGTTRSEEDSSYEGNLELRLREADFPIGQNQPHMGVSLKNSKDDLTVTANLRHQVSVGRQTKVTTFVSLDSKRTGCFTVRTNSSDQLQIAVMALLLLAM
Prediction   CCCSSSSSSSSSSSSSSSSSCCCCCCCSSCCSSSSSCCSSSCCCCCSSSSSSSSSSSCCCCCCCCCCHHHHSSSSCCCCSSSSSCCSSSSSCCCCCSSSSSCCCCCCCCCSSSSSCCHHHHHHHHSSSSSCC
Conf.Score   972562107899860123300102303023599962244316851220378999505888876642011101000146758860237767037883599996116776550899945337677364335564l9
              H:Helix; S:Strand; C:Coil

C

Pred CCCCCCCCCCEEEECCCCCCCCCCCCCCCCEEEEEEEEECCCCCCCCCCEEEEEEEEEEECCCCCCCCCCCCEEEEEEEEECCCEEEEEEEEEECCCCCEEEEEEECCCCCCEEEEEEECCHHHHHHHHHHHHHHHC
AA   MGTMVLPWQGFISREVSNLCNSRRNELTLGGLVTFFGTTRSEEDSSYEGNLELRLREADFPIGQNQPHMGVSLKNSKDDLTVTANLRHQVSVGRQTKVTTFVSLDSKRTGCFTVRTNSSDQLQIAVMALLLLAM

Legend:
Strand
Helix
Coil
Conf: Confidence of prediction
Cart: 3-state assignment cartoon
Pred: 3-state prediction
AA: Target Sequence

Figure A2: Secondary Structure Prediction for OEP15-1 by Three Tools. Secondary structure as predicted by A. JPred4, B. I-TASSER, and C. PSIpred for ease of comparison.
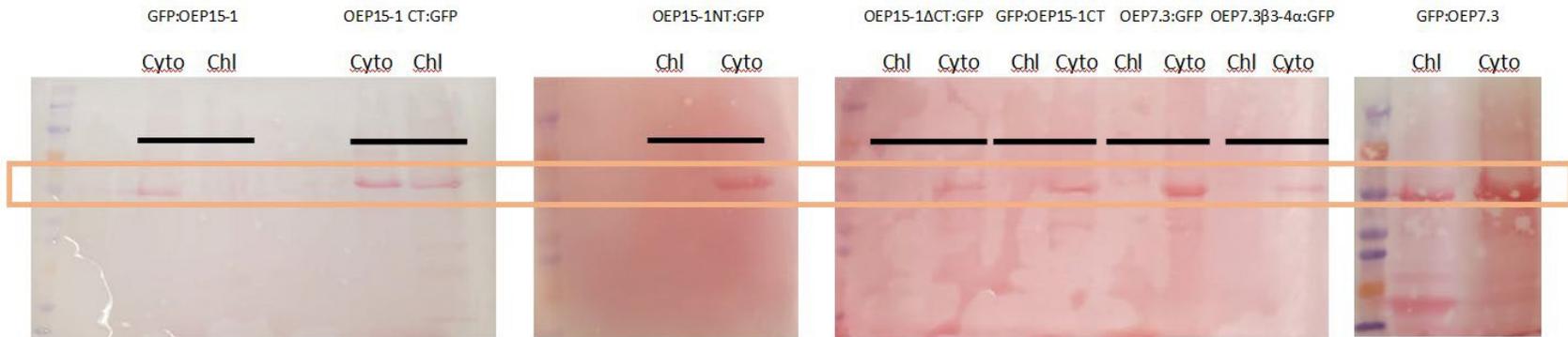
131

Figure A3: Ponceau Stains of Chloroplast vs Cytosolic Protein Fractions. Ponceau stained protein transfers for all fractionated Western blots in figure 3.12. Orange box highlights proteins at the molecular weight expected for RbcL – 55kDa. Amount RbcL in cytosolic fraction used to estimate chloroplast lysis.
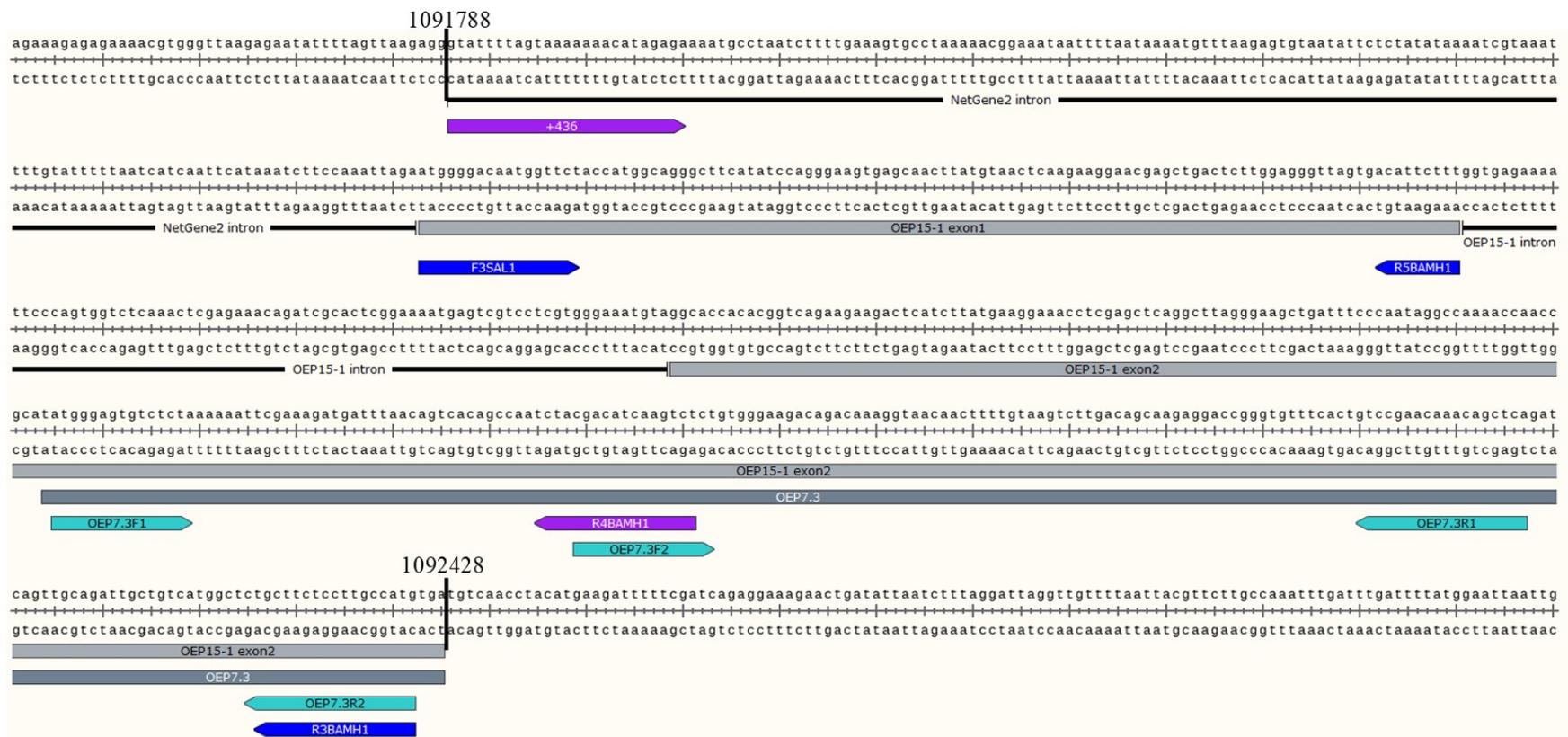
Figure A4: Primer Design. Partial sequence of Chromosome 4 of *A. thaliana* highlighting features predicted between nucleotides 1091788 and 1092428. Primers shown in teal were used to generate OEP7.3 truncations. Primers shown in purple were used to test cDNA contamination with genomic DNA in Figure 3.1B. Primers shown in Blue were used to amplify OEP15-1 cDNA in Figure 3.1A (F3SAL1 and R3BAMH1) or when generating the OEP15-1NT:GFP clone (F3SAL1 and R5BAMH1).

Table A3: Primers and Annealing Temperatures Used in Transcript Detection PCRs

Experiment [1]	Forward Primer [2]	Reverse Primer [2]	Annealing Temp. (^{O}C)
OEP15-1 transcript detection (**A**)	F3SAL1	R3BAMH1	52
Genomic Contaminant Detection (**B**)	+436	R4BAMH1	50
3' UTR amplification (**C**)	OEP7.3F2	dT$_{23}$VN (randomly anchored oligo dTs)	48

[1]Experiment column includes a brief description of the experiment goal as well as the frame letter of Figure 3.1 in which the results can be viewed in brackets (A, B, or C)

[2]Primer locations excluding the anchored oligo dTs can be viewed in Figure A4