NOVEL AUTOPROCESSING DOMAINS FOR STOICHIOMETRIC PROTEIN CO-EXPRESSION IN PLANTS

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MOLECULAR BIOSCIENCES AND BIOENGINEERING

MAY 2015

By

Bei Zhang

Dissertation Committee

Wei Wen Su, Chairperson
Jon-Paul Bingham
John Hu
Yong Soo Kim
Ho Leung Ng

Keywords: Polyprotein Expression, Plants, Kex2p, Intein, FMDV 2A

ACKNOWLEDGMENT

I would like to thank my advisor Dr. Wei Wen Su for his generous support and mentoring during the entire process of my Ph.D. study. Also, I would like to thank my other committee members: Dr. John Hu and his former student Dr. Kishore Dey for providing wild type *N. benthamiana* seeds and suggestions on plant maintenance and transient expression via agroinfiltration; Dr. Ho-Leung Ng for his guidance on structural analysis of the intein mutants; Dr. Jon-Paul Bingham for his suggestions on the protein purification and mass spectrometry analysis; Dr. Yong Soo Kim for his helps with mammalian cell culture and transfection. I am grateful for the help of my former lab mates, Dr. Maribel Zaporteza, Dr. Kun Yang, Ms. Reiko Tanaka and Mr. Ryan Okimoto, for teaching skills in protein purification and protein characterization. Particularly, I appreciate Mr. Leyang Huang for his investigation of the RKG-HDEL construct; Mr. Rosanto Paramban and Dr. Bob Bugos for creating the pBIN-GG construct; Dr. Monto Kumagai for his assistance in creating the TMV viral vector; and Dr. Madhusudhan Rapolu for assistance with the research. I also thank Dr. Zhenlin Han for helping in construction of pSKDuetGM vector.

I am very grateful to the assistance from other interdepartmental colleagues, particularly to Dr. Philip William from Chemistry Department and his former student Mr. Zhibin Liang for their assistance in ESI-TOF mass spectrometry analysis; Drs. Johann Urschitz and Stefan Moisyadi from Institute of Biogenesis Research for providing the HEK293T cell line and advices on cell maintenance and transfection; Dr. Paul Patek from Microbiology Department for preserving human TF1 cells and advices on cell maintenance; Dr. Kheng Cheah from Tropical Plant and

Soil Sciences Department for advices on *N. tabacum* tissue culture and transformation; and Dr. Qing Li from Department of Molecular Biosciences and Bioengineering for letting me use instruments in his laboratory.

I would also like to thank external colleagues Drs. Bernahard Hube from Friedrich-Schiller University, John Rogers from Washington State University and Stanton Gelvin from Purdue University for providing *Saccharomyces* kex2p expressing *Phichia* strain, pLJ607 vector and pMSP-1 series binary vectors (pE1226 and pE1775), respectively.

I am thankful to my friends, Margaret Baker, Joycelyn Chun, Camila Ortega-Ramirez, Archana Pal, Vishal Negi and Jannai Yafuso, as well as to the Chinese community at University of Hawaii for their companion and support.

Lastly, I would like to dedicate this dissertation to my husband, Zhibin Liang, my parents, Hongwei Gao, Ke Zhang, and other family members who always be the source of the keenest love and strength during the joyful and frustrating time of my Ph.D. study.

ABSTRACT

Coordinate expression of multiple proteins in plants plays important roles in both fundamental research and commercial practice in plant biotechnology. However, current technologies amenable to precise, efficient and stoichiometric multi-protein expressions are not yet available. This study aims to advance current state-of-the-art in multi-protein co-expression by developing novel polyprotein-based expression vector systems. These vectors enable stoichiometric expression of multiple proteins from a single open reading frame (ORF) that encodes a polyprotein precursor which is composed of multiple proteins connected by an autoprocessing domain. The unique autoprocessing domains allow individual proteins to be separated from the polyprotein precursor via either co-translational or post-translational cleavage by endogenous host proteases or self-catalytic cleavage activity. As a result, nearly stoichiometric expression of multiple proteins can be achieved using a single expression cassette. In this study, three independent polyprotein cassettes mediated by respective autoprocessing domains: (1) kex2p-like protease substrate sequence, (2) dual-intein and (3) intein-FMDV2A domain, were devised and investigated for coordinate co-expression of multiple proteins in plants. Efficient release of individual protein moiety was demonstrated in tobacco cells as well as in different plant species using immunoblotting analysis. Cellular cleavage mechanisms of the polyprotein precursors were elucidated by N-terminal amino acid sequencing along with mass spectrometry analysis. Functionality of released proteins was characterized by fluorescence spectroscopy and bioactivity assays. Specific cellular targeting of released proteins can be achieved by incorporation of suitable signal sequences. Collectively, our results have

successfully demonstrated the utility of different engineered autoprocessing domains for efficient co-expression of multiple proteins from novel polyprotein vectors. This technology should be highly useful in a wide range of applications, from improving fundamental studies of molecular mechanisms, to enhance crop yield, stress/disease tolerance, nutritional values, and production of high-value proteins.

TABLE OF CONTENTS

Acknowledgment ii

Abstract iv

List of Figures viii

List of Abbreviations xii

CHAPTER 1 1

Introduction 1

1.1. Overview of Gene Stacking in Plants 1

1.2. Stacking of Multiple Monocistronic Expression Cassettes in the Same Expression Vector 3

1.3. Polycistronic Cassette Based on the Internal Ribosomal Entry Sequence (IRES) 5

1.4. Foot-and-Mouth Disease Virus 2A Peptides 7

1.5. Putative Protease Substrate Sequences 10

1.6. Objectives and Experimental Design 12

CHAPTER 2 21

Coordinate Expression of Multiple Proteins in Plant Cells by Exploiting Plant Endogenous Kex2p-Like Protease Activity 21

2.1. Introduction 22

2.2. Materials and Method 25

2.3. Results 34

2.4. Discussion 46

CHAPTER 3 57

A Dual-Intein Autoprocessing Domain that Directs Synchronized Protein Co-Expression in Both Prokaryotes and Eukaryotes 57

3.1. Introduction 58

3.2. Materials and Method 61

3.3. Results and Discussion 68

3.4. Conclusion 80

CHAPTER 4 94

Improving the Self-Cleaving Polyprotein Expression System by Harnessing the Synergy Between An Intein and A FMDV 2A Sequence 94

4.1. Introduction 95

4.2. Materials and Method 98

4.3. Results 104

4.4. Discussion 112

CHAPTER 5 135

Conclusion and Recommendations for Future Studies 135

Appendices 139

References and Notes 149

LIST OF FIGURES

Figure 1.1. Gene stacking in plants via co-transformation of multiple monocistronic cassettes on the same or separated vectors. 16

Figure 1.2. Gene stacking in plants via polycistronic vector mediated by Internal Ribosome Entry Sites (IRES)..... 17

Figure 1.3. Gene stacking in plants via polyprotein cassette mediated by ribosomal recoding capable 2A sequence from Foot-and-Mouth Disease Virus (FMDV 2A)..... 18

Figure 1.4. Coordinate gene stacking in plants by exploiting the endogenous plant protease activity..... 19

Figure 1.5. Coordinate gene stacking in plants via the built-in TEV-NIa protease. 20

Figure 2.1. Organization of the different vector constructs used in this study..... 49

Figure 2.2. RKG cells display both green and red fluorescence..... 50

Figure 2.3. *In vivo* processing of RKG and GKG fusion proteins and secretion of the cleaved protein products. 51

Figure 2.4. Processing of RKG and GKG fusion proteins in different tissues of tobacco plants. 52

Figure 2.5. Analysis of RKG fusion protein and the cleaved protein products..... 53

Figure 2.6. Bioactivity of GMCSF processed from GKG fusion protein produced by the GKG cells. 54

Figure 2.7. Presence of the kex2p linker sequence is necessary for polyprotein cleavage *in planta*. 55

Figure 2.8. Cleavage of RKG fusion protein in different plant species. 56

Figure 3.1. Synchronized co-expression of multiple proteins using the DI-based polyprotein vector system.. 82

Figure 3.2. Mechanism of intein mediated protein splicing. 83

Figure 3.3. DI mediated autoprocessing of GFP_{172}-DI-RFP_{Strep} polyprotein in *E. coli*, mammalian HEK293T, and tobacco NT1 cells. 84

Figure 3.4. Processing of the GFP_{172}-DI-RFP_{Strep} polyprotein in plants, probed with western blots. 85

Figure 3.5. Characterizing autoprocessing of the GFP_{172}-DI-RFP_{Strep} polyprotein. 86

Figure 3.6. Effects of cellular redox environment and culture temperature on DI autocleavage activity. 87

Figure 3.7. Processed POIs from the DI-based polyprotein precursor preserve proper function 88

Figure 3.8. Relative expression levels of the constituent proteins released from the GFP-DI-RFP polyprotein precursor in different expression hosts. 89

Figure 3.9. DI polyprotein system is also applicable to co-expressing proteins other than fluorescent reporters. 90

Figure 3.10. Comparison of protein expression levels using the DI based polyprotein vector vs. the duet vector in *E. coli* BL21(DE3) or SHuffle strain under the control of T7 promoter 91

Figure 3.11. DI domain without the N-terminal linker extension preserves high autocleavage activity 92

Figure 3.12. Inefficiency of single intein mediated polyprotein processing in *E. coli* 93

Figure 4.1. Organization of the IntF2A based polyprotein constructs used in this study.....119

Figure 4.2. Organization and cellular processing of the IntF2A based polyprotein. 120

Figure 4.3. *In vivo* processing of the IntF2A based polyprotein expressing two fluorescent proteins from pND-1 in tobacco NT1 cells..... 121

Figure 4.4. *In vivo* processing of IntF2A based polyprotein (ND-1) in transgenic *N. tabacum* cv. Xanthi plants. 122

Figure 4.5. *In vivo* processing of IntF2A based polyprotein in different plant species.. 123

Figure 4.6. Proper fluorescence property was detected from POIs processed from the ND-1 polyprotein precursor. 124

Figure 4.7. Co-expression of three POIs using the IntF2A based polyprotein. 125

Figure 4.8. Expression level of POIs (GFP & RFP) processed from the IntF2A based polyprotein compared to that from the single protein vector..... 126

Figure 4.9. Application of IntF2A based polyprotein for co-expressing multiple proteins other than fluorescent reporters in plants. 127

Figure 4.10. Bioactivity of $_{His}$hGMCSF processed from the IntF2A based polyprotein precursor 128

Figure 4.11. Analysis of cellular cleavage sites within the IntF2A polyprotein precursor ND-1 129

Figure 4.12. Subcellular targeting of constituent POIs using the IntF2A based polyprotein expression system in plants..... 130

Figure 4.13. Incorporation of CSC flanking C-terminus of the intein domain is sufficient to

promote efficient released of upstream POI without adding additional residues to its C-terminus..... 131

Figure 4.14. Proteins with F2A sequence at their C-terminus may be unstable..... 132

Figure 4.15. C-terminal F2A extension causes mistargeting of POI to vacuole when a ER targeting signal is included on the N-terminus of the POI..... 133

Figure 4.16. Determination of cellular destination of released IntF2A fragment by western blot analysis..... 134

LIST OF ABBREVIATIONS

2,4-D	2,4 Dichlorophenoxy acetic acid
AMV	Alfalfa Mosaic Virus
BAP	6-Benzyl Amino Purine
BCIP	5-Bromo-4-Chloro-3-Indolyl Phosphate
BisTris	Bis(2-hydroxyethyl)-amino-Tris(hydroxymethyl)-methane
BMGY	Buffer Minimal Glycerol Medium
BMMY	Buffer Minimal Methanol Medium
CaMV 35S	Cauliflower Mosaic Virus 35S
CAT	Chloramphenicol Acetyltransferse
CMV	Cytomegalovirus
DEAE	Diethylaminoethyl Cellulose
DI	Dual-Intein
DMEM	Dulbecco's Modified Eagle's Medium
DsRed	*Discosoma* sp. Red Fluorescence Protein
DTNB	5,5’-Dithio-bis (2-Nitrobenzoic Acid)
DTT	Dithiothreitol
EDTA	Ethylene Diamine Tetraacetic Acid
ER	Endoplasmic Reticulum
ESI	Electrospray Ionization

FBS	Fetal Bovine Serum
FMDV	Foot-and-Mouth Disease Virus
GFP	Green Fluorescence Protein
GUS	Beta-Glucuronidase
HEK	Human Embryonic Kidney
hGMCSF	Human Granulocyte Macrophage Colony Growth Stimulating Factor
HPLC	High Performance Liquid Chromatography
IEC	Ion Exchange Chromatography
IMAC	Immobilized Metal Ion Affinity Chromatography
IntF2A	Intein-FMDV 2A
IP	Immunoprecipitation
IPTG	Isopropyl-β-D-thiogalactopyranoside
IRES	Internal Ribosome Entry Site
LB	Luria-Bertani Broth
LUC	Luciferase
MALDI	Matrix-Assisted Laser Desorption/Ionization
mKO1	monomeric Kusabira Orange Fluorescence Protein
MS	Mass Spectrometry
MS medium	Murashige and Skoog medium
NBT	Nitroblue Tetrazolium
(ocs)3/mas	Mannopine/Octopine Synthase

ORF	Open Reading Frame
PBS	Phosphate Buffer Saline
PCR	Polymerase Chain Reaction
PMSF	Phenylmethylsulfonyl Fluoride
POI	Protein of Interest
PVDF	Polyvinylidene Fluoride
RBS	Ribosome Binding Site
RFP	Red Fluorescence Protein
RPMI-1640	Roswell Park Memorial Institute-1640 Medium
SDS-PAGE	Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis
TEV	Tobacco Etch Virus
TOF	Time-of-Flight
Tris	Tris(hydroxymethyl)aminomethane
Trx	Thioredoxin
TSP	Total Soluble Proteins
TVMV	Tobacco Vein Mottling Virus
Ub	Ubiquitin
WT	Wild-type

CHAPTER 1

INTRODUCTION

1.1. Overview of Gene Stacking in Plants

In the past decades, much success has been gained in the area of transgenic modification of plants. For example, transgenic plants with modified flower color as well as resistance to insects, herbicides, and pathogenic microbes have been reported and are commercially available. Since the first commercial introduction of transgenic crops in 1996, the global development and cultivation of biotech crops have been constantly rising.[1] In 2012, transgenic crops were adopted by over 17.3 million farmers on over 170.3 million hectares, globally. The global value of biotech seeds alone was $15 billion in 2012, which represents 35% of the $34 billion in the commercial seed market.[1] In addition to the economic benefits, biotech crops also contribute in part to the reduction of agricultural impact on the global environmental issues.[2] From 1996 to 2012, the use of pesticides was reduced by more than 473 million kg as a result of planting insect resistance crops. The greenhouse gas emissions from pesticide manufactures were reduced by 23.1 billion kg in 2012 alone, which is equivalent to the total amount of exhaust gas from 10.2 million cars per year.[1]

Currently, most commercially transgenic crops have a single insect resistance gene (e.g., Bt toxin). However, there is growing concern of pathogen resistance when the single mode of inactivation is excessively used. Also, modification of plants with complex traits often requires introduction of an entire metabolic pathway involving multiple protein components that function

synergistically. Thus, biotech crops containing multiple transgenes are showing increasing market demands.[3-4] In 2012, planting of biotech crops with multiple traits occupied 26% of the total 170 million hectares in 13 countries around the world.[1] As advance in the field of plant biotechnology aims to meet demands for higher quality, yield and new crop varieties, the need for effective transgenic trait stacking will continue to grow.[5]

Traditional transgenic trait stacking requires cross fertilization of parents transformed with individual genes. This process is laborious and time consuming, particularly as the number of required transgenes increases. Excessive screening for the homozygous progenies is also required to prevent the loss of transgenes from the loci segregation during subsequent propagations.[4-5] Moreover, the stacking number of transgenes is limited by the scarcity of effective marker gene for transformation selection and strong plant promoter for overproduction of target proteins.

With years of investigation in plant biotechnology, novel techniques have been developed to overcome the current limitations. Numerous desirable traits have been reported to be successfully stacked in various crop species by the use of these superior technologies. It is conceivable that development of new and superior techniques to facilitate efficient gene stacking in plants will be continuously significant in future with a growing demand of higher quality crops.

1.2. Stacking of Multiple Monocistronic Expression Cassettes in the Same Expression Vector

Before sophisticated polyprotein vectors were invented, researches involving synergistic incorporation of multiple protein domains mainly relied on the iterative introduction of multiple monocistronic expression cassettes on the same or separated vector (Fig. 1.1). Owing to its simple construct design and transformation process, this approach has been widely applied to many encouraging studies including phytoremediation,[6-7] crop protection,[8-14] biomass enhancement, as well as production of valuable pharmaceutical agents.[15] Co-transformation of multiple vectors showed advantage for simplification of the construct cloning process, particularly for the studies on differential combinations of variational components.[16] It has been reported that co-transformation of two rice originated cystatin genes successfully conferred Colorado potato beetle resistance in two different potato species (Desiree and Dragačevka) by reducing adult body weight and shortening the maturation time of pests.[17] However, co-transformation efficiency was relatively low and excessive screening was required to identify the lines that receive all the desired transgenes.[18-19] More importantly, genotype dependent T-DNA integration might occur during co-transformation process, causing the extremely low co-transformation efficiency for certain plant species.[20] For example, when the two Colorado potato beetle resistance genes OCI and OCII on separated plasmids attempted to be co-transformed into potato cultivar Jelica, co-existence of both genes was found in only 2% of transformants.[17]

The low co-transformation efficiency using multiple separated monocistronic vector can be simply resolved by iterative transformation of the subsequent transgenes into the previously

established transgenic background. Using this stepwise transformation strategy, five genes from *Escherichia coli* glycolate catabolic pathway were introduced into *Arabidopsis thaliana* to increase photosynthesis by reducing the photorespiration loss in chloroplast.[21] Improvement of biomass accumulation in transgenic *A. thaliana* was observed as increased photosynthesis enable accumulation of more glyoxylate in comparison with the non-transformed counterpart. In another example, tuber starch of transgenic potatoes was doubled by iterative transformation of two metabolite translocators that are responsible for increasing the carbohydrate transportation towards tuber.[22]

Despite the encouraging achievements reported previously, a number of limitations are noted in the application of co-transformation or iterative strategy to multiple genes stacking in plants. One of the major problems is the requirement of different selection markers for the screening of each transgene. Although this problem can be temporarily resolved by the vigorous development of novel selection markers,[23] synchronous overproduction of multiple antibiotic resistance genes might induce metabolic burden that retards the growth of downstream transformants. While selectable marker can be removed by introducing the bacterial recombination system after each transformation, modified chromosome or irregular genotype resulted from expression of these bacterial recombination systems was also noted in plants.[24-27] Aside from the selectable marker genes, excessive use of the same promoter for protein expression in co-transformation strategies may cause homology-dependent gene silencing, which might hamper from the generation of homozygote transformations harboring all the transgenes.[28] This hemizygosity might lead to the segregation of transgenes in progenies, which could be eventually lost in subsequent generations. Furthermore, coordinate co-expression of the individual transgene is difficult to achieve when

they are expressed separately under the control of different promoters. Even though the same promoter was used, expression level of the individual transgene could be different due to the interference of the flanking integration sequence within the plant genome. This imbalance in protein expression is particularly problematic in the production of hetero-oligomeric protein complexes as their overall yield is highly dependent on the lowest expressed subunit. Instead, connecting the constituent proteins into a single fusion protein can remedy this imbalanced protein expression. For example, Molke and co-workers introduced an entire photosynthetic carbon fixation pathway into potato plastids by fusing different subunits of *E. coli* glycolate dehydrogenase via a long and flexible linker.[29] Increased biomass production of transgenic potato was confirmed and the accumulation of carbohydrate content showed correlation with expression level of the recombinant protein complex. While this stoichiometric accumulation of multiple protein subunits indeed maximized the production yield of the protein complex, it is not practical to produce proteins with exposed termini critical for proper function or those aimed to be targeted in different cellular compartments. Thus, new approaches that enable production of equal molar and physically separated proteins are necessary to compensate the aforementioned limitations.

1.3. Polycistronic Cassette Based on the Internal Ribosomal Entry Sequence (IRES)

Internal Ribosomal Entry Sequences (IRESs) promote non-canonical cap-independent protein translation initiation of the viral proteins upon transfection in host cells.[30-31] Although first discovered in picornavirus, different IRES variants have been identified in a broad range of

eukaryotes such as yeast, insects and animals. Because of their unique protein translation initiation activities, the IRES sequence has been exploited in the polycistronic expression cassette for co-expression of multiple proteins in a variety of eukaryotic systems (Fig. 1.2). In this case, the upstream protein is translated in cap-dependent manner whereas the downstream protein is translated through the internal translation process mediated by the IRES. The IRES activity in plants was first demonstrated by successfully co-express two discrete reporter proteins, a GFP and a luciferase (LUC), in transgenic *Nicotiana tabacum* from the bicistronic vector mediated by the encephalomyocarditis virus IRES (EMCV IRES).[32] While both GFP fluorescence and LUC activity can be detected in leaf tissue, no LUC activity was detected in root and hypocotyl. When the same EMCV IRES sequence was used to deliver two nematode resistance protease inhibitors in *N. tabacum* plants, expression of the downstream protein was noticeably reduced in all tested tissues based on ELISA measurement and protease activity assay. More reduction of downstream protein expression was detected in roots compared to that in leaf tissue.[33] These results indicate that the IRES mediated translation initiation is less efficient compared to the canonical cap-dependent protein translation and is found to be tissue specific. In addition to plants, similarly imbalanced protein production has been consistently demonstrated when the IRES based polycistronic expression system was used in different eukaryotic hosts.[34-36] The internal translation initiation activity also varies between IRES sequences originated from different viruses when expressed in various cell types. Some IRES sequences (e.g., IRES from c-myc proto-oncogene) failed to maintain constant expression of downstream protein in mammalian cells.[37] It has been revealed that the tertiary structure of certain conserved motifs within IRES sequence, for example domain 3 in FMDV IRES, played a crucial role in the

efficiency of internal translation initiation.[35] Such unique tertiary structures maintained by the internal sequence interaction are sensitive to the condition changes of reaction environment when test *in vitro*.[38] Besides, several cellular RNA binding proteins have been identified to enhance IRES activity by *in trans* interaction of translation factors.[39] Lower binding efficiency of essential cellular *trans*-acting elements (e.g., polypyrimidine-tract-binding protein (PTB) and La proteins) with eukaryotic mRNA might contribute to attenuating IRES based cap-independent translation efficiency in eukaryotic system.[40] The diverse IRES activities in different tissues and in developmental stages of cells were thought to be ascribed to the variable cellular level of IRES specific *trans*-acting factors during the growth cycles of different cell types.[41] The imbalanced protein expression using the IRES based polycistronic expression system would be favorable for expression of protein complexes assembled by protein domains with different oligomeric stages. However, attenuated cap-independent translation initiation efficiency mediated by the IRES limits its application when stoichiometric accumulation of co-expression partners is required. Such limitations of the IRES based polycistronic protein expression system spur needs for development of more efficient protein elements capable of coordinate production of multiple proteins from a single transcript.

1.4. Foot-and-Mouth Disease Virus 2A Peptides

The 2A region initially found in the natural polyprotein of picornaviruses was identified to be responsible for the primary viral polyprotein processing.[42] Unlike the 3C protease that shares sequence similarity among different genera of picornavirus, 2A region show no significant sequence homology or processing manner. The 17 kDa 2A region from entero- and rhinoviruses,

structurally similar to the trypsin-like serine proteinases, is capable of proteolytic self-cleavage from the polyprotein precursor from the N-terminus.[43-44] The 2A region from cardiovirus (EMCV) possesses a similar size, but performs self-cleavage from the C-terminus.[45] In the case of foot-and-mouth disease virus (FMDV), cleavage of the 2A sequence also occurs at its C-terminus, however, it consists of only 16-20 amino acids.[46]

Study of the processing properties of FMDV revealed that cleavage of the polyprotein precursor is solely mediated by the short 2A oligopeptide region since the expected cleavage was still observed when different portions of flanking sequences were deleted.[42] When the 19 amino acid 2A oligopeptide sequence was inserted between two reporter proteins (Fig. 1.3), 90% of the polyproteins were cleaved in both *in vitro* transcription/translation system and transfected mammalian cells. Processing of this artificial FMDV polyprotein occurred in a manner similar to that observed for native FMDV proteins. Cleavage of the remaining 10% could not be achieved through prolonged incubation.[47] Furthermore, the upstream protein accumulated to a higher level compared to the downstream protein upon 2A peptide mediated polyprotein cleavage even though they were synthesized from a single open reading frame.[48-49] It was demonstrated that this imbalanced accumulation could not be attributed to the different degradation rates of protein products or RNA transcripts; instead, it was due to the different rates of protein synthesis. Molecular dynamic modeling of F2A peptide revealed that it adopts an amphipathic α-helical structure over most of the sequence, with a reverse turn at the C-terminal region. During protein translation, the amphipathic α-helical structure tightly interacts with the ribosome which constrains the conformational space of the tight turn at the C-terminal region of the 2A sequence, causing a subtle alteration of the ribosomal conformation. This conformational shift disfavors

peptide bond formation between the last two amino acid residues, glycine and proline, thereby releasing the nascent upstream protein from the ribosome in a manner that allows for continued translation of the downstream protein.[50-51] Previous mutation studies showed that serial restoration of the native upstream sequence of 2A dramatically improved cleavage efficiencies. Extension by 5 amino acids upstream of the minimum 2A sequence increased the activity to 96%, and complete cleavage was observed when extended by 14 amino acids.[51] More recently, Minskaia and co-workers showed that upon extension beyond 29 amino acids from the cleavage site, the cleavage activity became insensitive to the upstream protein.[52] These results were consistent with the previous atomic structural analysis that showed the ribosomal exit tunnel can accommodate 35 amino acids of the nascent peptide in a helical structure.[53] Therefore, the C-terminal region of the upstream protein that resides in the ribosomal exit tunnel appears to affect the ribosome-2A interaction and the subsequent cleavage activity of the 2A sequence.

Owing to the small size and effective autonomous C-terminal cleavage activity, various lengths of 2A peptides have been widely used as tools for coordinate expression of multiple proteins in the practice of biomedicine and biotechnology.[4-5,54] However, the 2A sequence which remains as the C-terminal extension of the processed upstream protein might hamper its function if the authentic C-terminus is critical for protein activity. Randall *et. al* reported an altered morphology of zein protein bodies when formed by β- and δ-zein processed from a polyprotein precursor using a 2A sequence, compared to their wild-type counterparts, which was thought to be caused by the altered β-zein tertiary structure due to the 2A overhang left on the C-terminus.[55] In addition, a remaining 2A overhang on the C-terminus of the processed upstream protein may lead to mistargeting of the secretory protein for degradation.[54,56-57] To eliminate the problems

generated by the C-terminal 2A overhang, a cleavable linker peptide from the naturally occurring polyprotein precursor found in the seeds of *I. balsamina* was incorporated between the upstream protein and 2A sequence when two different plant defensins were co-expressed in *A. thaliana*.[57-58] Removal of the C-terminal 2A overhang was achieved, however, the processed upstream protein failed to target to the designated location.[57] In another study, the furin substrate sequence was reported to successfully remove the C-terminal 2A extension in mammalian cells[59-60] but it was only applicable to proteins designated to the secretory pathway and only in a limited number of the host cells harboring furin protease activity. A more robust method for removing the 2A extension is demanded to eliminate the adverse effects caused by the C-terminal 2A extension.

1.5. Putative Protease Substrate Sequences

Aside from the co-translational protein processing mechanisms, simultaneous expression of multiple proteins in plants can also be achieved by exploiting specific protease activity. In this case, multiple proteins of interest were connected, in tandem, via a short substrate sequence that can be recognized by an endogenous or exogenous protease (Fig. 1.4 and Fig. 1.5). After protein translation, individual component proteins can be separated from the polyprotein precursor inside the cells upon proteolytic digestion. Urwin and coworkers co-expressed two nematode resistance protease inhibitors connected by a pea metallothionein propeptide sequence in *Arabidopsis*.[61] Although the constitutive proteins were efficiently released from the polyprotein precursor, inhibition effect on the growth of nematode was not detected in transgenic plants. The lack of detectable resistance activity was resulted from the malabsorption of the processed proteins sequestered in the membrane by the attached hydrophobic substrate sequence after proteolytic

cleavage. In another study, two antifungal proteins were coordinately expressed using a naturally occurring polyprotein linker sequence from *Impatiens balsamina* in *Arabidopsis*.[62] Efficient cleavage was observed using this method, however, the cellular location and corresponding proteases for polyprotein cleavage remained obscure. Thus, with this strategy, it is difficult to control the *in vivo* processing of polyprotein precursors, especially when differential subcellular targeting of multiple proteins is required. Moreover, this endogenous protease dependent polyprotein cleavage system is not applicable in the plant species lacking the required proteases.[4] A refinement to this strategy is to incorporate the protease required for polyprotein processing. Several studies have demonstrated the feasibility of this approach by co-expressing two proteins separated by tobacco potyviral NIa protease together with the requisite recognition sequence in cell-free systems or in tobacco plants.[63-64] While efficient polyprotein cleavage and subcellular targeting were achieved, low protein expression levels were observed. This energetic wasteful process by incorporating the required proteases increases the overall size of the polyprotein, posing the problem of proper protein folding, thus subsequently decreases the protein production yield. More recently, a ubiquitin-based polyprotein vector for co-expressing two proteins in yeast or tobacco was reported.[65] In this method, the proteins of interest were connected by a C-terminal motif of ubiquitin, followed by a full-length ubiquitin monomer. Following protein translation, the polyprotein precursor was processed accurately by cellular deubiquitinase, and the protein moieties were released at a similar level. However, the processed proteins from this ubiquitin-based polyprotein accumulated less than if proteins were expressed individually.

1.6. Objectives and Experimental Design

Currently, one of the major challenges facing the development of advanced protein biologics and next generation bioengineered crops is coordinated expression of multiple proteins in a way that is precise, efficient, and simple to implement. This technology has a myriad of applications in biotechnology and biomedicine including production of high value multi-domain pharmaceutical protein complexes, redirection of cellular metabolism pathways, elucidation of fundamental cellular processes, as well as improvement of stress/disease tolerance or nutritional values, and other important agronomic traits. The aim of this dissertation research is to fulfill this demand by developing effective polyprotein systems for efficient expression of multiple proteins in plants. These systems will serve as a platform for development of genetically modified organisms whose functions rely on coordinate expression of multiple protein components for biomedical or biotechnological applications.

The dissertation research seeks to: (1) exploit plant endogenous kex2p-like activity for polyprotein expression; (2) develop a polyprotein expression system based on the dual-intein autoprocessing domain; and (3) further enhance the self-cleaving efficiency of the polyprotein expression system by harnessing the synergy between an intein and a FMDV 2A sequence (IntF2A).

The approach described in Chapter 2 exploits the endogenous kex2p-like proteolytic activity in tobacco and a broad variety of other plant species to co-express multiple secretory proteins via a short linker comprising three copies of the putative kex2p substrate sequence.[66-69] Upon translation, the entire polyprotein precursor is targeted to the secretory pathway, cleaved in the

trans-Golgi compartment where the endogenous kex2p-like activity is located, and subsequently released the two protein constituents. Different from the notable TEV NIa approach, this kex2p polyprotein system utilizes endogenous host proteases. In this way the protein folding problem generated from incorporating a bulky exogenous protease as a part of the polyprotein can be eliminated. Processing of the kex2p based polyprotein in undifferentiated cell culture, whole tobacco plants as well as in different plant species were investigated and the processed proteins characterized.

With the *trans*-Golgi localization of plant kex2p-like proteases, it can be conceived that this kex2p based polyprotein is only applicable to expression of secretory proteins. Moreover, if the co-expressed protein components require being separated and assembled into an intact bioactive hetero-oligomer in the secretory pathway before *trans*-Golgi, the protein function would probably be impaired.

Therefore, in Chapter 3, an autoprocessing dual-intein fusion domain was developed for coordinate expression of multiple protein constituents that aims to resolve limitations of the kex2p based polyprotein system. This dual-intein fusion domain comprises an *Ssp* DnaE mini-intein N159A mutant and an *Ssp* DnaB mini-intein C1A mutant connected in tandem via a flexible peptide linker, to confer efficient N- and C-terminal cleavage activity, respectively.[70-72] Release of the constitutive proteins from the polyprotein precursor does not depend on host proteases or enzyme co-factors, but solely on the spontaneous cleavage activity of the intein mutants in the dual-intein fusion domain. Thus, different from the kex2p based polyprotein system, it can be used to co-express proteins designated to different cellular compartments. In addition, as inteins are widely distributed among all three domains of life, the utility of this novel

dual-intein based polyprotein expression system is not limited in plants, but can also be applied to a wide variety of expression hosts including yeast, mammals and prokaryotes. Application of the dual-intein polyprotein for coordinated expression of multiple proteins in *E. coli*, mammalian HEK293T, tobacco, as well as in different plant species were investigated using fluorescence reporters, thioredoxin (Trx), as well as chloramphenicol acetyltransferase (CAT). Different factors governing cleavage activity of the dual-intein fusion domain were investigated. Our results demonstrated efficient release of both upstream and downstream proteins in all tested hosts by the hyperactive dual-intein cleavage. However, C-terminal intein cleavage was less effective than N-terminal intein cleavage and showed temperature sensitive in *E. coli*. Moreover, the relatively large size of the dual-intein cassette (~36 kDa) might result in unfavorable folding and reduce the yield of the protein, especially when more than two proteins are to be co-expressed.

In Chapter 4, a smaller and more effective C-terminal cleavage domain, the FMDV 2A (F2A), was designed to replace the C-terminal cleaving intein mutant in the dual-intein fusion domain for simultaneous expression of multiple proteins. This IntF2A approach harnesses the synergy between the unique attributes of the N-terminal cleaving mini-intein mutant and the F2A peptide which is capable of unique ribosomal recoding. Upon transcription, encoded protein moieties are processed through a two-stage co-translational/post-translational auto-cleavage process that is applicable in a wide range of eukaryotes. By further optimizing the intein sequence, the authentic termini of released protein constituents can be preserved without excrescent peptide appendages (except for the first proline on the downstream protein as the result of F2A cleavage). Also, differential cellular targeting of the processed proteins can be achieved when suitable signal

peptides are incorporated. Our results showed that stoichiometric co-expression of two or three proteins was successfully achieved using the IntF2A processing domain in plants with rather high efficiency. These encouraging findings prompt us to further investigate different aspects of the IntF2A system, including, processing patterns in whole transgenic plants, the destination of the IntF2A autoprocessing fragment, and expression level of the target proteins, to assist in the further perfection of this promising system for coordinate multiple protein expression.

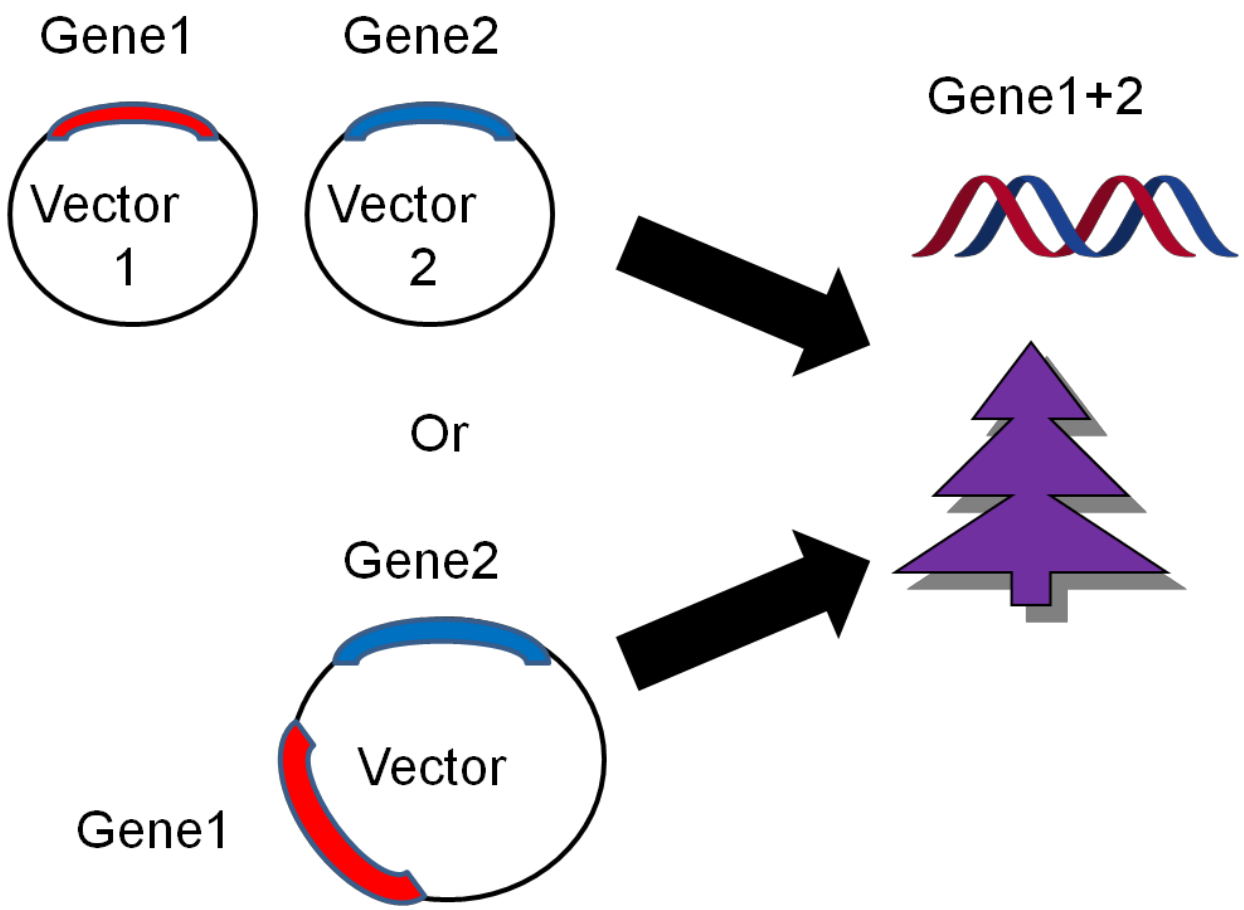


Figure 1.1. Gene stacking in plants via co-transformation of multiple monocistronic cassettes on the same or separated vectors.

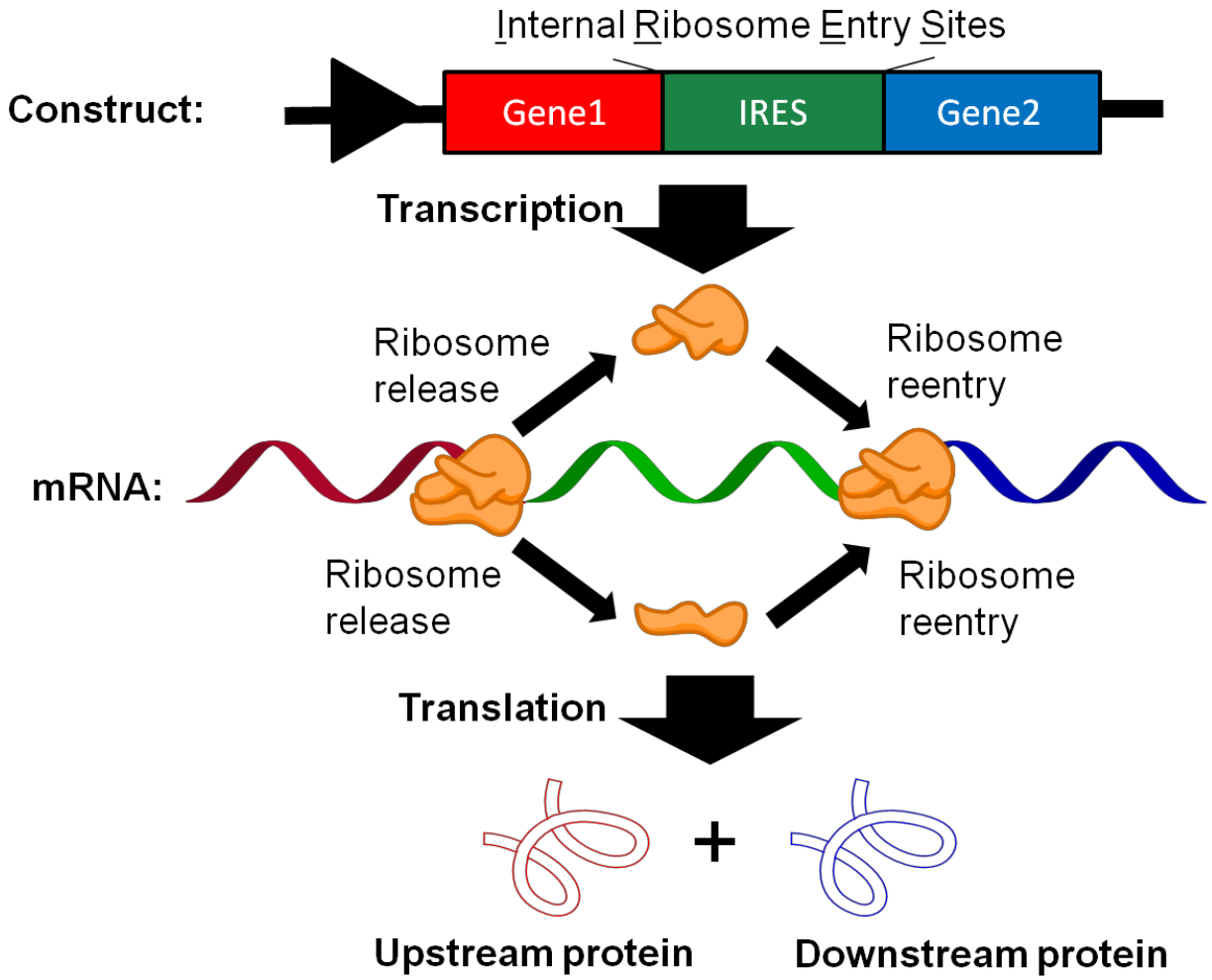


Figure 1.2. Gene stacking in plants via polycistronic vector mediated by Internal Ribosome Entry Sites (IRES).

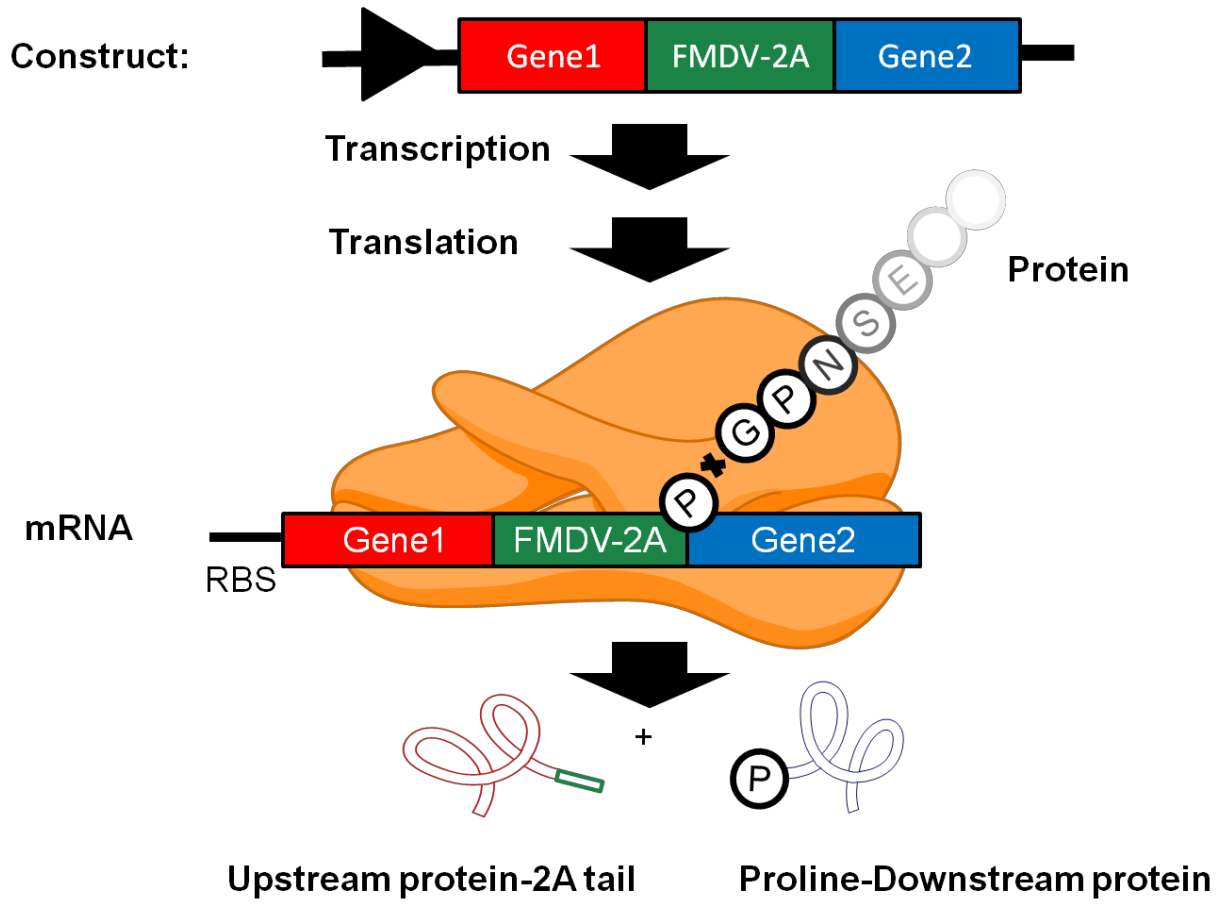


Figure 1.3. Gene stacking in plants via polyprotein cassette mediated by ribosomal recoding capable 2A sequence from Foot-and-Mouth Disease Virus (FMDV 2A).

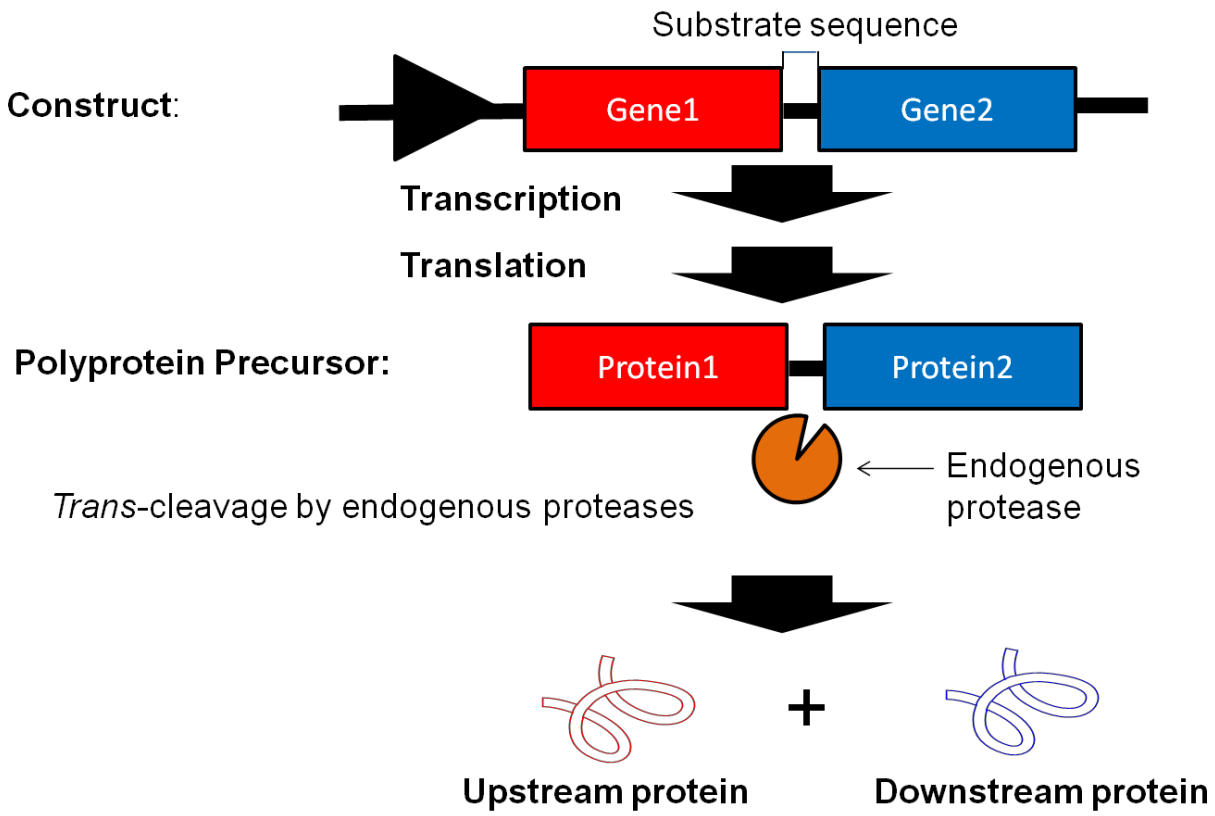


Figure 1.4. Coordinate gene stacking in plants by exploiting the endogenous plant protease activity. Multiple individual proteins are connected by putative protease substrate sequences in a polyprotein precursor. Individual proteins are separated by proteolytic cleavage of the intervening substrate sequence by the endogenous protease activity inside the cell

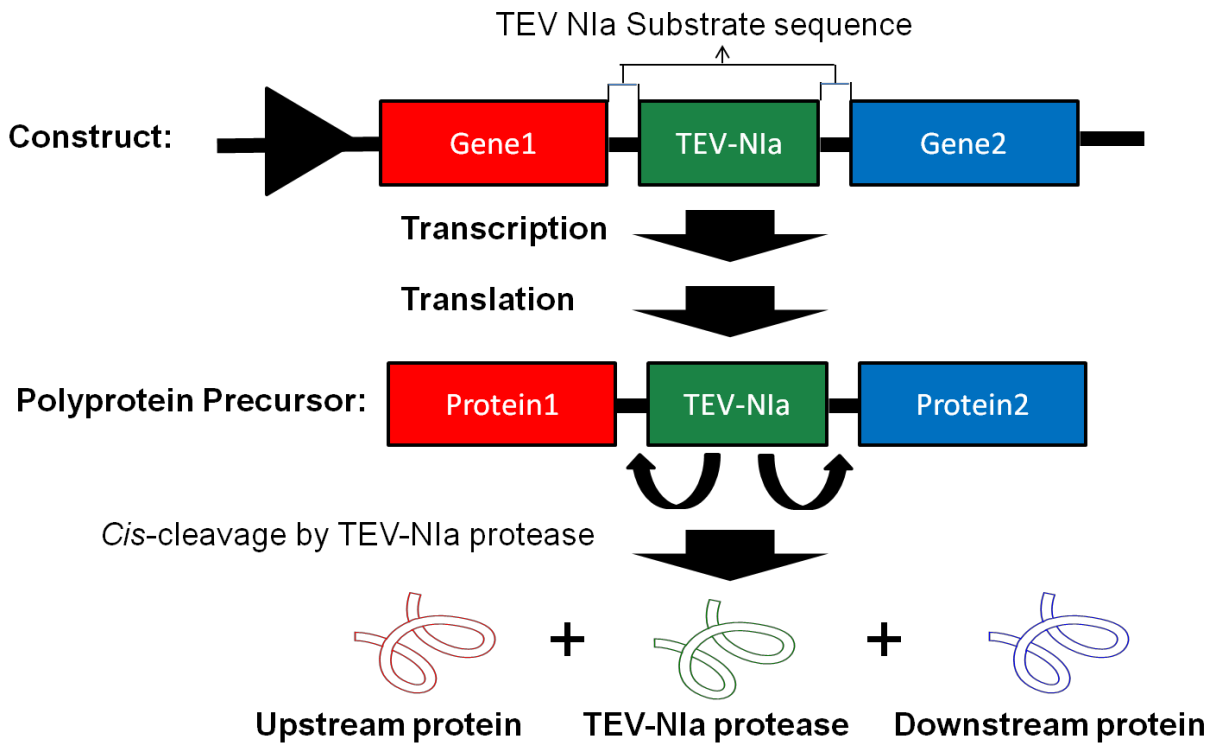


Figure 1.5. Coordinate gene stacking in plants via the built-in TEV-NIa protease. Multiple individual proteins, together with the TEV-NIa protease, are connected by the protease substrate sequences in a polyprotein precursor. After protein expression, individual proteins are released by the *cis*-acting proteolytic cleavage of the built-in TEV-NIa protease.

CHAPTER 2

COORDINATE EXPRESSION OF MULTIPLE PROTEINS IN PLANT CELLS BY EXPLOITING PLANT ENDOGENOUS KEX2P-LIKE PROTEASE ACTIVITY*

Abstract

Simultaneous expression of multiple proteins in plants finds ample applications. Here, we examined the biotechnological application of native kex2p-like protease activity in plants for coordinate expression of multiple secretory proteins from a single transgene encoding a cleavable polyprotein precursor. We expressed a secretory red fluorescent protein (DsRed) or human cytokine (GMCSF), fused to a downstream green fluorescent protein (GFP) by a linker containing putative recognition sites of the kex2p-like protease in tobacco cells, and referred to them as RKG and GKG cells, respectively. Our analyses showed that GFP is cleaved off the fusion proteins and secreted into the media by both RKG and GKG cells. The cleaved GFP product displayed the expected fluorescence characteristics. Using GFP immunoprecipitation and fluorescence analysis, the cleaved DsRed product in the RKG cells was found to be functional as

---

*Reproduced in part with permission from Zhang, B.; Rapolu, M.; Huang, L. and Su, W. W. Coordinate expression of multiple proteins in plant cells by exploiting endogenous kex2p-like protease activity. *Plant Biotechnology Journal* **2011**, *9*, 970-981. Copyright © 2011, Rights managed by Society for Experimental Biology and Association of Applied Biologists.

well. However, DsRed was not detected in the RKG culture medium, possibly due to its tetramer formation. Cleaved and biologically active GMCSF could also be detected in GKG cell extracts but secreted GMCSF was found to be only at a low level, likely because of instability of GMCSF protein in the medium. Processing of polyprotein precursors was observed to be similarly effective in tobacco leaf, stem, and root tissues. Importantly, we also demonstrated that, via agroinfiltration, polyprotein precursors can be efficiently processed in plant species other than tobacco. Collectively our results demonstrate the utility of native kex2p-like protease activity for the expression of multiple secretory proteins in plant cells using cleavable polyprotein precursors containing kex2p linker(s).

2.1. Introduction

Crops can be genetically manipulated to improve nutritional value, to resist diseases and stresses, and to produce high-value products such as secondary metabolites, therapeutic agents, vaccines, and industrial enzymes. First generation genetically modified crops involve simple traits (e.g., herbicide resistance) that call for introduction of a single gene (in addition to the selection marker gene). In order to introduce multiple traits or more complex traits, however, it often requires coordinated regulation of multiple endogenous genes or introduction of multiple transgenes. Two notable examples are the production of biodegradable plastics polyhydroxybutyrate in *Arabidopsis*[73] and the synthesis of β-carotene in "golden rice".[74] In producing multimeric proteins, it is also necessary to achieve similar levels of expression for the individual monomeric protein constituents, as in the development of plants producing recombinant antibodies, where co-expression of antibody heavy and light chains is required.[75]

As value-adding traits become ever more important to increase the profitability of agriculture and energy crops, the needs to produce multiple co-products from the same crop will continue to grow, which will most certainly benefit from a technology that enables efficient co-expression of multiple proteins.[65]

A handful of strategies have been reported for co-expression of multiple proteins in plants.[54,76-78] One promising strategy is to encode multiple proteins, connected in tandem, on a single open reading frame (ORF) in the form of polyproteins. Specific proteolytic processing of the polyprotein precursor *in vivo* leads to the release of the individual proteins. This approach enables coordinate expression through the use of a single promoter. Beachy's group demonstrated the feasibility of this approach by co-expressing two proteins separated by the tobacco etch virus (TEV) NIa protease recognition sequence (heptapeptide cleavage recognition sequence ENLYFQS) together with the Nla proteinase.[63,79] Dasgupta *et al*.[64] reported a similar polyprotein vector using a NIa proteinase from the tobacco vein mottling virus (TVMV) and demonstrated subcellular localization signals can be effectively recognized in the context of a polyprotein. The TEV NIa proteinase domain is about 48 kDa (while a truncated version is about 27 kDa) which adds to the overall size of the polyprotein. The large size of the polyprotein could cause protein misfolding and result in low protein production yield.[64,80] More recently, Walker and Vierstra described a modified ubiquitin (Ub)-based vector for co-expression of two reporter proteins LUC and GUS in tobacco.[65] While proper polyprotein processing was achieved using the Ub vector, the protein expression level was lower than those when the same protein was expressed alone. Walker and Vierstra suggested that the large size of the polyprotein (152 kDa) may have caused unfavorable folding and lead to low protein yield.[65]

In lieu of co-expressing a heterologous protease as a constituent of the polyprotein which is energetically wasteful and could complicate proper protein folding, it is plausible to utilize endogenous plant proteases to process the polyprotein precursor *in vivo* to release the individual protein elements. François *et al*. utilized linker peptide sequences originating from a natural polyprotein occurring in seeds of *Impatiens balsamina* in constructing polyprotein constructs for expressing two antifungal proteins in *Arabidopsis thaliana* and demonstrated post translational polyprotein precursor cleavage.[81-82] Urwin *et al.* linked two protease inhibitors via a propeptide derived from a pea metallothionein like protein and showed that the chimeric polyprotein precursor could be partially cleaved in *A. thaliana*.[61] The actual protease(s) that are responsible for the polyprotein precursors reported in these studies have yet to be identified. Another plausible strategy in this type of approach is to use linker sequences that are putative substrates of known endogenous plant proteases. An example to this end is to exploit the bacterial subtilisin or yeast kexin-like serine protease activities that have also been noted in plant systems. Yeast or mammalian kexins are type I integral membrane endopeptidases that reside in the *trans*-Golgi network.[83] The presence of kex2p-protease like activity in tobacco was reported by Bruenn's group by demonstrating the processing of a virally encoded antifungal preprotoxin in tobacco.[66] Subsequently Rogers' group showed that the substrate specificity of plant kex2p-like protease is similar to those of the fungal and yeast kex2p proteases.[67] Although no gene encoding plant kex2p proteases has been cloned from tobacco, expression of subtilisin-like serine proteases have been reported in a number of plant species including *Arabidopsis*[69] and tomato.[68] The findings from these fundamental research studies prompted us to examine the biotechnological

applications of kex2p-like activities in plants for expression of multiple secretory proteins from a single transgene encoding a cleavable polyprotein precursor.

In the present study, we created polyprotein constructs that encode a red fluorescent protein variant (DsRed), or a human cytokine (granulocyte macrophage colony stimulating factor; GMCSF), linked to a green fluorescent protein variant (GFP) by three copies of a putative recognition site of the kex2p-like protease, and characterized the expression and processing of the polyprotein precursors to determine the general applicability of the kex2p substrate linker in cleavable polyprotein precursors for co-expression of multiple proteins in plants. Here GMCSF serves as a model pharmaceutical protein and we have exploited the unique benefit of stoichiometric expression using the cleavable polyprotein constructs to estimate the production level of the processed GMCSF based on the GFP fluorescence without requiring direct covalent fusion of GMCSF to GFP as in conventional GFP fusion protein approaches. By encoding GMCSF and GFP connected via a kex2p-substrate linker in the polyprotein construct, it provides a novel alternative for monitoring protein expression using fluorescent protein fusion, and it is especially useful if permanent fusion hampers the protein function. [84]

2.2. Materials and Method

Construction of TTOSA1 RKG-HDEL viral vector and plant transfection.

A hybrid tobamoviral vector construct TTOSA1 RKG-HDEL encoding a fusion protein of DsRed, followed by a linker sequence, LEAGG(IGKRGK)$_3$EF, and a GFP with a HDEL (His-Asp-Glu-Leu) sequence at its C-terminal for ER-retention was constructed as follows. GFP coding region was amplified from pBIN-mgfp5ER[85] using a forward primer (mGFPS2E) and a

reverse primer (mGFPL242AE). The amplified 750-bp GFP fragment was digested with *EcoR*I and cloned into pLJ607 that contains the kex2p linker.[67] The linker-ligated GFP was reamplified using a forward primer (LJ607R4X) with a 5'-*Xho*I site and a reverse primer (mGFPL242AE) with an *Avr*II site and the codons for C-terminal ER-retention signal HDEL. DsRed sequence was amplified from pGDR[86] using a forward primer (DSRED2M1S) with *Sph*I site and a reverse primer (DSRED2HDELX) with *Xho*I site. The amplified fragments of DsRed and the linker-ligated GFP were digested with the respective restriction enzymes and ligated simultaneously into the viral vector TTOSA1-103SPEKCD43 that had been digested with *Sph*I and *Avr*II, behind the rice α-amylase signal peptide, to yield the plasmid TTOSA1-RKG-HDEL (Fig. 2.1). The viral vector was transformed into *E. coli* C600. *In vitro* RNA transcripts derived from TTOSA1-RKG-HDEL clone were mechanically inoculated onto *Nicotiana benthamiana* plants for systemic infection. Proteins extracted from the leaves 1 to 2 weeks after the infection were used for analysis.

Construction of plant binary expression vectors.

A binary construct pE1226-RKG was created from the TTOSA1-RKG-HDEL construct by eliminating the ER retention signal HDEL at the C-terminal of the fusion protein. The upstream DsRed along with the secretory signal peptide coding sequence was amplified from the TTOSA1-RKG-HDEL construct using a forward primer (OsAASP-F) containing a 5'-*Sal*I site and a reverse primer (DsRed-R) containing a 3'-*Xho*I site. Similarly, the downstream GFP along with the kex2p linker sequence was amplified by a forward primer (LJ607R4X) containing a 5'-*Xho*I site and a reverse primer (GFP-R) containing a 3'-*Xba*I site. The two fragments were then successively ligated into pBluescript SK- after digestion with the respective restriction enzymes

to make pBluescript-RKG. Upon confirming the DNA sequence, the entire polyprotein coding sequence was excised from pBluescript-RKG by digesting with *Sal*I and *Spe*I and ligated into pE1226 digested with the same enzymes to get pE1226-RKG. Expression of the RKG coding sequence is under the control of the chimeric mannopine/octopine synthase $(ocs)_3$/mas promoter[87] which gives equivalent or stronger expression than the commonly used CaMV 35S promoter.[87-88]

A direct fusion of GMCSF in tandem with GFP (without the linker) was expressed under the control of a double 35S promoter with an AMV enhancer[89] using a binary vector pBIN-GG (Figure 1). An *Arabidopsis* basic chitinase signal sequence was used to direct secretion of the fusion protein. Construction of the pBIN-GG vector is described elsewhere.[90] To construct the GKG expression vector, GMCSF along with the *Arabidopsis* signal peptide was amplified from pBIN-GG by PCR using a forward primer (*Sal*I-GMCSF-F) containing a *Sal* I site followed by the Kozak sequence, and a reverse primer (GMCSF-*Xho*I-R). The fragment was then digested with *Sal*I and *Xho*I, and moved into the pBluescript-RKG that was digested with the same restriction enzymes to make pBluescript-GKG. Finally the entire fusion sequence was moved into pE1226 to get pE1226-GKG. The construction of the secretory GFP vector (pBIN-GFP, Fig. 2.1) is described in previous publication.[91]

***Agrobacterium*-mediated plant stable transformation.**

The binary constructs were transformed into *Agrobacterium tumefaciens* LBA4404 containing a disarmed Ti plasmid pAL 4404 (Invitrogen, Grand Island, NY) by electroporation, and selected on YM medium (0.04% yeast extract, 1.0% mannitol, 1.7 mM NaCl, 0.8 mM $MgSO_4$, 2.2 mM K_2HPO_4) containing 100 mg/l kanamycin and 100 mg/l streptomycin. After

confirming by PCR, the transformant bacteria were used to infect tobacco (*N. tabacum* cv Xanthi) leaves. Leaf discs from three-week-old plants were inoculated with the transformant *Agrobacterium* in the presence of 0.2 mM acetosyringone, and placed on MS medium[92] containing 1 mg/l 6-benzyl amino purine (BAP) for shoot induction, 300 mg/l kanamycin for selection, and 300 mg/l cefotaxime to inhibit the bacterial growth. The resultant shoots were then transferred onto medium without the BAP for root growth, and then screened for the expression level of the recombinant protein by Western analyzes. Calli were initiated from stem sections of select high-expressing transgenic lines on MS media containing 3% sucrose, 1 mg/l 2,4 dichlorophenoxy acetic acid (2,4-D) and 0.1 mg/l kinetin. Cell suspension cultures were initiated from the callus tissues in the same medium.

Protein extraction.

Protein extraction was carried out essentially as reported earlier[90] with minor changes. Briefly, intracellular soluble proteins were extracted from leaves or calli by grinding in liquid nitrogen with an equal volume of an aqueous buffer containing 0.25 M sodium borate (pH 8.0), 0.25 M NaCl, 1% caffeine, 1% ascorbic acid, 1 mM DTT, 5 mM EDTA, and 1 mM PMSF. Extracellular fluids were concentrated from the spent medium (of suspension culture) or from the intercellular space of calli, using 10 kDa molecular weight cut-off centrifugal membrane filter (Millipore, Billerica, MA). All extract samples were prepared fresh and used immediately for subsequent western, chromatography, or bioassay experiments.

SDS-PAGE and western blot analysis.

Extracted proteins were separated on SDS-PAGE and blotted onto a PVDF membrane (Immobilon-P; Millipore, Billerica, MA). Protein-blot membranes were blocked and hybridized

with rabbit anti-GFP (Invitrogen, Grand Island, NY), rabbit anti-DsRed (Clontech, Mountain View, CA) or mouse anti-GMCSF antibody (R&D system, Minneapolis, MN) followed by goat anti-rabbit or goat anti-mouse alkaline phosphatase conjugated secondary antibodies (Southern Biotech, Birmingham, AL) and detected by the BCIP-NBT coupling reaction. Partial denaturing SDS-PAGE used for determining the oligomeric structure of the RKG polyprotein precursor and processed DsRed was based on the observation that DsRed remains as a tetramer in SDS-PAGE without sample boiling.[93] Slight modification to the method of Baird *et al*. was made.[93] Here, samples were mixed 4:1 with 5× SDS sample buffer (containing 10% β-mercaptoethanol) and immediately loaded to 10% SDS-PAGE gel without boiling. After electrophoresis, the samples were further analyzed by immunoblotting. In all western blots probed for GFP reported here, except in Fig. 2.3e and 2.7a in which a recombinant GFP protein from Clontech was used as standard, an *E. coli* expressed His_6-GFP-E3 fusion protein of about 30 kDa was used as GFP standard. The 21-amino acid E3 is a helical coil peptide motif.[94]

GFP purification for N-terminal amino acid sequencing.

Secreted GFP derived from liquid cultured tobacco cells expressing RKG or GFP alone[91] was purified for N-terminal amino acid sequencing. Spent media of liquid suspension cultures were separated from the cells by filtration through a nylon gauge (mesh size 10 μm) and clarified by adding ammonium sulfate to 30% saturation. The clarified medium was then concentrated with a 10 kDa molecular weight cut-off centrifugal filter device (Millipore, Billerica, MA) followed by SDS-PAGE separation and blotted onto a PVDF membrane. The candidate bands were excised from the membrane after staining with a Coomassie blue dye. The N-terminal amino acid sequencing was conducted using the Edman degradation technique on a Perkin Elmer Applied

Biosystems Procise 494 protein/peptide sequencer with an on-line Perkin Elmer Applied Biosystems Model 140C PTH Amino Acid Analyzer, performed by the Protein Facility at Iowa State University.

DsRed purification for N-terminal sequencing, MALDI-TOF MS, and ESI-TOFMS analysis.

DsRed was purified from RKG calli using a three-step column chromatography procedure. RKG callus extract was first subjected to copper affinity chromatography described by Deo and coworkers[95] with some modifications. Briefly, 0.1 M $CuSO_4$ was loaded onto a HiTrap Chelating HP column (GE Healthcare, Piscataway, NJ) connected to a Biologic Duo Flow chromatography system (BioRad, Hercules, CA). The column was washed with deionized water, followed by 20 mM sodium acetate (pH 4.0) buffer containing 1 M NaCl and equilibrated with a binding buffer (50 mM sodium phosphate containing 300 mM NaCl, 5 mM imidazole at pH 8). After loading the sample, the column was washed with the binding buffer and eluted with a linear gradient of imidazole from 5 to 100 mM in the binding buffer over 20 column volumes. Fluorescence fractions were concentrated and further separated on Shodex IEC DEAE-825 column (Showa Denko America, New York, NY) connected to a Shimadzu HPLC system equipped with a fluorescence detector. Prior to sample injection, the IEC column was pre-equilibrated with 20 mM Tris-Cl buffer (pH 8.0) (buffer A). After sample loading, the column was isocratically washed with a mixture of buffer A and the same buffer containing 500 mM NaCl (buffer B) at a volume ratio of 95:5 for 10 min, followed by elution with a linear gradient of buffers A and B from 95:5 to 50:50 over 45 min. A flow rate of 1 ml/min was used during the entire purification procedure.

After the ion-exchange separation, the DsRed fluorescent fractions were concentrated and separated on a BIOSEC S-2000 size exclusion column (Phenomenex, Torrance, CA) connected to a Shimadzu HPLC system equipped with a fluorescence detector, using a mobile phase of 20 mM sodium phosphate buffer (pH 7.0) containing 200 mM NaCl at a flow rate of 1 ml/min. The DsRed fluorescent fractions were concentrated and desalted with Zeba spin desalting columns (Thermo Scientific, Rockford, IL). N-terminal sequencing of the processed DsRed was performed using the same procedure as for GFP described above. The MALDI-TOF mass spectrometry was performed by the Protein Facility at Iowa State University using a Voyager DE-Pro MALDI-TOF mass spectrophotometer (Applied Biosystems, Grand Island, NY) operated at a linear mode. Electrospray Ionization (ESI) mass spectrometry was performed on an Agilent 6210 LC/MS-TOF system fitted with an ESI source operated in positive ion mode. MassHunter software was used for data acquisition. ESI scanning was performed with an m/z range from 150 to 3200. Samples were acidified in 1% formic acid for 10 min to disassociate DsRed tetramers into monomers before being applied to the mass spectrometer with a solvent consisting of 60% acetonitrile in water plus 0.1% formic acid.

Expression and purification of yeast kex2p for *in vitro* proteolytic digestion.

Pichia pastoris expressing *Saccharomyces cerevisiae* kex2p gene (Sckex2p) was a kind gift from professor Bernhard Hube of the Friedrich-Schiller-University, Jena, Germany. Expression of Sckex2p in *P. pastoris* and subsequent purification were performed essentially as described by Hube and coworkers.[96] Briefly, *P. pastoris* cells were grown in buffered minimal glycerol medium (BMGY) at 30°C overnight, then in buffered minimal methanol medium (BMMY) for 16 hr. The culture supernatant was then collected and concentrated with a 30 kDa molecular

weight cut-off centrifugal filter and desalted with a PD-10 column. The desalted crude enzyme sample was loaded onto a HiTrap ANX FF anion exchange column (GE Healthcare), rinsed with the 50mM BisTris buffer (pH 4.5) containing 10 mM NaCl and then eluted with the same buffer containing 100 mM NaCl. The purity of the recovered yeast kex2p protein was verified by SDS-PAGE. For *in vitro* proteolytic digestion, purified yeast kex2p was incubated with RKG and GKG protein extracts at room temperature for 1 hour. The enzymatic digestion was stopped by rapid heating to 100°C for 4 min in the SDS-PAGE loading buffer.

Separation of cleaved DsRed and RKG fusion protein.

RKG calli were extracted and loaded onto a Q-Sepharose HP cation exchange column (GE Healthcare, Piscataway, NJ) connected to a Biologic Duo Flow chromatography system (Bio-Rad), washed with 20 mM Tris-Cl (pH 8.0) and eluted with a linear gradient from 0 to 500 mM NaCl over 20 column volumes. Fractions with DsRed fluorescence were collected and further separated on a BIOSEC S-2000 size exclusion column (Phenomenex, Torrance, CA) connected to a Shimadzu HPLC system equipped with a fluorescence detector. The mobile phase was 50 mM sodium phosphate buffer (pH 6.8) and the flow rate was set at 1 ml/min. The corresponding molecular weights of peaks displaying DsRed fluorescence were calculated based on a calibration curve established with gel filtration standards (Bio-Rad, Hercules, CA). By combining the ion exchange and size exclusion chromatography, the uncleaved RKG protein could be separated from the cleaved DsRed and GFP products. In the experiments to confirm the functionality (fluorescence) of DsRed processed from the RKG fusion protein, uncleaved RKG fusion protein and processed GFP in the RKG protein extracts were removed by GFP immuno-

precipitation using Dynabeads® protein G (Invitrogen, Grand Island, NY) conjugated with anti-GFP antibody (Cell Sciences, Canton, MA) following the manufacturer's instructions.

Agroinfiltration.

Transient expression was carried out in leaf discs by vacuum-infiltrating the recombinant *Agrobacteria*.[97] Briefly, fresh streaks were made from the glycerol stocks of *Agrobacterium tumefaciens* C58C1 transformed with the respective binary vector on LB-agar plate containing 100 μg/ml of kanamycin, and 15 μg/ml of tetracyclin and incubated for about 48 hours at 28°C. A single colony was then inoculated into a starter culture of 5 ml of YEP medium (10 g/l yeast extract, 10 g/l peptone, and 5 g/l sodium chloride) containing 100 μg/ml of kanamycin, and 15 μg/ml of tetracyclin and incubated for about 48 hours at 28°C with shaking at 200 rpm. The starter culture was in turn inoculated into 250 ml of the same media containing the antibiotics and incubated under the same conditions until the optical density of the culture at 600 nm reached about 0.5. The *Agrobacterium* cells were then collected by centrifuging the cultures at 3000 × g for 5 min at room temperature and resuspended in 125 ml of sterile distilled water containing 200 μM acetosyringone and 0.01% Tween-20. Leaves from about 3-weeks old *Nicotiana tabacum*, or *N. benthamiana* plants grown in the laboratory, or *Lactuca sativa L. var. longifolia* (Romaine lettuce) purchased from a local market were used for the infiltration. Leaf discs of about 1.3 cm diameter were punched out using a cork-borer. Approximately 2 g of leaf discs were used per 100 ml of *Agrobacterium* suspension. 250-ml Erlenmeyer flasks with baffles were used with 125 ml of *Agrobacterium* suspension per transformation. The flasks were placed in a desiccator connected to vacuum. About 250 mbar vacuum was applied for 20 min with gentle shaking (150 rpm) of the desiccator. The vacuum was then quickly released to facilitate

efficient infusion of the bacteria into the tissue. The leaf discs were drained off of *Agrobacterium* suspension, and placed on sterile moistened filter paper in 10-cm Petri plates. The plates were sealed with paraffin film and incubated in the dark at 26°C for 3 to 6 days.

Cell proliferation assay of GMCSF.

The biological activity of GMCSF expressed in GKG calli was assayed using TF-1 cells[98]. Cryo-preserved stock culture of TF-1 cells (ATCC CRL-2003) was inoculated into RPMI-1640 medium supplemented with 10% fetal bovine serum, 5 ng/ml GMCSF and 1× Pencillin-Streptomycin, and incubated at 37°C under 5% CO_2. The cells were subcultured into fresh medium at least three passages before using for the assay. The assay procedure was essentially as described previously.[90] Briefly, TF-1 cells of at least 90% viability were collected by brief centrifugation and suspended in the above medium without GMCSF to a density of $2\text{-}3\times10^5$ cells/ml, and supplemented with different amounts of GKG callus extracts or RKG callus extract spiked with different amounts of pure GMCSF protein. The callus extraction buffer consisted of 20 mM sodum phosphate buffer (pH 7.4), 150 mM NaCl and 1 mM PMSF. 96-well culture plates were used with 100 μl of culture per well. After 44 hour of incubation, AQueous One Solution (Promega, Madison, WI) was added and the color change was monitored at 490 nm for 4 hours. All treatments were in triplicates.

2.3. Results

Design of polyprotein expression constructs.

In order to exploit the endogenous kex2p-like activity in plants as a general strategy for processing heterologous polyprotein precursors, we developed two binary expression vectors

(respectively termed RKG and GKG herein) each encodes a secretory signal peptide at the N-terminal, an upstream protein, followed by a linker that contains putative cleavage sites for the kex2p-like protease (the kex2p linker), and a downstream protein (Fig. 2.1). The upstream protein is DsRed or GMCSF, for RKG and GKG, respectively, whereas the downstream protein is GFP in both constructs. The rice α-amylase signal sequence and the *Arabidopsis* basic chitinase signal sequence were used in directing protein secretion for RKG and GKG, respectively. In both constructs, the kex2p linker sequence is LEAGG(IGKRGK)$_3$EF (amino acids Leu/Glu and Glu/Phe were incorporated in the linker as a result of introducing the *Xho*I and *Eco*RI cloning sites, respectively) and it contains three copies of a putative recognition site of the kex2p-like protease (i.e., IGKR) which was modeled after the naturally occurring kex2p-substrate sequence (IGKRGKRPR) in the *Ustilago maydis* KP6 preprotoxin. It was reported that when a gene encoding the KP6 preprotoxin was expressed in transgenic tobacco the preprotoxin was correctly processed into α and β subunits at the kex2p-cleavage site[66]. In a later study, a linker containing only a single kex2p cleavage site (either IMRKY or IGKRG) was used to link two reporter proteins – a mutated proaleurain and a truncated vacuolar sorting receptor BP-80, yet no cleavage of the linker was observed when the reporter construct was expressed in transgenic tobacco.[99] Subsequently the reporter construct was modified to incorporate a linker having three tandem copies of the IGKRG motifs to improve access of the peptide substrate by the kex2p-like proteases, and cleavage of the new linker was confirmed when the construct was expressed transiently in tobacco protoplasts.[67] The kex2p linker was examined in this study for its cleavability in stably transformed *N. tabacum* calli, suspension cells, and plant tissues, as well as in *Agrobacterium tumefaciens* infiltrated leaves of *Lactuca sativa L. var. longifolia* (Romaine

lettuce) and *N. benthamiana*. In addition, a construct similar to GKG but lacking the linker (i.e., GMCSF directly fused to the N-terminus of GFP, termed GG herein, Fig. 2.1)[90] was assembled to examine cellular processing of the fusion protein in the absence of the kex2p cleavage sites.

Over twenty independent RKG, GKG, or GG transgenic tobacco lines were developed and screened using PCR and GFP western blot. One line each was selected for initiating callus and cell suspension cultures. The RKG callus cells displayed bright GFP and DsRed fluorescence when examined using an epifluorescence microscope fitted with GFP and DsRed filter set, respectively (Fig. 2.2). Strong fluorescence was seen in the periphery of the cell and around the nucleus, indicating cell wall and perinuclear ER localization, which is similar to those observed in *Arabidopsis* and tobacco cells expressing a secretory GFP.[85,91] GFP fluorescence is also visible in GKG and GG cells. These results suggest *in planta* expression of the fusion protein constructs yielded properly folded proteins. More detailed characterizations were conducted subsequently to examine the cellular processing of the fusion proteins.

Processing of polyprotein precursors *in planta*.

Proteolytic cleavage of the RKG and GKG polyprotein precursors in *N. tabacum* was investigated using western blot analysis of protein extracts from calli, suspension cells, leaf, stem, and root tissues of transgenic plants, agroinfiltrated leave discs, as well as intercellular fluid from RKG and GKG calli or spent media from RKG and GKG cell suspension cultures. Antibodies against GFP, DsRed, and GMCSF were used in western blots to probe the protein processing. As shown in lane 1 of Figures 2.3a & c, a strong immunoreactive doublet corresponding to GFP (~27 kDa) cleaved from the polyprotein is clearly visible in both RKG and GKG callus extracts. The doublet likely indicates GFP with and without the kex2p-linker peptide. Some uncleaved

polyprotein precursors were still present in the RKG extracts (~55 kDa) but GKG processing appeared to be almost complete. It is known that kex2p activity is inhibited by PMSF and EDTA[96] which are included in our protein extraction buffer. Therefore any post-extraction cleavage by kex2p should be inhibited, and what is observed on western blots (e.g., Figs. 2.3 & 2.4) should reflect the kex2p-mediated processing *in vivo*.

In concentrated spent media of either RKG or GKG suspension cultures, a strong anti-GFP antibody reacting band matching the GFP molecular weight (~27 kDa) was also detected (Figs. 2.3a & c, lane 2). Concentrates of spent media from RKG and GKG cultures exhibited fluorescence and absorption spectra that are characteristic of GFP. Similar to the processed GFP, an immunoreactive band that matches the expected molecular size of the processed DsRed protein (~26 kDa) could be seen on the western blot of RKG callus extract probed with a anti-DsRed antibody, while a second, less intense band matching the molecular weight of the RKG fusion protein (~55 kDa) was somewhat visible (Fig. 2.3b, lane 1). The processed DsRed produced in the RKG cells (Fig. 2.3b, lane 1) has a slightly lower molecular weight than the bacteria-expressed DsRed standard which is known to migrate on SDS-PAGE at a higher than expected molecular weight.[100] The intensity of the immunoreactive bands recognized by the DsRed antibody was not as strong compared with the bands seen on the GFP western blots. This might be caused by the lower affinity of the DsRed antibody since we noted that DsRed standard at 100 ng generated a less intense band than that of GFP standard at just 50 ng (cf. Figs. 2.3a & b, lane 3). This also explains why almost no RKG fusion protein is visible in the anti-DsRed western blot (Fig. 2.3b, lane 1). Unlike GFP, however, no DsRed protein was detected in the spent media of RKG suspension cultures (Fig. 2.3b, lane 2). It is known that native DsRed is a

tetramer[93] and the tetrameric structure is required for its fluorescent function.[101] The lack of detectable DsRed in the spent media is therefore likely resulting from formation of tetramers that are confined to the cell wall/apoplastic space. In subsequent experiments we confirmed that the processed DsRed as well as the RKG fusion proteins indeed formed tetramers as to be discussed further in the next section.

For the GKG culture, both processed GFP and GMCSF could be detected in callus extract (Figs. 2.3c and d, lane 1). The processed GFP product exhibited expected molecular size, while for GMCSF, two immunoreactive bands (~18 kDa and ~28 kDa) were noted on the western blot (Fig. 2.3d, lane 1). A similar dual band pattern was also reported for GMCSF expressed in rice seeds.[102] The higher molecular weight GMCSF band might be caused by glycosylation or dimerization as reported in transgenic tobacco cell culture,[103] transgenic rice seeds,[102] and rice cell suspension culture[104] expressing human GMCSF. Whereas essentially no secreted GMCSF can be detected in the spent media of GKG cell culture, a putative secreted GMCSF is observed on the anti-GMCSF blot in the concentrated GKG callus apoplastic fluid sample, though at a level disproportionately lower than the amount of secreted GFP (cf. Figs. 2.3c & d, lane 2). The putative extracellular GMCSF product also displayed a size smaller than its intracellular counterpart. This, along with the very low level of extracellular accumulation, could be due to the high susceptibility of GMCSF to proteolytic degradation in tobacco cell suspension culture as previously reported.[104]

Since yeast kexin-like serine protease activities are likely to be responsible for the observed processing of the RKG and GKG polyprotein precursors in tobacco, and that yeast or mammalian kexins reside in the *trans*-Golgi network,[83] we have exploited the unique properties

of the H/KDEL receptor and ER-retention signal in order to gain further insights into the cellular location of the plant kex2p-like activity. We infected *N. benthamiana* leaves with an RNA-viral vector containing coding sequence for the RKG polyprotein tagged with an ER-retention signal (HDEL) at the C-terminal (i.e., the TTOSA1 RKG-HDEL viral vector, Fig. 2.1). The mesophyll tissues of the viral-infected leaves exhibited both green and red fluorescence with high intensity, indicating that the fusion protein was expressed in these tissues at high levels. However, western blot analyses with anti-GFP and anti-DsRed antibodies showed only the RKG fusion protein but no cleaved proteins of either GFP or DsRed (Figs. 2.3e and f). Since the HDEL receptor resides in the *cis*-Golgi and hence the fact that the RKG polyprotein cleavage did not occur supports the notion that the tobacco kex2p activity is localized in the secretory pathway past the *cis*-Golgi stage. This finding is consistent with that of Jiang and Rogers who reported kex2p-like activity was attenuated by treating with Brefeldin-A which is known to induce destruction of Golgi stacks.[67]

Tissue specificity of the kex2p-like activity on polyprotein processing was then analyzed in transgenic tobacco RKG and GKG plants. Leaf, stem, and root tissues from the transgenic RKG and GKG plants were collected and analyzed by western blot probed with GFP antibody. As shown in Figure 2.4, cleaved GFP along with some uncleaved polyprotein precursor can be detected in all tested tissues. While the extent of processing does not appear to differ much among the three types of tissues tested, the highest expression was seen in the root tissues for both RKG and GKG (note that equal total protein was loaded in each lane). This is likely due to the use of the $(ocs)_3$/mas promoter which was reported to give higher expression in roots compared to other parts of the plant.[105] It is worthy of noting that comparing with the root

tissues, expression and processing in the leaf tissue are also quite good (Fig. 2.4, lane 2). Also noted was that both young developing leaves as well as mature developed (fully expanded) leaves showed similar processing efficiency. Considering that leaf agroinfiltration has gained favor as a preferred industrial molecular farming platform,[106] the endogenous kex2p-like activity in leaf tissues could be exploited for multi-protein expression at a large scale.

In summary, the majority of the RKG and GKG polyprotein precursors expressed in *N. tabacum* was indeed cleaved *in vivo*, yielding GFP that was secreted and either DsRed or GMCSF for RKG and GKG cells, respectively, inside the cells. We further verified that the observed cellular cleavage of RKG and GKG fusion proteins was indeed due to the presence of the kex2p linker as discussed in the following section.

Characterization of proteins processed from the polyprotein precursors.

Using western blots we were able to demonstrate *in vivo* processing of the RKG and GKG proteins in tobacco cells. To determine whether the processed DsRed product was functional, we performed GFP immunoprecipitation (IP) to remove uncleaved RKG protein and processed GFP to allow analysis of the size and functionality of the processed DsRed product. After subjecting RKG callus extracts to immunoprecipitation with an anti-GFP antibody, almost all of the GFP fluorescence was lost (Fig. 2.5a; indicating removal of both RKG fusion protein and processed GFP product) but about two-thirds of DsRed fluorescence remained in the solution (Fig. 2.5b). These results indicate that the cleaved DsRed is functional (i.e., displays expected fluorescence characteristics). To investigate whether the RKG fusion protein and the released DsRed protein form tetramers, RKG callus extract was subjected to GFP IP, followed by DsRed western blot analyses with partial denaturing (non-boiling) SDS-PAGE which has been shown to preserve the

tetrameric structure of DsRed.[93] As shown in Figure 2.5c, processed DsRed migrated as a tetramer (lane 4; ~100 kDa) while the uncleaved RKG fusion protein also appeared to be a tetramer (lane 2 upper band; ~200 kDa) for unboiled samples. Upon sample boiling, the corresponding monomeric RKG and processed DsRed became visible (Fig. 2.5c, lanes 1 & 3).

Because in the kex2p linker-based polyprotein construct the protein moieties are translated from a single open reading frame and then released post-translationally by kex2p-like protease activity, the individual protein moieties should in principle be accumulated at approximately stoichiometric levels inside the cell, provided that they have similar protein stability. Using RKG cell extracts, we determined the concentrations of the processed DsRed and GFP based on fluorescence calibration curves established using pure fluorescent proteins. An approximately equimolar quantities of DsRed (monomer) and GFP were noted, demonstrating the feasibility to achieve stoichiometric equivalence of individual protein moieties using the kex2p-linker based polyprotein constructs.

Next, the biological activity of processed GMCSF from the GKG cells was determined using GMCSF-dependent human TF1 cell proliferation assay. Utilizing the concept of stoichiometric expression of the protein moieties encoded in the polyprotein construct, as demonstrated in the expression of DsRed and GFP in RKG, the GMCSF concentration in the GKG extract was estimated based on the GFP fluorescence of the GKG extract. Since *in vivo* GKG cleavage was found to be highly efficient (Figs. 2.3c & d), biological activity of the GKG extract should derive primarily from the processed GMCSF. Different amounts of GKG callus extract that gave equivalent GMCSF concentrations from 0 to 20 ng/ml were added to the growth medium of human TF1 cells and the extents of cell proliferation were measured and compared. To cancel out

potential effect from the tobacco callus extract, we used RKG extracts spiked with 0, 0.5, 1.0 and 2.0 ng/ml of purified recombinant GMCSF for comparison. The amount of the RKG extract used was chosen to match the total protein concentration of the GKG extract sample that gave 20 ng/ml of equivalent GMCSF. We obtained similar results when the background RKG extract protein amount was cut in half. As shown in Figure 2.6, increased TF1 proliferation (indicated by the increasing absorbance at 490 nm due to bioreduction of a tetrazolium compound and formation of colored formazan by live cells) with increasing GMCSF concentrations was noted for both GKG extract as well as RKG extract spiked with pure GMCSF. The growth stimulating effect tapered off at a GMCSF concentration exceeds about 5 ng/ml. This result demonstrates that the processed GMCSF produced in the GKG cells is biologically active. Furthermore, provided that the specific GMCSF activity in GKG extract is similar to that of the pure GMCSF spiked into the RKG extract, the GMCSF concentration estimation based on GFP fluorescence of the GKG extract turned out to be quite accurate, evident by the closeness of the two curves in Figure 2.6. At a GMCSF concentration of 2 ng/ml, the absorbance is differed by less than 10% between the GKG extract and the RKG extract spiked with GMCSF.

To demonstrate that the observed cellular cleavage of polyprotein precursor was indeed due to the presence of the kex2p linker, we transformed tobacco plants with a binary expression construct coding for GMCSF and GFP in direct tandem fusion, lacking the kex2p linker, but with a signal peptide for secretion (i.e., pBIN-GG, Fig. 2.1). As shown in Figure 2.7a, lane 2, no cleavage occurred in the direct fusion protein. In addition, no GFP or GMCSF-GFP (GG) fusion protein was detected in the spent culture media (Fig. 2.7a, lane 3). Note that the *Arabidopsis*

basic chitinase secretion signal used in pBIN-GG was also used in pE1226-GKG which did lead to secretion of the GFP-processing product (Fig. 2.3c, lane 2).

We have taken several approaches to examine the location of cleavage in the polyprotein precursor. First, we purified the secreted GFP from the spent media of RKG suspension culture (expressing pE1226-RKG) and analyzed its N-terminal amino acid sequence using Edman degradation. We found that the N-terminus of the processed and secreted GFP was short of five amino acids residues (GKEFS) from the anticipated linker cleavage site, i.e., at the C-terminus of IGKR (Appendix 2). Of the five missing amino acids, Ser is the first amino acid of native GFP (Appendix 2). We also purified and sequenced secreted GFP from cultured transgenic tobacco cells that express GFP alone with a secretory signal (i.e., pBIN-GFP, Fig. 2.1).[91] Interestingly, about half of the GFP population was also missing the N-terminal Ser residue exactly as the processed and secreted GFP product from the RKG cells, while the remaining population had its signal peptide removed at exactly the expected site, leaving two extra amino acids, Glu and Phe (resulting from introduction of the *Eco*RI cloning site), in front of the first GFP amino acid Ser (Appendix 2). Cellular processing of secreted GFP proteins expressed from pBIN-GFP vs. pE1226-RKG shared certain common features. For the secretory GFP derived from pBIN-GFP, the signal sequence is removed by the signal peptidase in the ER lumen, and the processed GFP travels through the secretory pathway and is destined for secretion. For the secretory GFP derived from pE1226-RKG, the RKG polyprotein precursor travels through the secretion pathway, cleaved post *cis*-Golgi, releasing the processed GFP which is then secreted out of the cells. The secreted GFP from both of these two systems exhibited removal of the N-terminal Ser residue. This finding suggests initial proteolytic cleavage by signal peptidase or kex2p-like

protease (for pBIN-GFP and pE1226-RKG, respectively) followed by trimming the exposed N-terminal peptide extension via additional proteases in the secretory pathway. Another piece of evidence that supports this notion is from the GFP western blots of RKG and GKG calli in which two distinctive immunoreactive bands around 27 kDa are always visible while the lower band matches the size of the GFP band secreted into the medium (Figs. 2.3a & c, lane 1 vs. lane 2). In another experiment, we digested partially purified polyprotein precursors from extracts of RKG suspension cells (by removing the cleaved DsRed and GFP products using Q-Sepharose and Size-exclusion chromatography) with purified recombinant *S. cerevisiae* Sckex2p from transgenic *P. pastoris* and analyzed the resulting extracts using western blot probed with GFP and DsRed antibodies. Intriguingly, the results (Figs. 2.7b and c) indicated that both *in vitro* (Sc-kex2p) processed GFP and DsRed were slightly larger than the corresponding *in planta* processed products in the RKG callus extracts (cf. lanes 2 and 3) . This result not only confirmed that the $(IGKRGK)_3$ motif is indeed a suitable substrate for kex2p, it also supported the above polyprotein cleavage hypothesis.

In addition to analyzing the downstream GFP, we purified the upstream protein DsRed from the RKG cells and analyzed it using mass spectrometry (MS). Purification of the processed DsRed from the RKG callus extract was achieved using copper affinity chromatography,[95] followed by DEAE ion exchange chromatography and size exclusion chromatography. The purity was over 80% based on SDS-PAGE analysis. This purified DsRed was then analyzed using N-terminal sequencing, MALDI-TOF MS, and ESI-TOF MS. According to N-terminal sequencing, the DsRed product released from the RKG polyprotein lacks the first three residues (Met, Pro, and Ser; cf. Appendix 2). The observed molecular mass of the processed DsRed

protein measured using MALDI-TOF MS and ESI-TOF MS are 25,700 and 25,694 Da, respectively. This mass matches that of a DsRed protein lacking the three native N-terminal residues and having two additional amino acids (Leu and Glu originated from the kex2p linker) at its carboxyl terminus (calculated mass is 25,699 Da; Appendix 2). This finding, along with those presented above, suggest initial proteolytic processing within the kex2p-substrate linker, and subsequent digestion of the peptide linker extension by additional proteases in the plant cell secretory pathway. There have been an increasing number of reports on proteolytic removal of a stretch of linear amino acids (such as affinity tags) fused to the termini of proteins targeted to the secretion pathway in some plant species.[107-108] It is postulated that any stretch of nonstructural and exposed peptide sequences not protected by a compact tertiary protein structure may be particularly vulnerable to proteolytic degradation especially in cells of *Solanaceous* species.[107]

Processing of polyprotein precursors in other plant species.

To examine whether plant endogenous kex2p activity could be used in plant species other than *N. tabacum* to process polyproteins, we conducted transient expression of RKG polyprotein in *L. sativa L. var. longifolia* (Romaine lettuce) and *N. benthamiana* via agroinfiltration and analyzed protein processing patterns by western blotting probed with anti-GFP antibody. As shown in Figure 2.8, processed GFP can be detected in all three plant species upon agroinfiltration. This result suggests potentially general utility of plant endogenous kex2p-like activity for expressing multiple proteins from a single transgene *in planta*.

2.4. Discussion

The purpose of this study was to exploit the plant endogenous kex2p-like protease activity in developing a strategy to co-express multiple proteins via *in vivo* proteolytic processing of a polyprotein precursor. While such strategy has been hinted in the patent and review literature,[82,109-111] no detailed study has yet been published that investigated this approach as a general strategy for multi-protein expression in plant systems. In two previous studies, endogenous kex2p-like pro-protein convertase activity was implicated in the processing of a preprotoxin and a proaleurain reporter in tobacco, and the protease activity was reported to exhibit substrate specificity characteristic of yeast kex2p.[66-67] Using a purified recombinant *S. cerevisiae* kex2p protease, we confirmed the $(IGKRGK)_3$ peptide sequence used in the present study is indeed a suitable substrate for yeast kex2p. We also demonstrated that this peptide sequence can be effectively cleaved *in planta,* releasing the proteins flanking the sequence. RKG and GKG polyprotein precursors are efficiently processed in tobacco cells into DsRed and GFP, and GMCSF and GFP, respectively. Furthermore, the processed GFP product in both constructs was successfully secreted into the culture medium. The processed DsRed product was found to be functional, though not detectable in the culture medium. The DsRed released from the RKG polyprotein formed tetramers that are likely confined within the cell wall, as penetration through the cell wall is known to be the limiting step for secretion of many recombinant proteins in plant cells.[112] The GMCSF processed from the GKG polyprotein precursor showed biological activity in supporting proliferation of the GMCSF-dependent TF1 cells in a dosage-dependent manner. Only very low amounts of GMCSF were detected in the

GKG callus apoplastic concentrate, which might result from the proteolytic degradation of secreted GMCSF, as very high protease activity in apoplasts and spent culture medium of transgenic tobacco cells has been reported and GMCSF is known to be prone to proteolytic degradation in this environment.[104,113]

The processing of the RKG and GKG polyprotein precursors *in planta* may involve sequential actions of plant endogenous kex2p-like endoproteases, followed by additional exoproteases that trim off peptide extensions originated from the cleaved kex2p linker. This conclusion was supported by the slightly smaller molecular size of *in planta* processed protein products vs. the *in vitro* yeast kex2p processed protein products derived from the RKG polyprotein precursor (Figs. 2.7b & c), along with the N-terminal amino acid sequence information that indicated no extra amino acid originated from the kex2p linker sequence is present at the amino terminus of the processed and secreted GFP produced by the RKG cells, and data from N-terminal sequencing and MS-based mass analysis that revealed the processed DsRed contains only two extra residues originated from the kex2p-linker (Leu and Glu) at its C-terminus (Appendix 2). When GFP alone was expressed and directed for secretion by using a signal peptide, about half of the secreted GFP was found to lose its N-terminal Ser residue exactly as the processed and secreted GFP product from the RKG cells. We also noted that on the GFP western blots of RKG and GKG callus extracts, two distinctive immunoreactive bands around 27 kDa are always visible while the lower band matches the size of the GFP band secreted into the medium (Fig. 2.3). The upper band in this case is likely processed GFP that still contains a peptide extension originated from the kex2p linker. All of these findings indicate initial proteolytic cleavage of the polyprotein precursors by kex2p-like protease and subsequent

proteolytic trimming of the exposed peptide extension by additional enzymes in the secretory pathway. In fact, a similar observation has been reported for polyprotein vectors that utilize an intervening linker sequence derived from the natural polyprotein present in the *Impatiens balsamina* seeds.[62] It was reported that initial cleavage of the *I. balsamina* linker occurred by the action of an endoprotease, and unidentified exoproteases were responsible for the subsequent trimming of the cleaved protein products.[62]

We demonstrated in this study that the kex2p-linker based polyprotein construct can be used to achieve nearly stoichiometric expression of the individual proteins encoded in the construct. This useful feature was exploited in this study to accurately estimate the titer of GMCSF from simple GFP fluorescence measurement in the GKG cell extracts. Using a HDEL ER-retention signal we proved that cleavage of polyprotein precursors occurs en route the secretory pathway in a cellular compartment beyond *cis*-Golgi. Processing of polyprotein precursors was observed to be similarly effective in tobacco leaf, stem, and root tissues. We also demonstrate, via agroinfiltration, that polyprotein precursors can be efficiently processed in plant species other than tobacco. These findings point to potential general application of kex2p-substrate linkers in production of multiple proteins from a single transgene in large-scale molecular farming via agroinfiltration. Taken together, the kex2p-linker based polyprotein constructs provide an effective tool that simplifies coordinate expression of multiple proteins in plant cells and enables new applications such as recombinant protein monitoring without requiring permanent reporter protein fusion.

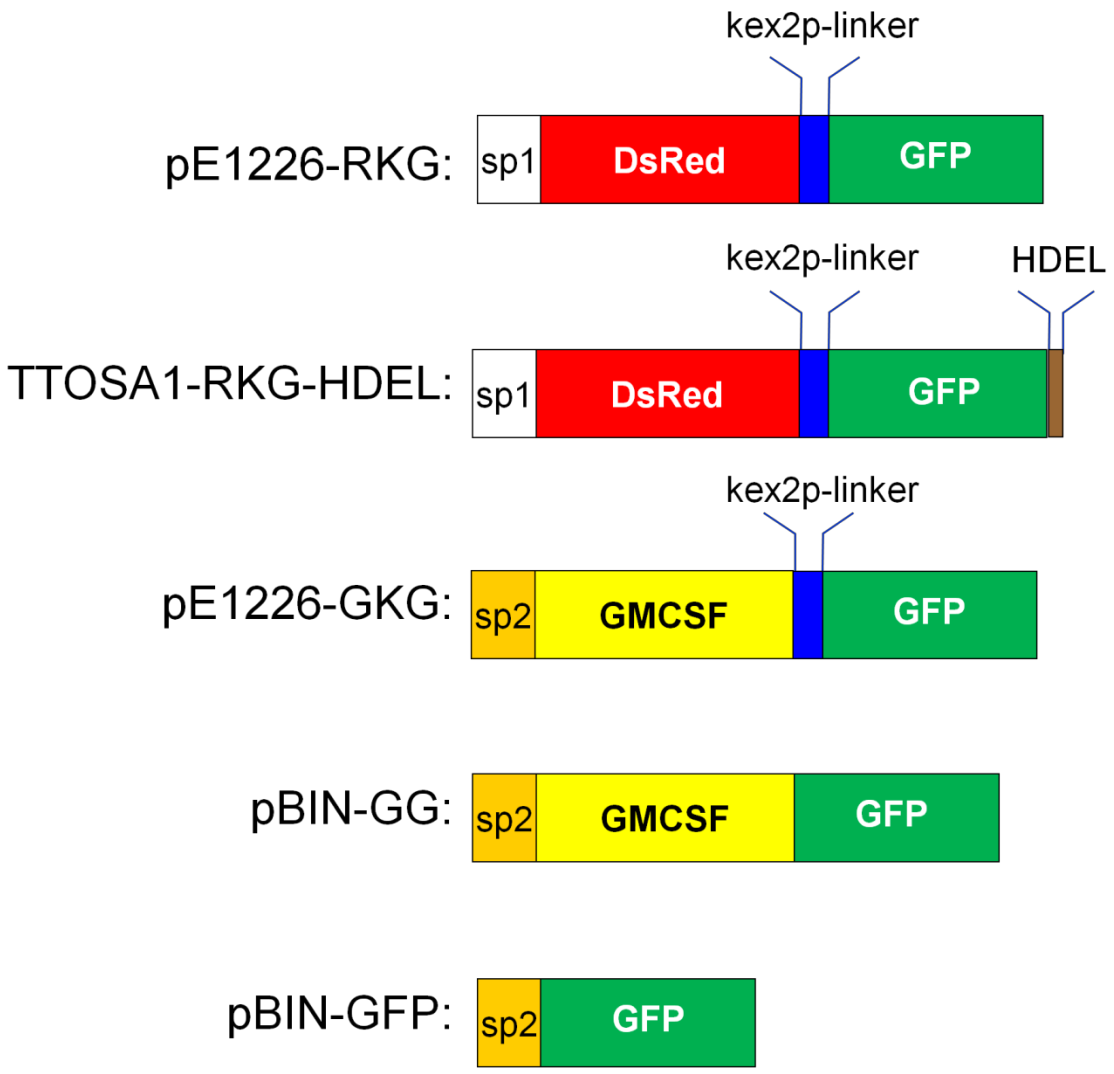


Figure 2.1. Organization of the different vector constructs used in this study. sp1: rice α-amylase signal peptide; sp2: *Arabidopsis* basic chitinase signal peptide.

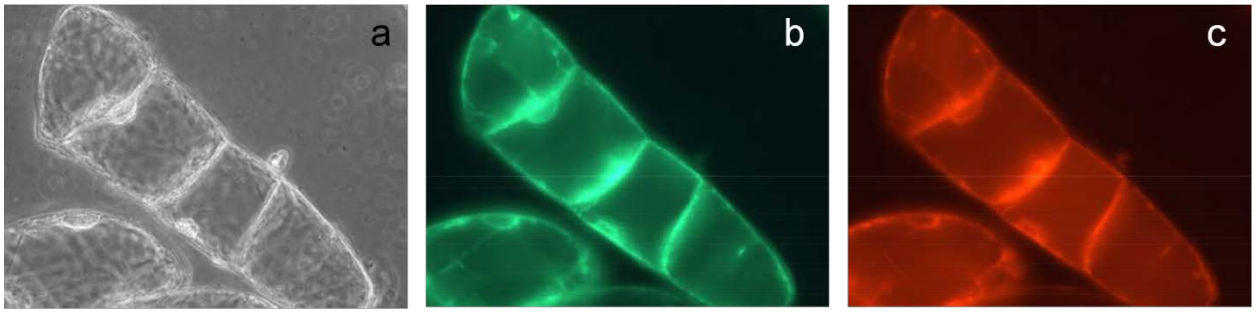


Figure 2.2. RKG cells display both green and red fluorescence. Phase-contrast image (a); fluorescence image using a GFP filter (b) or a DsRed filter (c) of the same cells.

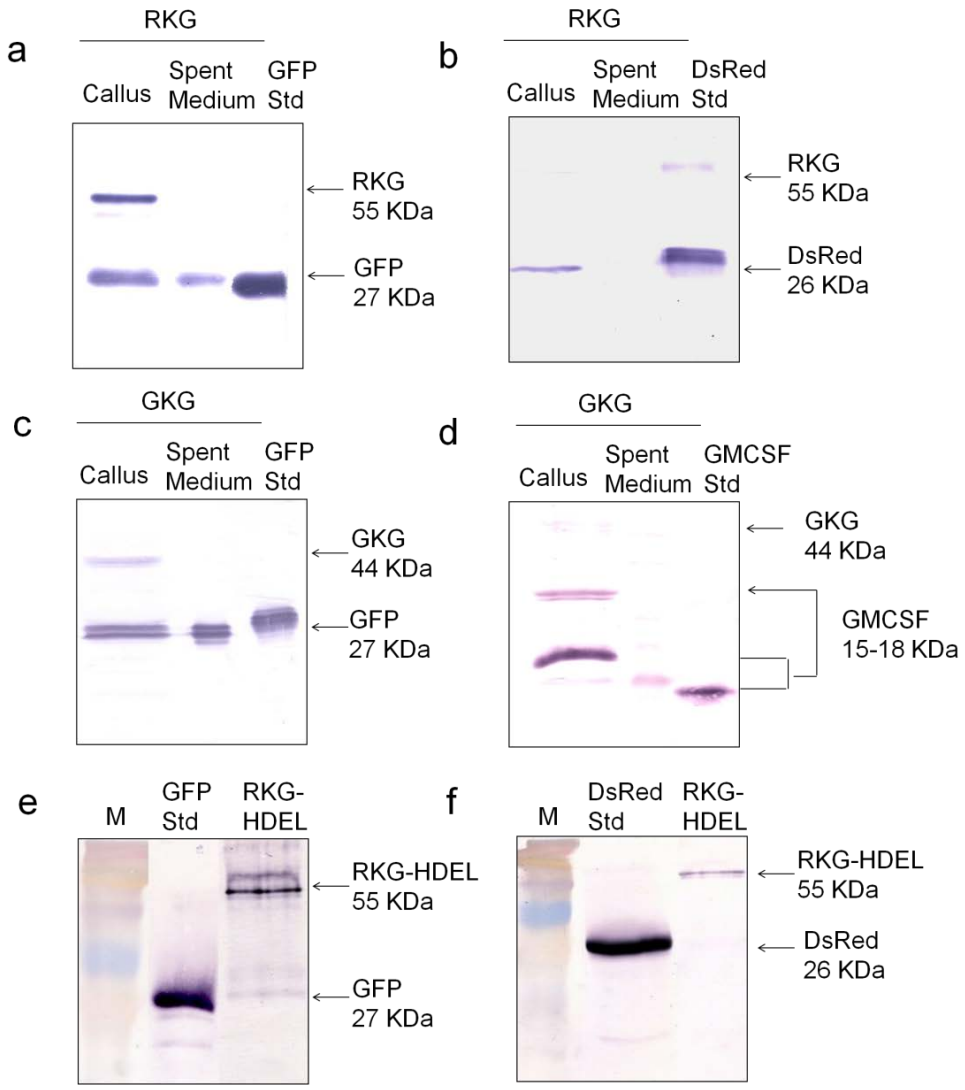


Figure 2.3. *In vivo* processing of RKG and GKG fusion proteins and secretion of the cleaved protein products. Proteins extracted or concentrated from the indicated samples were probed with anti-GFP antibody (a and c), anti-DsRed antibody (b), or anti-GMCSF antibody (d). Cleavage of polyprotein precursors via kex2p-like protease occurs *en route* the secretory pathway: western blot analysis of *N. benthamiana* leaves transfected with TTOSA1 RKG-HDEL viral vector probed with GFP antibody (e), or DsRed antibody (f) revealed no cleavage of the RKG-HDEL polyprotein.

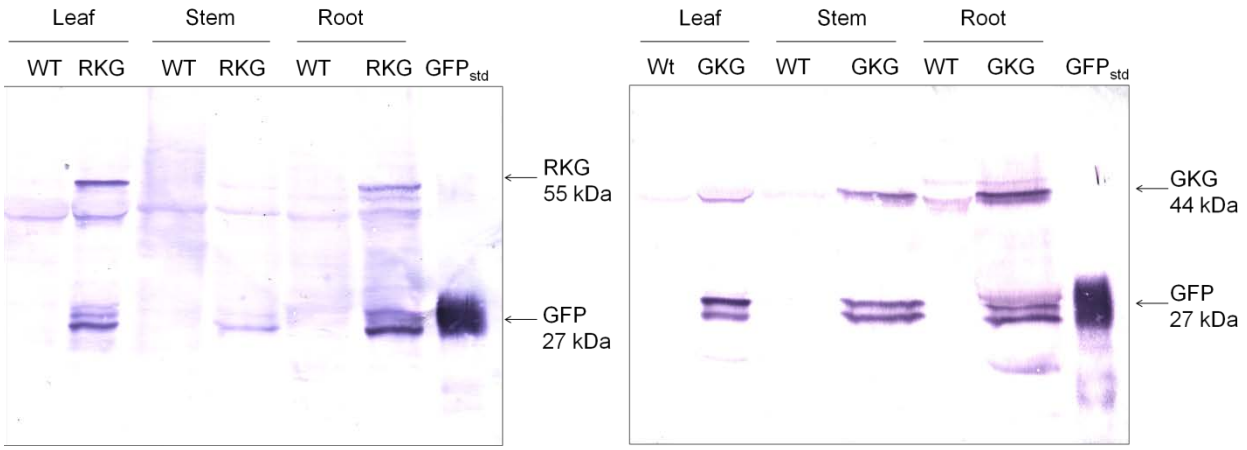


Figure 2.4. Processing of RKG and GKG fusion proteins in different tissues of tobacco plants. Equal amounts of soluble proteins extracted from leaf, stem, or root tissues of wild-type (Wt), RKG, or GKG tobacco plants were loaded for western blot analysis with anti-GFP antibody.

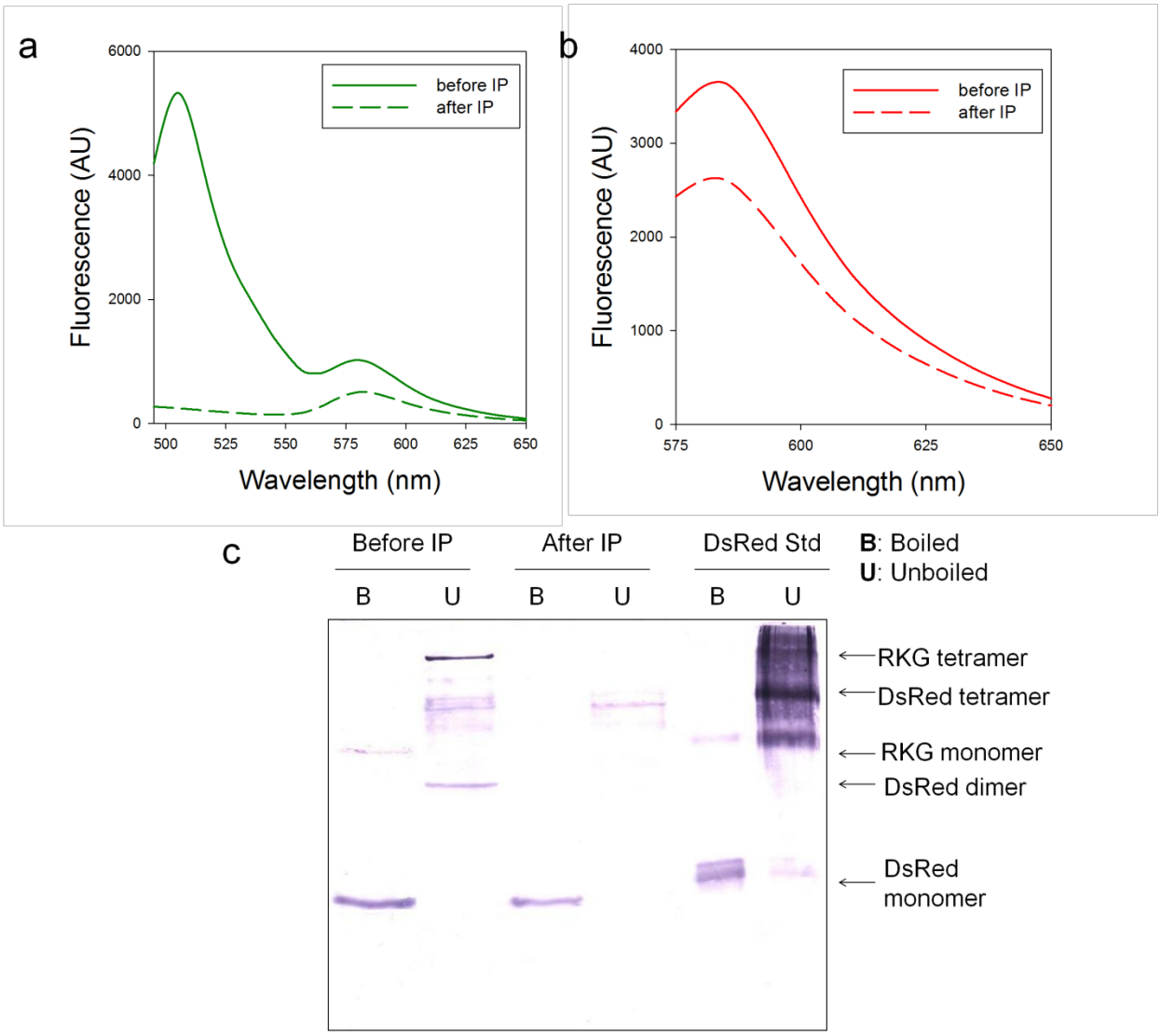


Figure 2.5. Analysis of RKG fusion protein and the cleaved protein products. GFP fluorescence with excitation at 470 nm **(**a**)**, and DsRed fluorescence with excitation at 556 nm (b) of RKG cell extract before, or after immunoprecipitation (IP) with anti-GFP antibody. (c) Western blot of RKG extract before, or after the IP, under complete (boiled) or partial (unboiled) denaturation, as indicated. The samples were subjected to SDS-PAGE, blotted onto a membrane and detected with anti-GFP antibody.

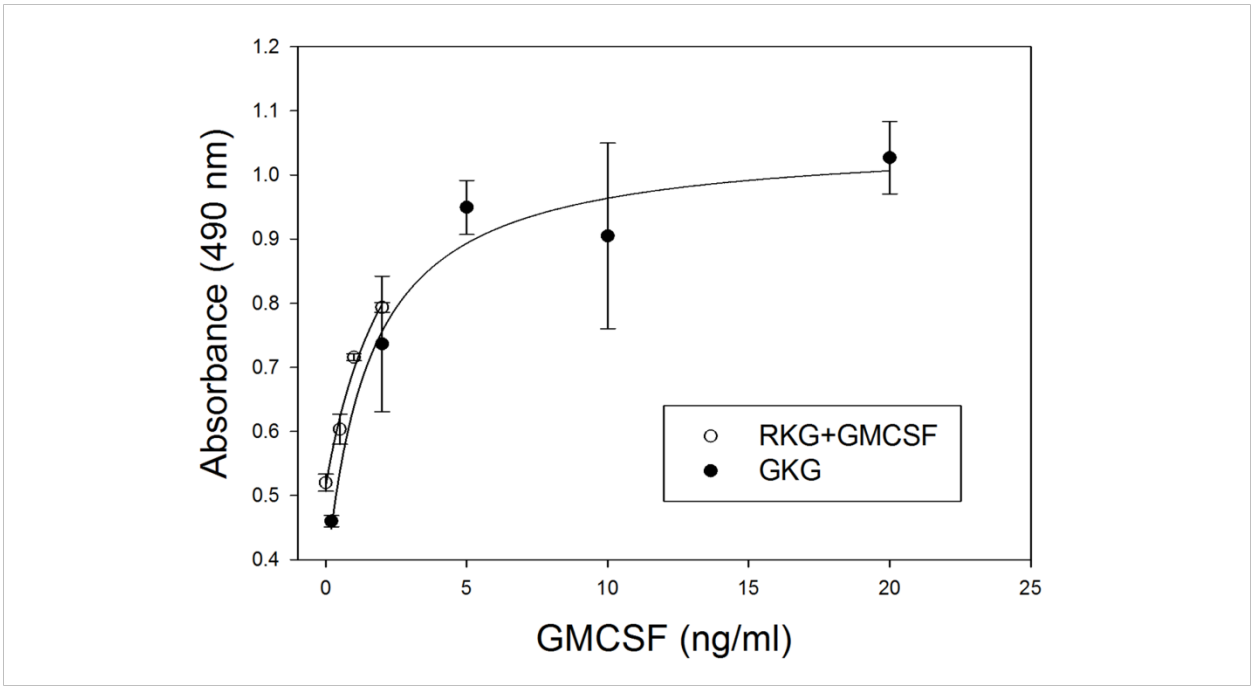


Figure 2.6. Bioactivity of GMCSF processed from GKG fusion protein produced by the GKG cells. GMCSF-dependent TF1 cells were cultured with different amounts of GKG callus extract to give the indicated concentrations of GMCSF. The amount of GMCSF in the extract was calculated from the fluorescence value of GFP. A stoichiometric ratio of 1:1 of GMCSF to GFP was assumed. To normalize any effect from the plant cell extract, pure GMCSF was spiked into RKG extract for comparison. Higher absorbance indicates higher cell proliferation.

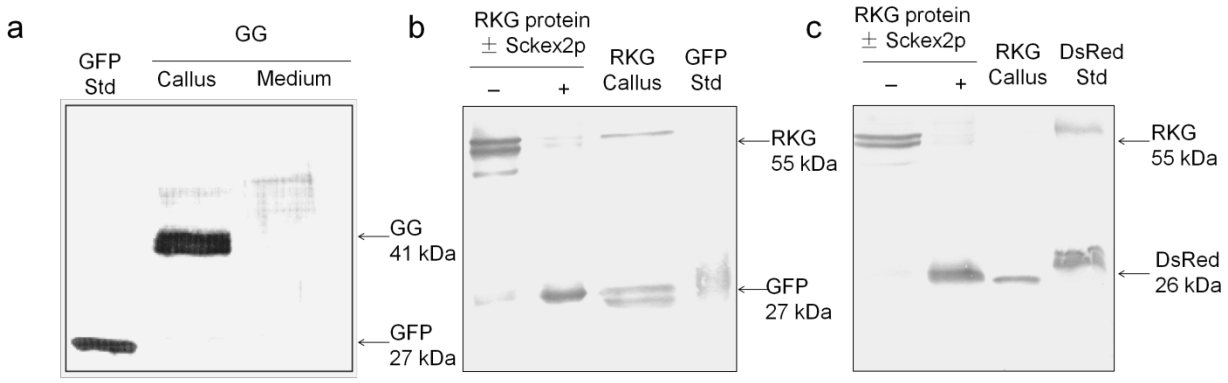


Figure 2.7. Presence of the kex2p linker sequence is necessary for polyprotein cleavage *in planta*. (a): Western blot of GMCSF-GFP direct fusion protein (GG), without the kex2p linker, expressed in tobacco cells and probed with anti-GFP antibody. (b & c): RKG fusion protein was purified from RKG tobacco cells, treated (+), or not (-), with yeast kex2p (Sckex2p) *in vitro* and analyzed by western probing with anti-GFP antibody (b), or anti DsRed antibody (c); for comparison, RKG callus extract was also probed with the antibodies to show cleavage of the RKG fusion protein *in-planta*.

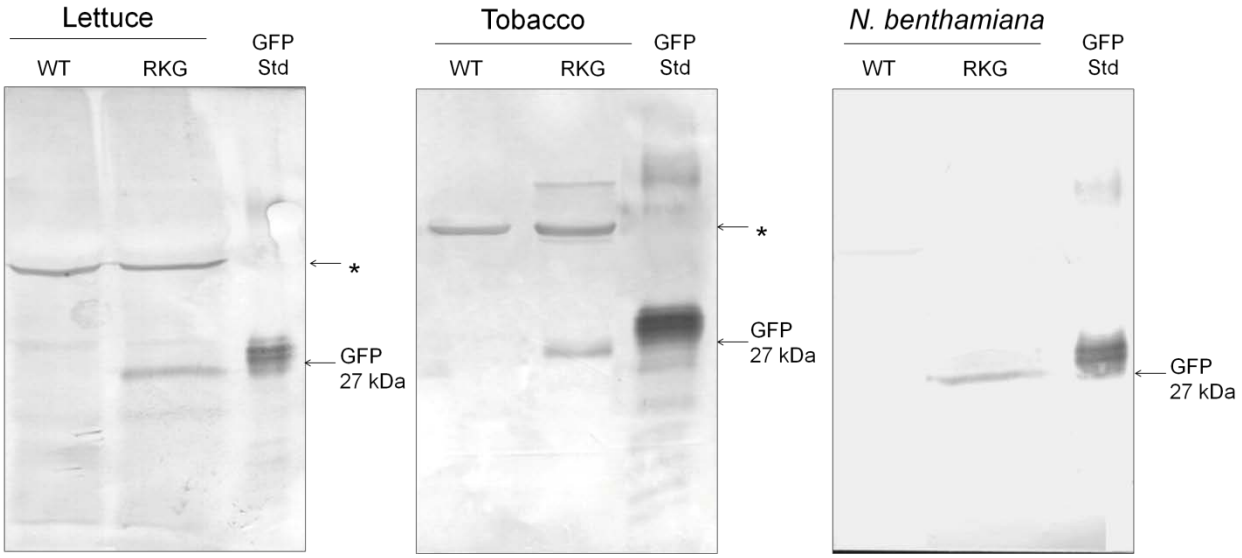


Figure 2.8. Cleavage of RKG fusion protein in different plant species. Leaves of lettuce, tobacco, or *N. benthamiana* were infiltrated with *Agrobacterium* transformed with the pE1226-RKG construct, or empty vector plasmid for control. Equal amounts of soluble proteins extracted from these leaves were analyzed by Western blot probing with anti GFP antibody. "*" indicates nonspecific band.

CHAPTER 3

A DUAL-INTEIN AUTOPROCESSING DOMAIN THAT DIRECTS SYNCHRONIZED PROTEIN CO-EXPRESSION IN BOTH PROKARYOTES AND EUKARYOTES*

Abstract

Being able to coordinate co-expression of multiple proteins is necessary for a variety of important applications such as assembly of protein complexes, trait stacking, and metabolic engineering. Currently only few options are available for multiple recombinant protein co-expression, and most of them are not applicable to both prokaryotic and eukaryotic hosts. Here, we report a new polyprotein vector system that is based on a pair of self-excising mini-inteins fused in tandem, termed the dual-intein (DI) domain, to achieve synchronized co-expression of multiple proteins. The DI domain comprises an *Ssp* DnaE mini-intein N159A mutant and an *Ssp* DnaB mini-intein C1A mutant connected in tandem by a peptide linker to mediate efficient release of the flanking proteins via autocatalytic cleavage. Essentially complete release of

---

*Reproduced in part with permission from Zhang, B.; Rapolu, M.; Liang, Z.; Han, Z.; Williams, P.G. and Su, W.W. A dual-intein autoprocessing domain that directs synchronized protein co-expression in both prokaryotes and eukaryotes. *Scientific Reports* **2015**, *5*, 8541. Copyright © 2015, Rights managed by Nature Publishing Group.

constituent proteins, GFP and RFP (mCherry), from a polyprotein precursor, in bacterial, mammalian, and plant hosts was demonstrated. In addition, successful co-expression of GFP with chloramphenicol acetyltransferase, and thioredoxin with RFP, respectively, further substantiates the general applicability of the DI polyprotein system. Collectively, our results demonstrate the DI-based polyprotein technology as a highly valuable addition to the molecular toolbox for multi-protein co-expression which finds vast applications in biotechnology, biosciences, and biomedicine.

3.1. Introduction

Intein is a naturally found protein splicing element that undergoes self-excision from a protein precursor with concomitant joining (splicing) of the flanking protein sequences called exteins. This protein splicing process is autocatalytic and does not require the presence of any exogenous host-specific proteases or co-factors. The discovery of intein has enabled a wide variety of innovative applications including enzyme activation, protein purification, expressed protein ligation, and production of cyclic proteins.[114-117] In addition to being an effective splicing element, inteins can be engineered to block the splicing activity and promote auto-excision at their N- or C- terminal, by mutating the crucial terminal residues, as well as the extein residues immediately flanking the intein sequence. Here we exploit the hyperactive intein-mediated self cleavage to develop a system for coordinating co-expression of multiple proteins from a single open reading frame (ORF).

Simultaneous and balanced co-expression of multiple proteins enables many important applications in biochemistry, life sciences, and biotechnology. Some notable examples are: (*i*)

production of multimeric pharmaceutical proteins such as antibodies and multi-subunit vaccines for disease diagnosis and treatment;[118-120] (*ii*) synthesis of protein complexes for biochemical, biophysical, and structural studies;[121] (*iii*) modification of metabolic pathways;[122] (*iv*) elucidation of how complex metabolic networks are organized and regulated;[123] as well as (*v*) engineering of crops with desirable agronomic traits.[5]

Currently three strategies are most commonly practiced for co-expression of multiple proteins: (*i*) multiple monocistronic expression cassettes on the same or separate vectors; (*ii*) polycistronic vectors based on ribosome binding site (RBS) or internal ribosomal entry site (IRES) for prokaryotic and eukaryotic hosts, respectively; and (*iii*) polyprotein vectors based on foot-and-mouth-disease-virus (FMDV) 2A-like peptide or protease substrate sequence for eukaryotic hosts.[77,124] With respect to the use of multiple monocistronic cassettes for co-expression, control over stoichiometric expression of the multiple transgenes cannot be readily achieved even after extensive promoter tuning. Also, driving the expression of each transgene cassette using the same type of promoter does not necessarily result in the same level of expression for each transgene.[125] When linked transgene expression cassettes are used on a single vector, because each gene cassette harbors its own set of regulatory elements, the size of the vector could become quite large especially if the transgene number increases. Such large vector size could substantially attenuate the transformation/transfection efficiency. As for the IRES-mediated translational initiation, it is known to be far less efficient compared with that of the 5'-cap mediated initiation. This leads to highly uneven co-expression, and the efficiency is shown to be cell-type specific.[125] Regarding the FMDV-2A or 2A-like viral peptide sequences, efficacy of protein co-expression is affected by the peptide sequence or protein structure immediately

upstream of the 2A motif in a manner that is poorly understood.[52,126-127] While many examples of successful 2A-mediated co-expression have been reported, there are also cases where a considerable portion of the polyprotein fails to separate into its constituent proteins.[126] Upon co-translational release, the 2A sequence remained on the carboxyl terminus of its upstream protein could cause erroneous subcellular targeting, or may interfere with the folding and function of certain target proteins.[57,77] Also notably, 2A- and IRES- based systems only work in eukaryotes.

Owing to the ease of genetic manipulation and ability to produce a large amount of recombinant proteins in a short period of time, *Escherichia coli* is commonly used as a test bed for functional studies of novel engineered proteins or protein complexes. Ultimate validation of the biological function of the protein of interest in a different host such as plant or mammalian cells would be necessary when the primary goal is to use or test the function of the protein in these hosts. It is therefore necessary to establish an efficient multi-protein co-expression platform that can be readily adopted in different hosts.

Here, we report a new polyprotein vector system that is based on a pair of self-excising mini-inteins fused in tandem, termed the dual-intein (DI) domain, to achieve coordinated co-expression of multiple proteins. The DI fusion domain comprises an *Ssp* DnaE mini-intein N159A mutant and an *Ssp* DnaB mini-intein C1A mutant connected in tandem by a peptide linker. This unique fusion domain harnesses the synergy between the N- and C- terminal auto-cleaving activity of DnaE (N159A) and DnaB (C1A) mini-intein variant, respectively, to mediate autocatalytic release of the flanking proteins *in vivo* (Fig. 3.1a). As intein autoprocessing can proceed efficiently in both prokaryotes and eukaryotes, we envisioned that the DI based polyprotein system can be applicable to a variety of expression hosts. In this study, we

investigated the DI system in bacterial, mammalian, and plant hosts, and confirmed its usefulness in all host systems tested.

3.2. Materials and Method

Vector construction.

The DI polyprotein cassettes, GFP_{172}-DI-RFP_{Strep}, GFP_{172}-DI-CAT_{Strep}, and Trx_{His}-DI-RFP_{Strep} were cloned into different vectors for expression in *E. coli*, mammalian HEK293T, and tobacco NT1 cells, respectively. To assemble the GFP_{172}-DI-RFP_{Strep} cassette, first the coding sequence of *Ssp* DnaE mini-intein (DnaE) N159A variant flanked by five bordering N-extein residues (KFAND) reported to accelerate N-terminal cleavage[128] and three native C-extein (CFN) residues was synthesized by Genscript (Piscataway, NJ) and ligated into pUC57 between *Xho*I and *Xba*I sites to yield pUC-DnaE(A_{159}CFN) vector. The coding sequence for GFP_{172} (a modified mGFP5 with six histidine residues inserted between amino acid residues 172 and 173)[129] was amplified from pGEM5z-GFP172 incorporating *Sal*I and *Xho*I sites by forward (*Sal*I-mGFP5) and reverse primers (GFP-*Xho*I-R), respectively. mCherry coding region was amplified from pETMD[130] to incorporate a 5'- *Apa*I site and a Strep tag followed by a *Sac*I site at the 3'-end by two step overlap PCR with the forward primer (mCherry-2A-F1) and reverse primers (mCherry-Strep and GKZ-R). C-terminal cleaving *Ssp* DnaB mini-intein (DnaB) with three bordering native N-extein residues (ESG) and C-terminal cleavage accelerating C-extein residues (SR)[131] was amplified from pTWIN1 (New England Biolab, Ipswich, MA) with primer pairs (DnaB-*Xba*I-F/DnaB-SRGP-R) incorporating *Xba*I and *Apa*I flanking the 5'- and 3'- junctions, respectively. PCR fragments of GFP_{172}, DnaB mutant and mCherry-Strep tag were successively ligated into

pUC57-DnaE backbone after digestion with respective restriction enzymes to yield pUC-GFP_{172}-DnaE-DnaB-mCherry-strep (called pUC-GFP_{172}-DI-RFP_{Strep} hereinafter). To replace GFP_{172} with GFP_{H6} (mGFP5 with a C-terminal hexa His tag) in pUC-GFP_{172}-DI-RFP_{Strep}, GFP_{H6} coding sequence was amplified from pBISN1-mGFP5-ER using primers GFP-*Kpn*I-F and mGFP5-His-*Xho*I-R to incorporate *Kpn*I and *Xho*I sites at the 5'- and 3'-ends, respectively, and replaced the *Kpn*I-GFP_{172}-*Xho*I fragment in pUC-GFP_{172}-DI-RFP_{Strep} to yield pUC-GFP_{H6}-DI-RFP_{Strep}. To assemble a vector for expressing thioredoxin-DI-mCherry (Trx_{His}-DI-RFP_{Strep}) in *E. coli*, the DI-RFP_{Strep} fragment from pUC-GFP_{172}-DI-RFP_{Strep} was amplified using PCR primers *Kpn*I-DnaE-F and GKZ-R to incorporate *Kpn*I and *Sac*I sites at the 5'- and 3'-ends, respectively. The PCR fragment was digested with *Kpn*I/*Sac*I and ligated into *Kpn*I/*Sac*I digested pET32a vector for in-frame fusion with an upstream thioredoxin sequence to create pET-Trx_{His}-DI-RFP_{Strep}.

For protein expression, the assembled GFP_{172}-DI-RFP_{Strep} cassette was mobilized into pET vector. In doing so, *Nde*I site within the GFP_{172} coding region was removed by silent mutation using overlap PCR with primers *Nde*I-G172-F/Nd172R and Nd172F/G172-*Kpn*I-R. The modified GFP_{172} incorporating a 5'-*Nde*I and a 3'-*Xho*I site was subsequently cloned into pET-Trx_{His}-DI-RFP_{Strep} vector to replace the upstream thioredoxin, yielding pET-GFP_{172}-DI-RFP_{Strep}. To replace the downstream RFP_{Strep} protein in DI polyprotein to CAT_{Strep}, CAT_{Strep} coding sequence was amplified from pET-CAT vector by two step PCR using forward primer (CAT-2A-F1) and reverse primers (CAT-STREP-R and GKZ-R) incorporating *Apa*I and *Sac*I at 5'- and 3'-end, respectively, and replaced the *Apa*I-RFP_{strep}-*Sac*I fragment in pET-GFP_{172}-DI-RFP_{Strep} to create pET-GFP_{172}-DI-CAT_{Strep}. To eliminate the C-terminal overhang on the upstream POI, the N-extein linker extension sequence (LEGGSKFAND) preceding the DI domain in the GFP_{172}-

DI-RFP_{Strep} polyprotein was removed by mutating the pET-GFP_{172}-DI-RFP_{Strep} vector via inverse PCR using primer pair G172-DnaE-F and G172-DnaE-R. The amplified plasmid pET-GFP_{172}-DI(-)-RFP_{Strep} was transformed into *E. coli* DH5α after *Dpn*I enzyme treatment. To compare protein expression using our DI polyprotein system with the commercial vector incorporating multiple monocistronic expression cassettes, GFP_{172} and RFP_{Strep} were inserted into the first and the second multiple cloning sites of the pSKDuet01 vector (Addgene plasmid 12172), respectively. To do so, RFP_{Strep} fragment was PCR amplified from pETMD vector using forward primer (XBF) and reverse primer (MBR) containing a 5'-*Nde*I and a 3'-*Bam*HI, respectively. Amplified *Nde*I-RFP_{Strep}-*Bam*HI fragment was ligated into pSKDuet01 vector digested with the corresponding restriction enzymes, to yield pSKM. GFP172 incorporating a *Sac*II site and a *Sac*I site at the 5'- and 3'-ends, respectively, was amplified from pGEM5z-GFP172 vector using primers GFS2 and GRS. The resulting *Sac*II-GFP_{172}-*Sac*I fragment was ligated into pSKM vector digested with *Sac*II and *Sac*I enzymes to yield pSKGM. To compare the co-cleavage efficiency at both N- and C-termini of our DI fusion domain with the single intein incorporating the C+1A mutation, *Xho*I-*Ssp* DnaE(A_{159}CFN)-*Xba*I fragment in pUC-DnaE(A_{159}CFN) vector was replaced with *Xho*I-*Ssp* DnaE(N_{159}AFN)-*Xba*I which was amplified using primers *Xho*I-Int-F and DnaENAFN-*Xba*I-R, to yield pUC-DnaE(N_{159}AFN). The resulting *Xho*I-*Ssp* DnaE((N_{159}AFN)-*Xba*I fragment along with a C-terminal linker sequence was excised from pUC-DnaE(N_{159}AFN) vector between *Xho*I and *Apa*I sites and was subsequently used to replace the DI fragment in pUC-GFP_{172}-DI-RFP_{Strep} vector, to yield pUC-GFP_{172}-Int-RFP_{Strep}.

The cleavage-inactive DnaE intein mutant was generated using primers (IDnaE-*Kpn*I-F/ DnaE-*Xba*I-R-2) that carry the C1A mutation (in addition to the N159A mutation). This inactive

DnaE replaced its active counterpart in pET-GFP_{172}-DI-RFP_{Strep} to create pET-GFP_{172}-DI(N-)-RFP_{Strep} that encodes an N-(cleavage) inactive DI polyprotein. Here, the N-extein and C-extein residues flanking the inactive DnaE were changed to EY and GG, respectively, to minimize any cleaving activity. Similarly, cleavage-inactive DnaB intein mutant was generated using primers (DnaB-*Xba*I-F/IDnaB-*Apa*I-R) that contains N154A mutation (in addition to the C1A mutation). In addition, C-extein residues bordering DnaB were changed from SR to GS. The inactive DnaB fragment was digested with *Xba*I/*Apa*I, and ligated into the pET-GFP_{172}-DI-RFP_{Strep} vector digested with the corresponding restriction enzymes, yielding a vector pET-GFP_{172}-DI(C-)-RFP_{Strep} that encodes a C-(cleavage) inactive DI polyprotein.

For tobacco NT1 cell transformation, coding sequence of GFP_{172}-DI-RFP_{Strep} was inserted between *Sal*I and *Sac*I sites in the binary vector pE1775 under the control of the mannopine/octopine synthase, $(ocs)_3$/mas, promoter.[132] For mammalian HEK293T transient expression, GFP_{172}-DI-RFP_{Strep} fragment was PCR amplified using primers *Sac*I-*Nhe*-DI-F3 & Strep-*Not*I-R to incorporate *Sac*I and *Not*I at 5'- and 3'- ends, respectively, and inserted into pcDNA3.1 vector at the corresponding digested sites.

Expression in *Escherichia coli*.

The pET vectors encoding DI-polyprotein sequences were transformed into *E. coli* BL21(DE3). Bacterial cells were cultured in LB medium (1% tryptone, 0.5% yeast extract, 1% NaCl) at 37°C except when indicated otherwise. Protein expression was induced with 0.1 mM of isopropyl-β-D-thiogalactopyranoside (IPTG) when cells were grown to an OD_{600} of 0.5. In addition to *E. coli* BL21(DE3), protein expression was performed in SHuffle *E. coli* strain (New England Biolab, Ipswich, MA) lacking the trxB and gor reductases along with an additional

suppressor mutation (ahpC) which restores viability.[133] These mutations lead to an altered redox state in the SHuffle cells to permit oxidative folding. Upon completion of induction, cells were pelleted by centrifuging at 6000×g for 5 min, and then rinsed using phosphate buffer saline for three times. Washed cell pellets were disrupted by ultrasonication at 5 Watt for 5 min in PBS protein extraction buffer (20 mM sodium phosphate pH 7.5, 150 mM sodium chloride, 1 mM phenylmethylsulfonyl fluoride, 200 μM lysozyme, 1 μg/ml leupeptin, 100 ng/ml pepstatin). Total soluble proteins were obtained by centrifuging for 15 min at 14,000 × g at 4 °C and used in subsequent analysis.

Mammalian HEK293T cell transfection, protein expression, extraction, and analysis.

DI mediated polyprotein processing in mammalian system was tested in human embryonic kidney (HEK293T) cells. HEK293T cells were cultured in DMEM medium (Invitrogen, Grand Island, NY) supplemented with 10% fetal bovine serum (FBS) and 50 mg/L penicillin-streptomycin, and incubated at 37°C under 5% CO_2. Transient transfection was performed using Xtreme Gene 9 (Roche, Indianapolis, IN) when the cell density reached 70-90% confluency, by following the product instruction. The transfection complex was subsequently transferred to 3 ml of HEK293T culture. After 24 hour of incubation at 37°C under 5% CO_2, transfected cells were washed and resuspended in PBS protein extraction buffer, and then lysed by ultrasonication at 5 watt on the ice for 5 min to extract the total soluble protein. Protein extract was clarified by centrifugation at 14,000 × g at 4 °C for 15 min and subjected to subsequent analysis.

Tobacco transformation and protein characterization.

The GFP_{172}-DI-RFP_{Strep} polyprotein binary vector was transformed into *A. tumefaciens* C58C1 by electroporation. After PCR confirmation, transformed *Agrobacteria* was co-cultivated

with the tobacco NT1 cells or tobacco leaves (*N. tabacum* cv Xanthi) using a modified protocol reported previously.[134-135] The resultant transformants were selected on MS agar media[92] supplemented with hygromycin. The high expressing lines, screened based on GFP fluorescence and western blot analysis, were selected for subsequent characterization. Transient expression was performed in leaf tissues of *N. benthamiana* and Romaine lettuce via vacuum assisted agroinfiltration as described in section 2.2.[136] Protein extraction from transgenic tobacco was carried out by following a protocol described in section 2.2.[90] Clarified supernatants were used for subsequent western blot, protein, fluorescence analysis, and chromatographic purification. Fluorescence measurement was performed using a Hitachi F-2500 fluorescence spectrophotometer with excitation at 470 and 575 nm for detection of GFP and mCherry fluorescence, respectively. Amount of GFP and mCherry was estimated based on the fluorescence calibration curve established using purified protein standards.

Gel electrophoresis and western blot analysis.

Protein extracts were mixed with SDS-PAGE loading buffer and denatured at 95ºC for 5 minutes before SDS-PAGE separation. SDS-PAGE and western blot of the denatured samples were carried out as described in section 2.2. Anti-GFP rabbit serum (Invitrogen, Grand Island, NY), mouse anti-His tag (Lamda Biotech, St. Louis, MO) and rabbit anti-Strep tag II (Genscript, Piscataway, NJ) were used to detect processed proteins. Goat anti-rabbit or goat anti-mouse conjugated with alkaline phosphatase was used as secondary antibody (Southern Biotech, Birmingham, AL).

Purification of cleaved GFP_{172} for ESI-TOFMS analysis.

Crude protein extracts from transgenic tobacco NT1 calli were pre-clarified by 30% ammonium sulfate precipitation before applying onto a Phenyl-Sepharose hydrophobic-interaction column (GE Healthcare, Piscataway, NJ) connected to a Biologic Duo Flow chromatography system (Bio-Rad, Hercules, CA). The loaded column was rinsed with a washing buffer (20 mM sodium phosphate pH 8.0, 1 M ammonium sulfate), and eluted with a linear gradient from 1 M to 0 M ammonium sulfate over 20 column volumes. Fractions with GFP fluorescence were subsequently purified with a HiTrap IMAC-Sepharose column (GE Healthcare, Piscataway, NJ) by washing with a binding buffer (20 mM sodium phosphate pH 7.4, 500 mM sodium chloride and 20 mM imidazole) and eluted with a linear gradient from 20 mM to 500 mM imidazole over 12 column volumes. Fractions containing GFP fluorescence were concentrated and desalted with a Zeba desalting spin column (Thermo Fisher., Rockford, IL) for electrospray time-of-flight mass spectrometry (ESI-TOFMS) analysis on an Agilent 6210 LC-TOFMS system fitted with an ESI source operated in positive ion mode as described in section 2.2.[136]

Purification of cleaved RFP_{Strep}.

Cleaved RFP_{Strep} was purified with a Strep-Tactin Superflow Plus 2 ml column according to the manufacture manual (Qiagen, Valencia, CA). Briefly, the column was charged with crude soluble protein extracts from transgenic tobacco NT1 cells then washed with NP buffer (50 mM sodium phosphate pH 8.0, 300 mM sodium chloride), followed by elution with NPD buffer (NP buffer containing 2.5 mM desthiobiotin). The fractions shown mCherry fluorescence were concentrated for N-terminal amino acid sequencing.

N-terminal amino acid sequencing.

The GFP_{172} and RFP_{Strep} liberated from the polyprotein precursor were purified by chromatography as described above, followed by separation on SDS-PAGE and transferred onto a PVDF membrane. The target band with the expected size was excised out after staining with Coomassie blue. The N-terminal sequencing was performed using Edman degradation on a Perkin Elmer Applied Biosystems Procise 494 protein/peptide sequencer coupled with an on-line Perkin Elmer Applied Biosystems Model 140C PTH Amino Acid Analyzer (Applied Biosystems, Carlsbad, CA), performed by the Protein Core Facility at the Iowa State University.

CAT assay.

Extracts of *E. coli* expressing GFP_{172}-DI-CAT_{Strep} were diluted in a reaction mixture containing 100 mM Tris buffer (pH 7.8), 8 mM of 5,5'-Dithio-bis (2-Nitrobenzoic Acid) (DTNB), and 16 mM of Acetyl Coenzyme A. After equilibration, the reaction was initiated by adding 0.005% of chloramphenicol. The enzyme activity was calculated based on the increase of A_{412} over 5 min at room temperature.

3.3. Results and Discussion

Designing the DI based polyprotein system for efficient autoprocessing.

Efficient autoprocessing (cleaving) of a polyprotein precursor *in vivo* into physically separated constituent proteins is essential in developing a polyprotein-based vector system for coordinating co-expression of multiple proteins from a single ORF. This is achieved here by incorporating engineered intein elements. Generally, intein splicing begins with the N-S/O acyl shift reaction at the amide bond preceding the N-terminal junction with the side chain of the first

intein residue (position 1, typically Cys or Ser) (Fig. 3.2). Thereafter, the resulting linear (thio)ester intermediate undergoes trans(thio)esterification by nucleophilic attack from the sulfhydryl or hydroxyl groups on the side chain of the first C-extein residue (position +1, typically Cys, Ser, or Thr) and subsequently forms a branched intermediate. Finally, the branched intermediate is resolved by Asn cyclization of the intein C-terminal to liberate the intein fragment and ligate N- and C-exteins via a (thio)ester bond, which can be spontaneously rearranged into a more stable amide bond.[137]

The rationale behind development of the DI system is that N- and C-terminal autocleavage reactions conferred by engineered inteins can be isolated and tuned separately.[137] Here, each intein domain in the DI cassette is optimized independently, to maximize the N-terminal and C-terminal autocleavage activity of *Ssp* DnaE and DnaB intein, respectively. As shown in Figure 3.1b, for the *Ssp* DnaE mini-intein moiety of the DI system (**1**), we created the Ala mutation at position 159 (N159A) to abolish Asn cyclization, and incorporated an Asp residue at position -1 (Asp-1) that has been reported to accelerate *Ssp* DnaE autocleavage at its N-terminal splicing junction.[128] With the N159A mutation, release of the N-extein is believed to proceed via hydrolysis of the labile thioester bond in the linear (**2**) and branched (**3**) intermediates following the general N-terminal intein autocleavage (Fig. 3.1b, Pathways A & B).[137] For the *Ssp* DnaB mini-intein moiety in **1**, we incorporated the Ala mutation at position 1 (C1A) that blocks the formation of thioester linkage preceding the N-terminus of intein. In addition, the Ser-Arg residues that were experimentally confirmed to give high C-terminal autocleavage activity, were incorporated after the C-terminal junction at position +1 and +2 of DnaB intein moiety, respectively.[131] Based on the engineered DI system, we assembled a panel of reporter polyprotein

constructs to systematically investigate key factors governing the autoprocessing activity of the DI domain. Each of these constructs encodes a polyprotein (Fig. 3.1a) in which the DI sequence is sandwiched between an upstream GFP variant (GFP_{172}; POI1),[129] and a downstream mCherry sequence carrying a C-terminal Strep tag (RFP_{Strep}; POI2). T7 promoter, human cytomegalovirus (CMV) promoter, and $(ocs)_3$/mas promoter were used to drive expression of the DI-based polyproteins in *E. coli*, mammalian HEK293T cells and *N. tabacum*, respectively.

Cellular processing of the dual-intein polyprotein in bacterial, mammalian and plant expression systems.

To investigate the extent of DI-mediated polyprotein autoprocessing, protein extracts from *E. coli*, HEK293T, and tobacco NT1 cells expressing the GFP_{172}-DI-RFP_{Strep} reporter construct were subject to analysis of anti-GFP and anti-Strep tag western blots. As shown in Figure 3.3, both GFP_{172} and RFP_{Strep} were nearly completely released from the polyprotein precursor in all tested host systems. The lower immuno-reactive band appeared in the anti-Strep tag western blot represents a partially hydrolyzed mCherry product that results from sample preparation for gel electrophoresis.[138] Processing of the DI polyprotein system at the whole plant level was investigated in transgenic *N. tabacum* cv. Xanthi plants expressing the GFP_{172}-DI-RFP_{Strep} reporter construct. As with undifferentiated tobacco NT1 cells, efficient cleavage on both N- and C-termini of the DI fusion domain was observed in leaf, stem and root tissues based on the western blot analysis (Fig. 3.4a). While no significant difference in the extent of cleavage was noted between different plant tissues, highest expression level was observed in root tissue. This is due mainly to the $(ocs)_3$/mas promoter used which gives highest expression in roots as reported in both tobacco and maize plants.[87,132] In addition to *N. tabacum*, efficient processing of

the DI polyprotein was also seen in leaf tissues of *N. benthamiana* and *Lactuca sativa L. var. longifolia* (Romaine lettuce) by transient expression via agroinfiltration (Fig. 3.4b & c). Demonstration of the efficacy of the DI system in *N. benthamiana* is of particular interests as agroinfiltration of this plant species offers a very promising large scale recombinant antibody production platform which was used to produce ZMapp, a promising drug to fight the Ebola outbreak.[139-140]

Determination of the cleavage sites in the DI polyprotein precursor.

To further investigate whether the observed cellular cleavage of the precursor polyprotein was specifically attributed to the autocatalytic cleavage activity of the DI fusion domain, inactive intein mutants with blocked N- or C-terminal auto-cleavage activity were used to replace the cleavage-active inteins in the DI domain. Specifically, the first (Cys, position 1 of DnaE intein) or the last (Asn, position 154 of DnaB intein) residue in the DI domain was mutated to Ala to create an N-inactive or a C-inactive DI, respectively. As shown in Fig. 3.5a, when the N-terminal cleavage-inactive DI was used, the upstream GFP_{172} was unable to separate from the DI domain, yet since the C-terminal cleavage was still active, the downstream RFP_{Strep} could still be released from the polyprotein precursor. Similarly, for the C-terminal cleavage-inactive DI, the upstream GFP_{172} was liberated from the DI domain but the downstream RFP_{Strep} was not (Fig. 3.5b). This result confirmed that the highly efficient polyprotein auto-cleavage observed with the active DI domain was indeed resulted from the specific action of the DI domain as described by the intein cleavage mechanism (Fig. 3.1b).

The exact cleavage sites within the GFP_{172}-DI-RFP_{Strep} polyprotein were determined by analyzing the released RFP_{Strep} and GFP_{172}, purified from the transgenic tobacco NT1 cells.

According to N-terminal sequencing, the RFP product released from the polyprotein contains an N-terminal Ser at position +1 of DnaB intein (followed by the Arg-Gly-Pro sequence), while the N-terminus of the released GFP_{172} was found to be blocked. As shown in Figure 3.5c, the Ser-Arg-Gly-Pro sequence is flanking the C-terminal splicing junction of the DnaB intein, and hence it confirmed the release of the downstream RFP is mediated by the C-terminal auto-cleavage action of the intein. Furthermore, the observed molecular mass of the released GFP_{172} product measured using ESI-TOFMS is 28,590.90 Da which matches that of an N-terminal acetylated GFP_{172} (calculated mass 28,593.08 Da). As the most common substrate for N-terminal acetyltransferase, Ser at N-terminus of the upstream GFP_{172} was likely to be acetylated after the initiator Met was removed by methionine aminopeptidase. Taken together, these findings supported that the observed *in vivo* processing of the DI polyprotein precursor indeed occurred at the anticipated sites bordering the DI domain.

Effects of cellular redox environment and culture temperature on the DI autocleavage activity.

Since kinetics of intein autocleavage could be affected by expression temperature and differential cellular redox environment, we investigated how these factors could affect the DI mediated polyprotein processing in *E. coli*. The GFP_{172}-DI-RFP_{Strep} reporter construct was mobilized into *E. coli* BL21 and SHuffle[133] strains engineered to have different cellular redox environments, and induced the protein expression at two different temperatures (25 vs. 37°C). Our results showed little difference in the polyprotein processing between the two strains tested (Fig. 3.6a & b). As shown in Figure 3.3a & b, efficient polyprotein processing was also noted in HEK293T and tobacco NT1 cells which had a more oxidative cytosolic environment compared

with *E. coli*. Therefore, with the highly efficient *Ssp* DnaE and DnaB mini-intein domains, self-excising activity of our DI system was not susceptible to differences in the host cell redox environment, at least within the differences found in the hosts tested.

As for the temperature effect, overnight induction of *E. coli* at 25°C did not affect the N-terminal cleavage efficiency of the DI domain (Fig. 3.6a). However, detection of the DI-RFP_{Strep} fragment on the western blot indicated less efficient C-terminal cleavage at 25°C, although the properly cleaved downstream RFP_{Strep} remained the predominant product based on the western blot analysis (Fig. 3.6b). Previous intein studies suggested that the N-terminal Cys or Ser residue at position 1 of intein played a crucial role in promoting Asn cyclization.[141] After trans(thio)esterification, the side chain of the liberated N-terminal Cys or Ser of intein nucleophilically attacked the scissile peptide bond at the C-terminal Asn, forming a transient macrocyclic intermediate that promotes Asn cyclization. When the N-terminal nucleophilic residue (position 1) was substituted with non-functioned Ala, reduction in C-terminal intein cleavage rate was observed.[142-143] In our DI design (Fig. 3.1b), C1A mutation was introduced into the C-terminal cleaving intein (*Ssp* DnaB in **1**) to enforce the intein reaction towards C-terminal cleavage, thus the effect of the first nucleophilic side chain was eliminated. To this end, the Asn cyclization was mainly initiated by protonation of the imidazole side chain of the penultimate His (position 153 of DnaB intein).[144-145] The attenuated C-terminal cleavage rate of the C1A substitution could be compensated by increased hydrolysis rate at elevated reaction temperature. To this end, after overnight induction at 25°C, the *E. coli* cells were resuspended in PBS buffer and incubated at 37°C for up to 12 hours. Nearly complete release of RFP_{Strep} from the DI-RFP_{Strep} fragment was observed after about only four hours (Fig. 3.6c & d). This post-induction

incubation strategy provides a solution to remedy the effect of the attenuated C-terminal cleavage at lower temperature. Since only the *Ssp* DnaB intein was tested in this study as the C-terminal cleavage domain, this intein might be replaced by another intein capable of more active C-terminal cleavage. Such intein variants have been reported, for instance by Ramirez *et al*.[146]

Characterization of proteins released from the DI polyprotein precursor.

To determine whether proteins of interest released from the polyprotein precursor remain functional, transgenic cells expressing the GFP_{172}-DI-RFP_{Strep} reporter were examined using fluorescence microscopy. As shown in Figure 3.7a, bright green and red fluorescence distinctive from the background autofluorescence were visible in all tested cell lines. These results, in consort with the western blot data (Fig. 3.3) that showed essentially complete cleavage of the polyprotein precursor, indicated that fluorescence detected in the transgenic cells was originated from the processed proteins and hence these proteins are indeed functional. As a side note, slightly different distribution of GFP and RFP can be seen in the plant cell images in Figure 3.7a. Unlike the cytoplasmic distribution of RFP fluorescence, GFP was prone to accumulation in the nuclear region of tobacco cells although no nuclear localization signal had been identified in its sequence.[147-148] Nuclear localization of GFP was thought to be resulted from the passive diffusion through the nuclear pores due to their low molecular mass.[148-149] Once inside the nucleus, GFP oligomerized into large homomultimers for the high local concentration, trapping the GFP in the nuclear boundary and leading the equilibrium towards nucleus import.[148] On the other hand, RFP (mCherry) tends to maintain its monomeric structure and traverse freely through the nuclear pores. Further analysis of the transgenic cell extracts using fluorescence spectrometry showed emission spectra distinctive to GFP and RFP (mCherry), indicating that not only the

released GFP_{172} and RFP_{Strep} displayed fluorescence, the spectral properties of the released fluorescent proteins were not affected by the DI-mediated *in vivo* cleavage (Fig. 3.7b).

In multi-protein co-expression, it is important to be able to control the relative stoichiometry of the gene products. To determine whether stoichiometric accumulation of the constituent proteins can be achieved using the DI polyprotein approach, cellular accumulation of cleaved GFP_{172} and RFP_{Strep} was analyzed using scanning densitometry of western blots, along with protein quantification based on fluorescence measurement. As shown in Figure 3.8, approximately equal molar accumulations of processed GFP_{172} and RFP_{Strep} were noted in mammalian HEK293T cells. However, production of GFP_{172} was about 50% lower than RFP_{Strep} in *E. coli* BL21 induced at 37°C. GFP_{172} variant contained an internal hexa-histidine sequence inserted between GFP residues 172 and 173. This GFP variant was found in our earlier study to fold less efficiently in *E. coli* at 37°C than at 25°C.[129] GFP with a C-terminal His tag (GFP_{H6}) improved the folding at 37°C, resulting in nearly equimolar accumulation of GFP_{H6} and RFP_{Strep} (Fig. 3.8). This result illustrated the fact that post-translational modification and/or differential protein stability could also play a role in deciding the relative accumulation levels of the constituent proteins released from the polyprotein. For proteins with similar stability or folding efficiency, we showed here that stoichiometric production could indeed be achievable in *E. coli* with the DI polyprotein vector system when the GFP_{H6} was used to replace the upstream GFP_{172}. Such stoichiometric expression is an important feature that makes the DI polyprotein system superior to expression vectors containing multiple monocistronic expression cassettes. To this end, we observed that GFP_{172} and RFP_{Strep} protein levels differed by over 10-fold when co-

expressed using the pSKDuet01 (Addgene plasmid 12172) vector which is derived from pRSFDuet-1[121] that contains two expression cassettes (Fig. 3.8).

In mammalian HEK293T cells, GFP_{172} apparently folded relatively efficiently at 37°C, reaching almost similar expression levels as RFP_{Strep} (Fig. 3.8). In plant hosts, for both undifferentiated cultured tobacco cells and leaf tissue of transgenic tobacco, the downstream RFP_{Strep} accumulated to a lesser amount than GFP_{172}. As discussed above, decline in C-terminal intein cleavage was observed in *E. coli* cells cultured at lower temperature (e.g., at 25°C). This attenuated cleavage efficiency was likely caused by the intrinsic C1A mutation in *Ssp* DnaB mini-intein domain that reduced reaction rate of C-terminal Asn cyclization in a temperature-dependent manner. Thus, similar reduction in C-terminal cleavage could also be expected in plants that preferably grown at room temperature. Failure in detection of the uncleaved DI-RFP_{Strep} fragment in plant extracts by western blot suggested its degradation inside the cells.

To further demonstrate the general applicability of the DI polyprotein vector system for multi-protein co-expression, upstream GFP_{172} and downstream RFP_{Strep} were replaced by a C-terminal His tagged thioredoxin (Trx_{His}) and a C-terminal Strep tagged chloramphenicol acetyltransferase (CAT_{Strep}), respectively, to create constructs Trx_{His}-DI-RFP_{Strep} and GFP_{172}-DI-CAT_{Strep}. As revealed by the anti-His tag and anti-Strep tag western blots, highly efficient processing of the polyprotein precursors with essentially complete release of the constituent proteins in *E. coli* BL21 was achieved in both cases (Fig. 3.9). The bioactivity of the released CAT enzyme was also confirmed by enzymatic assay. Based on scanning densitometry analysis of western blots, Trx_{His} and RFP_{Strep} were found to accumulate to approximately equimolar amounts in *E. coli* cells expressing Trx_{His}-DI-RFP_{Strep}. As for the GFP_{172}-DI-CAT_{Strep} vector,

GFP_{172} accumulated to a lesser amount than CAT_{Strep}, which again was most likely attributed to the less efficient folding of GFP_{172} in *E. coli* BL21 at 37°C, as discussed above. Nonetheless, when *E. coli* transformed with the GFP_{172}-DI-CAT_{Strep} vector was induced overnight at 25°C, the molar ratio between GFP_{172} and CAT_{Strep} became very close to one. These results point to potential universal applicability of the DI-polyprotein vector system for multi-protein co-expression, as the autocleavage activity of the DI fusion domain is apparently quite insensitive to the flanking protein sequences.

Protein expression levels using the DI polyprotein vectors.

Total soluble amounts of GFP_{172} generated using the DI-polyprotein vector (encoding GFP_{172}-DI-RFP_{Strep}) were compared with those using vectors harboring either a single non-fused GFP_{172} or RFP_{Strep}, or separate GFP_{172} and RFP_{Strep} expression cassettes, under the control of the same promoter. This comparison was carried out in *E. coli* and tobacco NT1 cells, using the T7 and the $(ocs)_3$/mas promoter, respectively. We investigated *E. coli* BL21(DE3) and the SHuffle[133] strain for GFP_{172} expression from the DI polyprotein vector vs. the pSKDuet vector (co-expressing GFP_{172} and RFP_{Strep} from separate monocistronic cassettes on the same vector). The SHuffle strain constitutively expresses DsbC which is a disulfide bond isomerase known to serve as a chaperone that assists in protein folding.[150] Interestingly, while GFP_{172} expression is lower when expressed using the DI vector than using pSKDuet in BL21(DE3) upon overnigh expression at 37°C, the trend is reversed at 25°C in both BL21 and SHuffle. At 37°C in the SHuffle strain, GFP_{172} expression is similar using either the DI vector or pSKDuet (Fig. 3.10). Chaperone co-expression has been reported to cause proteolytic degradation of folding-reluctant protein species,[151] which might explain the low expression levels seen in the SHuffle strain at

37°C, as GFP_{172} is known to fold less efficiently than the regular GFP at this temperature.[129] The lower expression temperature (25°C) facilitated folding of GFP_{172} and the DI domain in both strains.

In tobacco, GFP_{172} reached 0.28% and 0.04% of total soluble proteins (TSP) when expressed from the GFP_{172}-DI-RFP_{Strep} polyprotein vector in NT1 cells and *N. tabacum* leaf tissue, respectively. For comparison, expression of GFP_{172} or RFP_{Strep} alone using the same $(ocs)_3$/mas promoter in transgenic tobacco NT1 cells reached about 0.29% and 0.61% TSP, respectively. Although more proteins need to be tested, the encouraging results of reporter protein expression suggest that the DI-based polyprotein vector system could produce a comparable level of recombinant proteins as with the conventional single protein vectors.

Autoprocessing activity of the DI domain without its N-terminal extein linker extension.

In the data presented above, the DI domain employed contains an N-terminal extension derived from the native flanking N-extein sequence of the *Ssp* DnaE intein but with mutations at N-1 and N-2 positions to Asp and Asn, respectively (Fig. 3.5c). This linker extension was incorporated with the intention to accelerate autocleavage at the N-terminal of the DI domain.[128] While this version of the DI domain mediated highly efficient autocleavage as presented in the preceding sections, we have subsequently confirmed that comparable autocleavage can be attained by removing the DI N-terminal extension. As shown in the western blot result presented in Figure 3.11, co-expression of GFP_{172} and RFP_{Strep} using a polyprotein vector containing a modified DI domain omitting its N-terminal extension sequence, i.e., LEGGSKFAND (cf. Fig. 3.5c), led to very efficient release of both GFP_{172} and RFP_{Strep}. This result is rational since it is known that the key residues involved in N-terminal intein autocleavage reside inside and

immediately downstream of the intein sequence.[137] In this case, no extraneous amino acids are appended to the POI released from the N-terminal of the DI domain. As for the downstream POI, after it is released from the polyprotein precursor, and upon purification, the amino acid overhang at its N-terminus can be removed to preserve its authentic N-terminal residue using a suitable aminopeptidase. As an example, commercially available TAGZyme system based on the use of dipeptidyl aminopeptidase I (DAPase Enzyme) in combination with glutamine cyclotransferase (Qcyclase Enzyme) and pyroglutamyl aminopeptidase (pGAPase Enzyme) (Qiagen, Valencia, CA) may be considered. To this end, a DAPase stop point (e.g., a Gln residue) needs to be introduced in front of the native N-terminal residue. Upon removal of the first N-terminal dipeptide, DAPase removes dipeptides progressively till it encounters the Gln stop point which is then removed by Qcyclase and pGAPase.[152]

Autoprocessing of polyprotein containing a single intein domain.

In principle, self cleavage on both ends of an intein may be achievable by promoting hydrolysis of the linear (thio)ester intermediate and allowing more efficient intein C-terminal Asn cyclization.[142] It has been suggested that release of both N- and C-exteins from a single intein without subsequent extein splicing could occur by mutating the essential residue at position +1 (flanking the intein C-terminus) to Gly or Ala.[142] Here we investigated whether co-cleavage at both splicing junctions of a single intein with such mutation can result in efficient release of flanking POIs as in the case of DI-based vectors. A polyprotein construct that encodes an *Ssp* DnaE mini-intein with C+1A mutation sandwiched between a GFP_{172} and a RFP_{Strep} was created and expressed in *E. coli* BL21(DE3). Cell extract was probed with western blot. As shown in Figure 3.12, a substantial portion of uncleaved or partially cleaved polyprotein

fragments was detected on western blots of the *E. coli* extract, and only a small portion (about 10%) of N- and C-exteins resulted from co-cleavage at both ends of the intein was observed. By comparing this result with that shown in Figure 3.11 for a DI-based polyprotein system, it is clear that the DI domain is necessary to confer hyperactive autocleavage of the polyprotein precursor to achieve highly efficient release of the POIs.

3.4. Conclusion

In this study we demonstrated that incorporation of an engineered DI domain in a polyprotein system enabled highly efficient synchronized protein co-expression from a single ORF in different expression host systems. This approach has a number of unique advantages over existing methods. The DI system is based on sequences of non-viral origins and hence avoids biosafety, environmental, and compliance issues. The highly efficient and rapid *in vivo* protein processing is autocatalytic and it does not require host-specific factors. Since the multiple POIs are encoded by a single ORF, it potentially allows control over the relative stoichiometry of the gene products to achieve a balanced multi-protein synthesis, especially when the POIs have similar folding and stability characteristics. The DI technology is applicable to both prokaryotic as well as eukaryotic hosts. By integrating with homologous recombination-based cloning methods such as the popular Gateway technology, the new DI polyprotein cassette can be easily shuttled between different hosts for expression. On the downside, with the DI based polyprotein vector the multiple POIs are expressed from the same promoter and thus if expression of each POI needs to be driven by a different promoter, e.g., for expression in different tissues or to be expressed at different times, the DI polyprotein vector will not be well suited. Overall, the DI

based polyprotein approach has many unique features and is also highly complementary to existing protein expression techniques, and hence it is a highly valuable addition to the molecular toolbox for synchronized multi-protein co-expression.

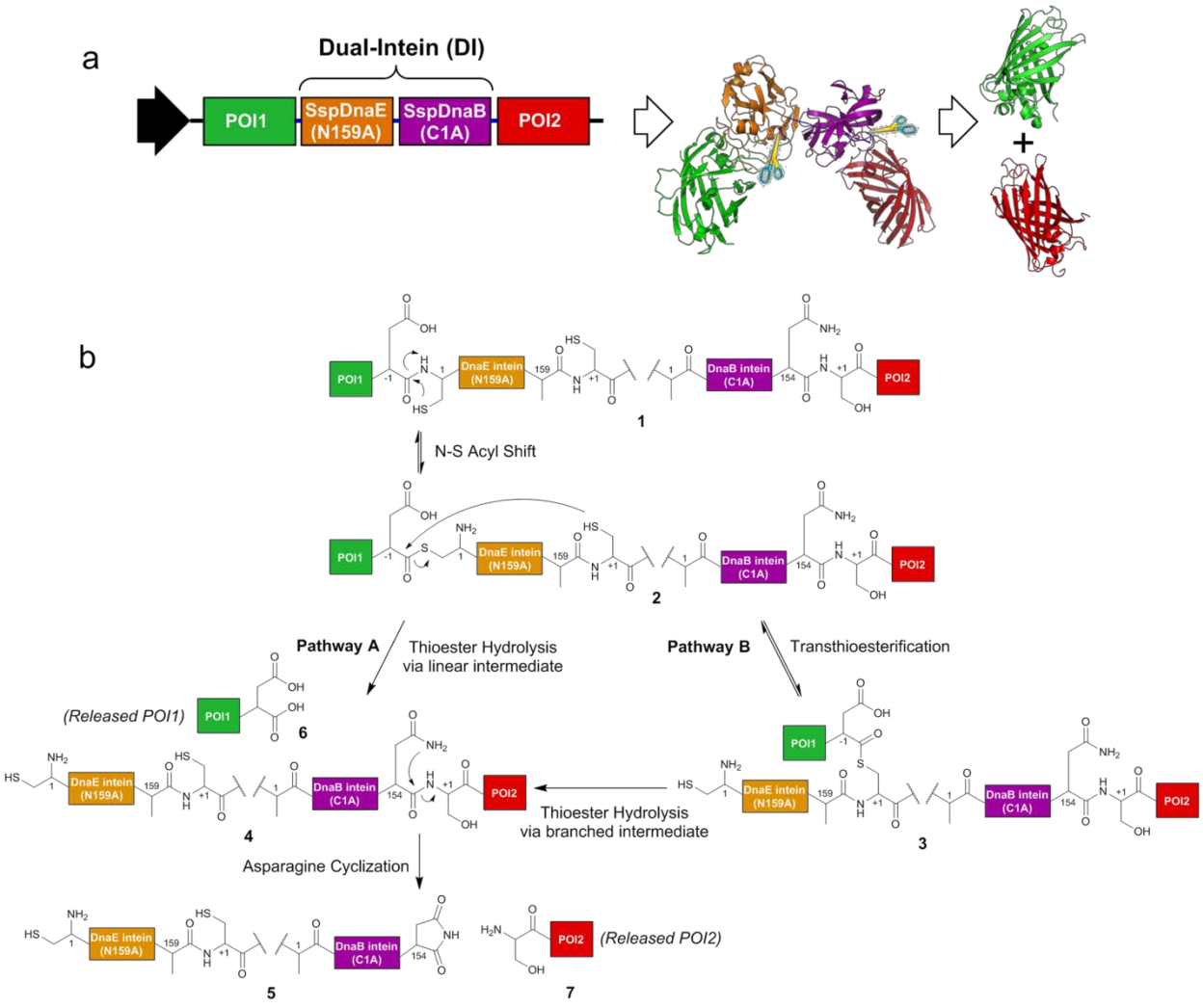


Figure 3.1. Synchronized co-expression of multiple proteins using the DI-based polyprotein vector system. (a) Organization of the DI-based polyprotein ORF, and autoprocessing of the polyprotein precursor into two separate proteins of interest (POI); illustrated for co-expression of GFP and RFP as POI1 and POI2, respectively. (b) Proposed mechanism of DI mediated autocleavage of a POI1-DI-POI2 polyprotein precursor. Arrows represent the general intein cleavage (pathways A & B).

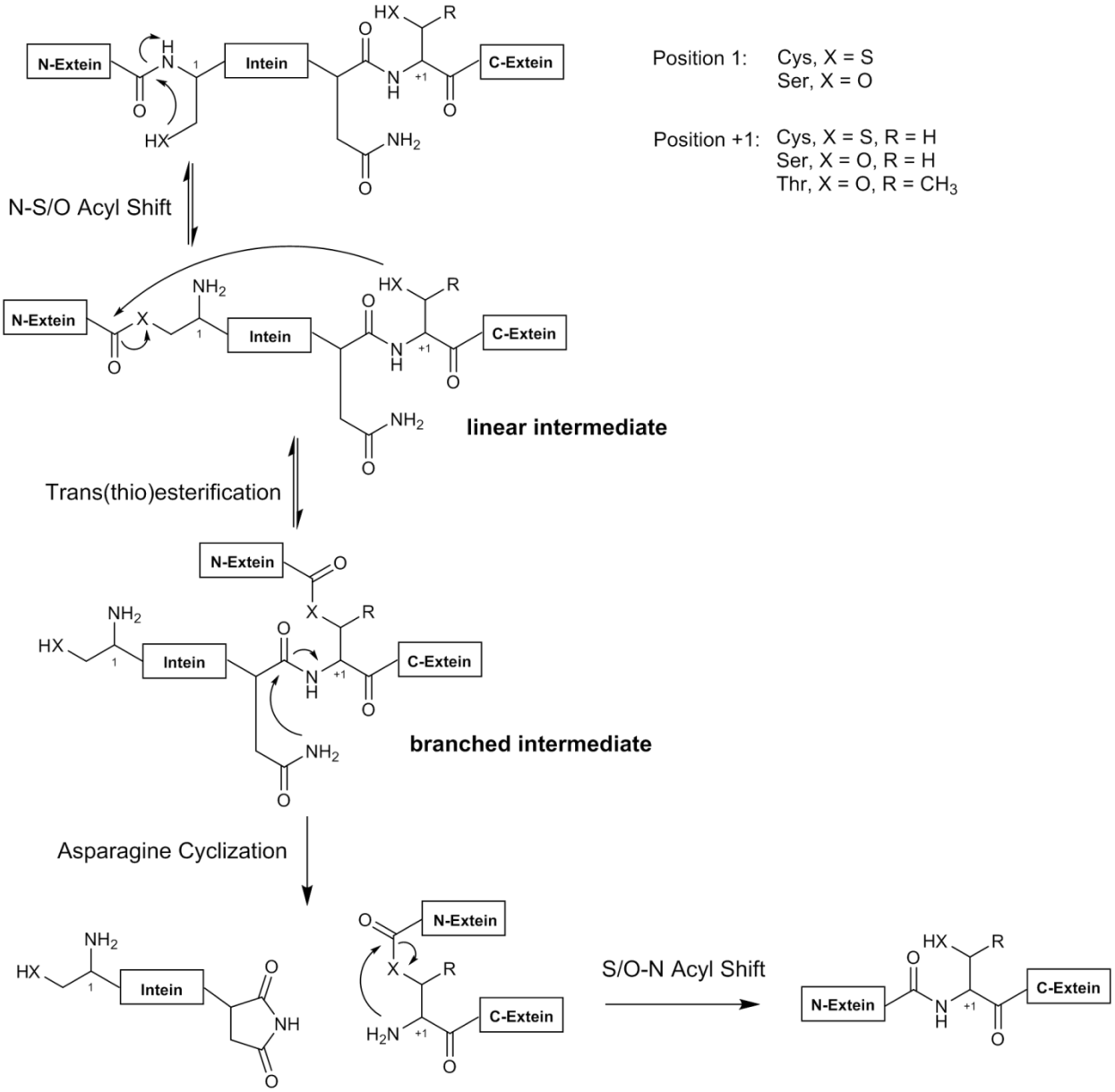


Figure 3.2. Mechanism of intein mediated protein splicing.

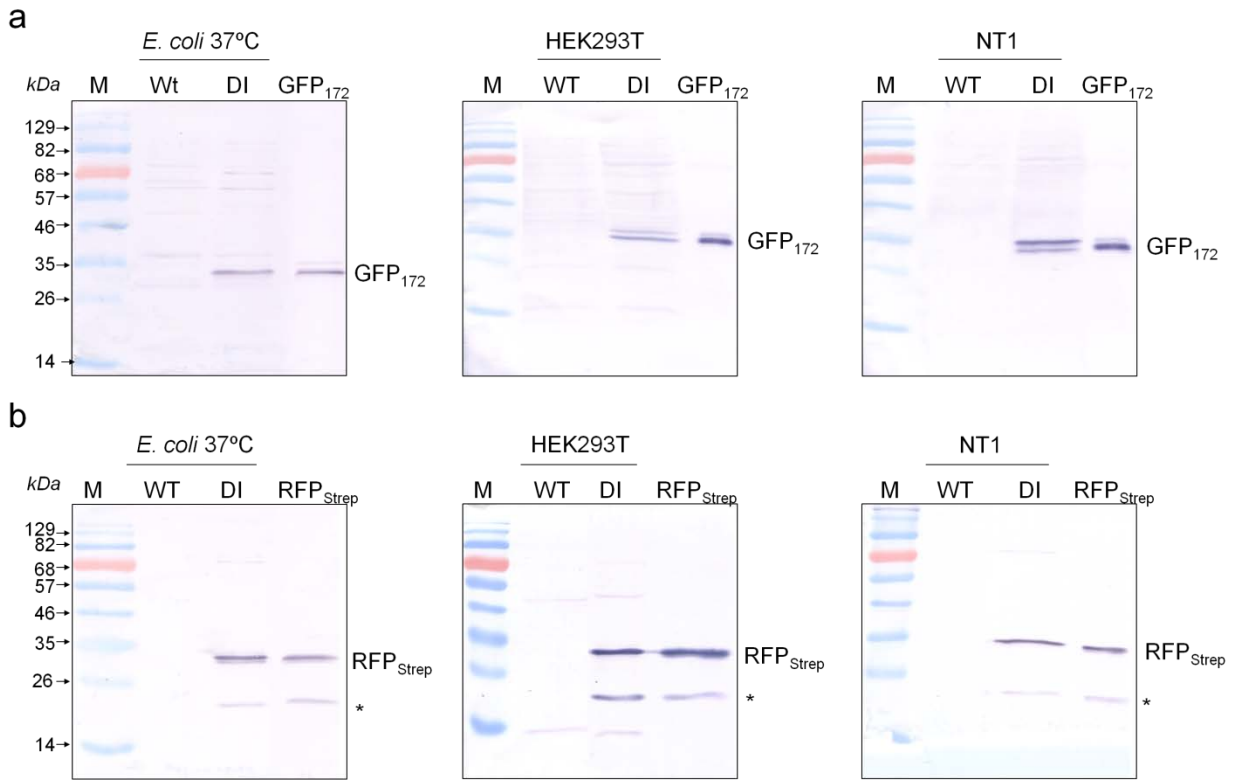


Figure 3.3. DI mediated autoprocessing of GFP_{172}-DI-RFP_{Strep} polyprotein in *E. coli*, mammalian HEK293T, and tobacco NT1 cells. Total soluble protein extracts were probed using western blots. (a) Release of upstream protein probed using an anti-GFP antibody. (b) Release of downstream protein was confirmed using an anti-strep tag antibody. Hereinafter, "M" represents protein markers, "WT" means non-transformed wild-type control, and "*" denotes a degraded RFP_{Strep} fragment resulting from sample preparation for gel electrophoresis (see text for details).

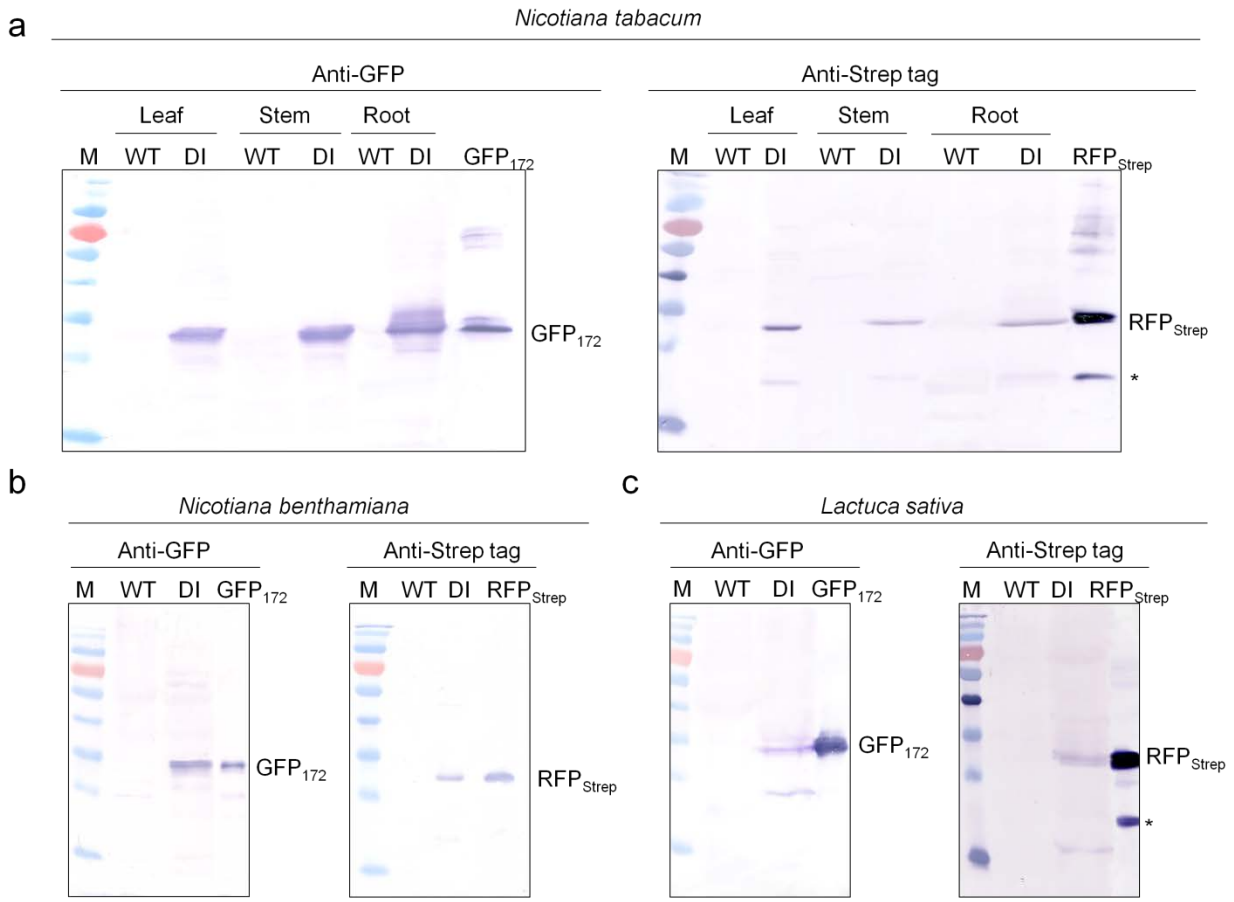


Figure 3.4. Processing of the GFP_{172}-DI-RFP_{Strep} polyprotein in plants, probed with western blots. (a) Leaf, stem, and root tissues of stably transformed *N. tabacum* plants. (b) Leaf tissues of agroinfiltrated *N. benthamiana*. (c) Leaf tissues of agroinfiltrated *L. sativa L. var. longifolia*.

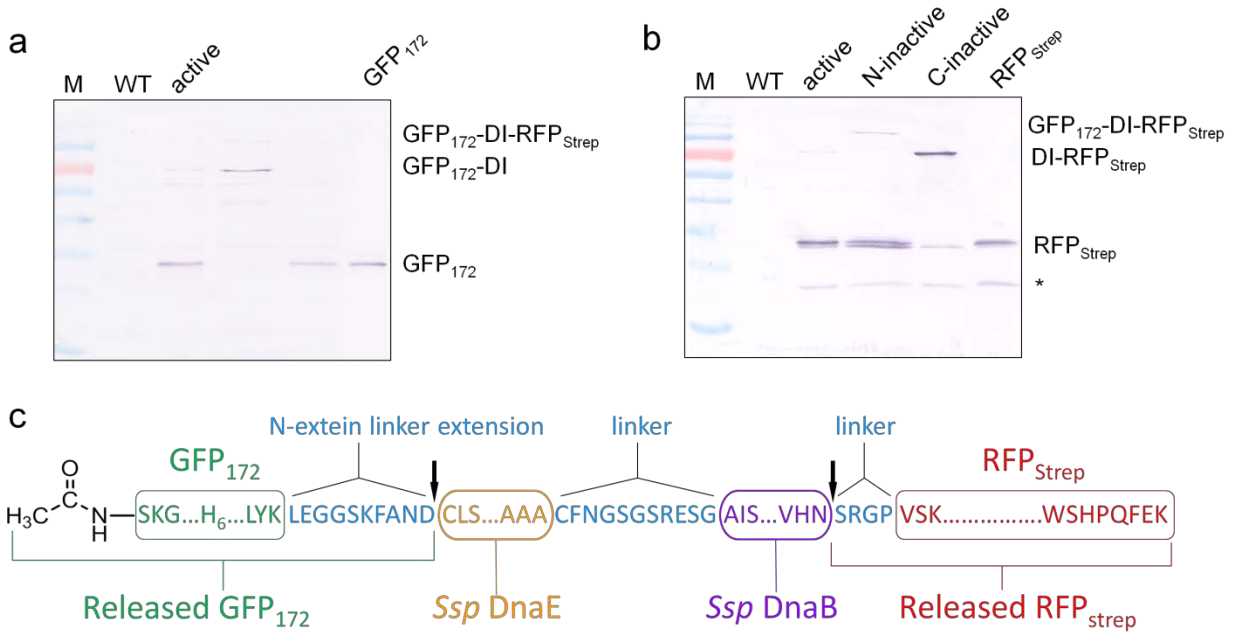


Figure 3.5. Characterizing autoprocessing of the GFP_{172}-DI-RFP_{Strep} polyprotein. (a, b) Western blot analysis of cleavage patterns of polyproteins carrying active vs. N- or C- inactive DI variants confirmed release of the constituent proteins (GFP_{172} and RFP_{strep}) was attributed to the autocatalytic cleavage activity of the DI domain. The N- or C-terminal cleavage activities of the DI domain was blocked by Ala mutation at the respective cleaving junction to create N- and C-inactive DI mutant. (c) Organization of the polyprotein precursor showing the main protein domains (boxed) and the linker sequences (in blue); N-terminal amino acid sequencing and ESI-TOFMS analysis of the released proteins revealed cleavage of the polyprotein at the expected sites indicated by the arrows.

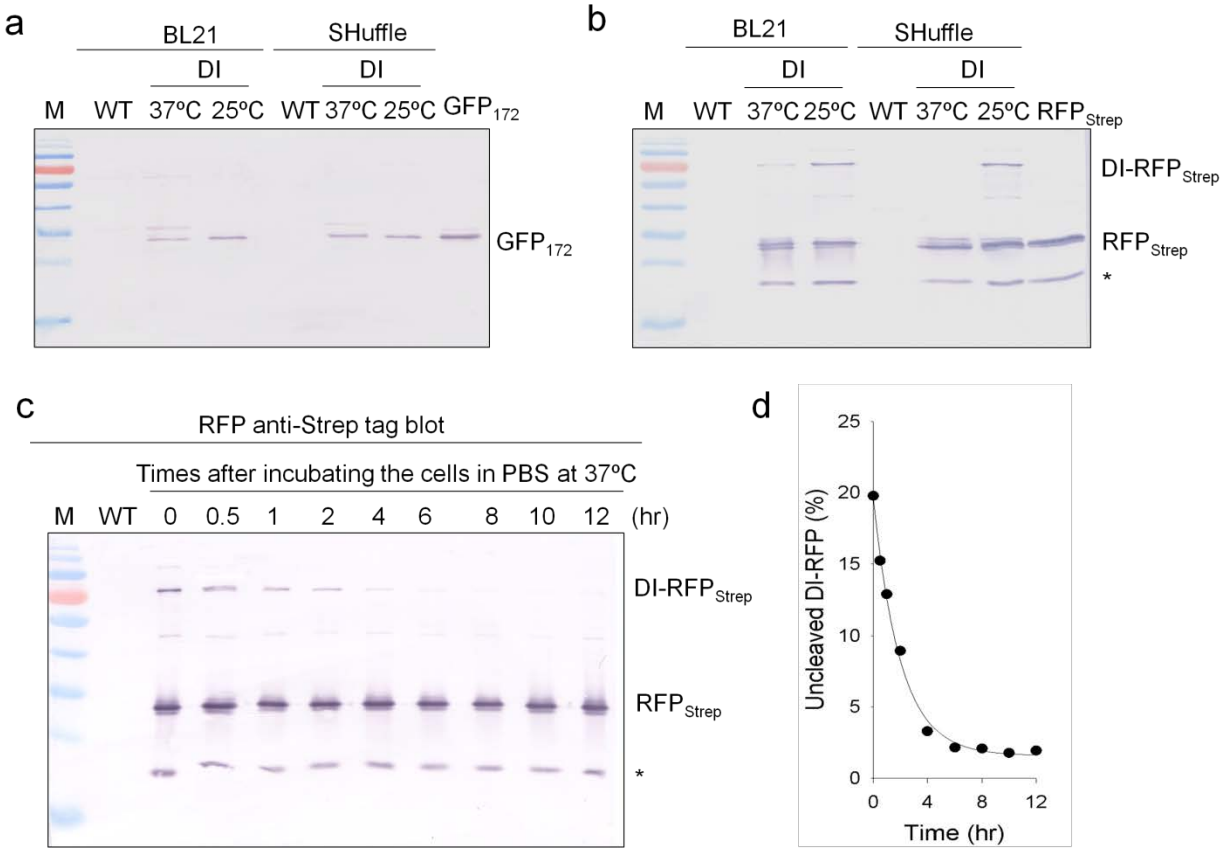


Figure 3.6. Effects of cellular redox environment and culture temperature on DI autocleavage activity. (a, b) Effect of cellular redox environment and expression temperature on the cleavage efficiency of the DI domain. (c) C-terminal cleavage efficiency of the DI domain in *E. coli* was improved by incubating at 37°C after overnight expression at 25°C. (d) Nearly complete cleavage was observed within 4 hours after incubation at 37°C.

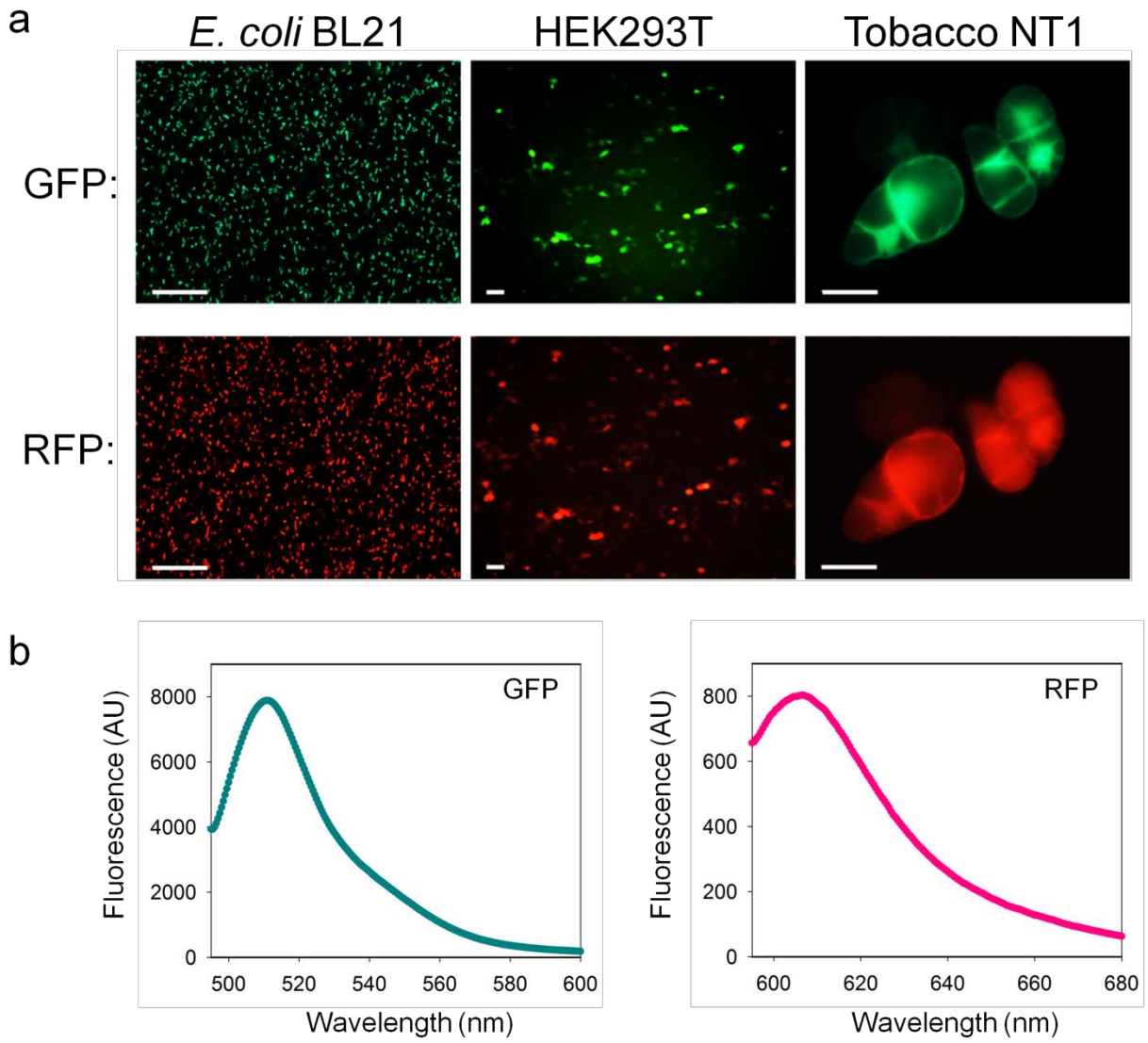


Figure 3.7. Processed POIs from the DI-based polyprotein precursor preserve proper function. (a) Cells expressing the GFP_{172}-DI-RFP_{Strep} reporter construct display both green (upper panel) and red (lower panel) fluorescence. Scale bar: 50 µm. (b) GFP_{172} and RFP_{Strep} released from the polyprotein precursor are functional as indicated by their respective fluorescence emission spectra.

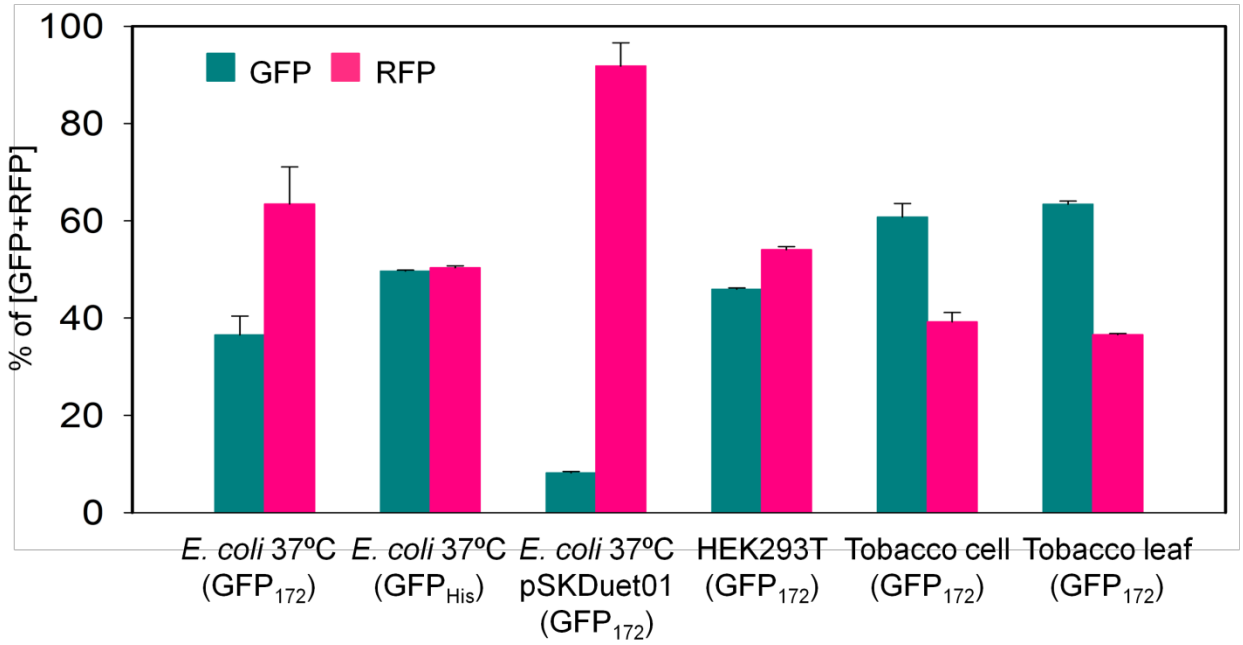


Figure 3.8. Relative expression levels of the constituent proteins released from the GFP-DI-RFP polyprotein precursor in different expression hosts.

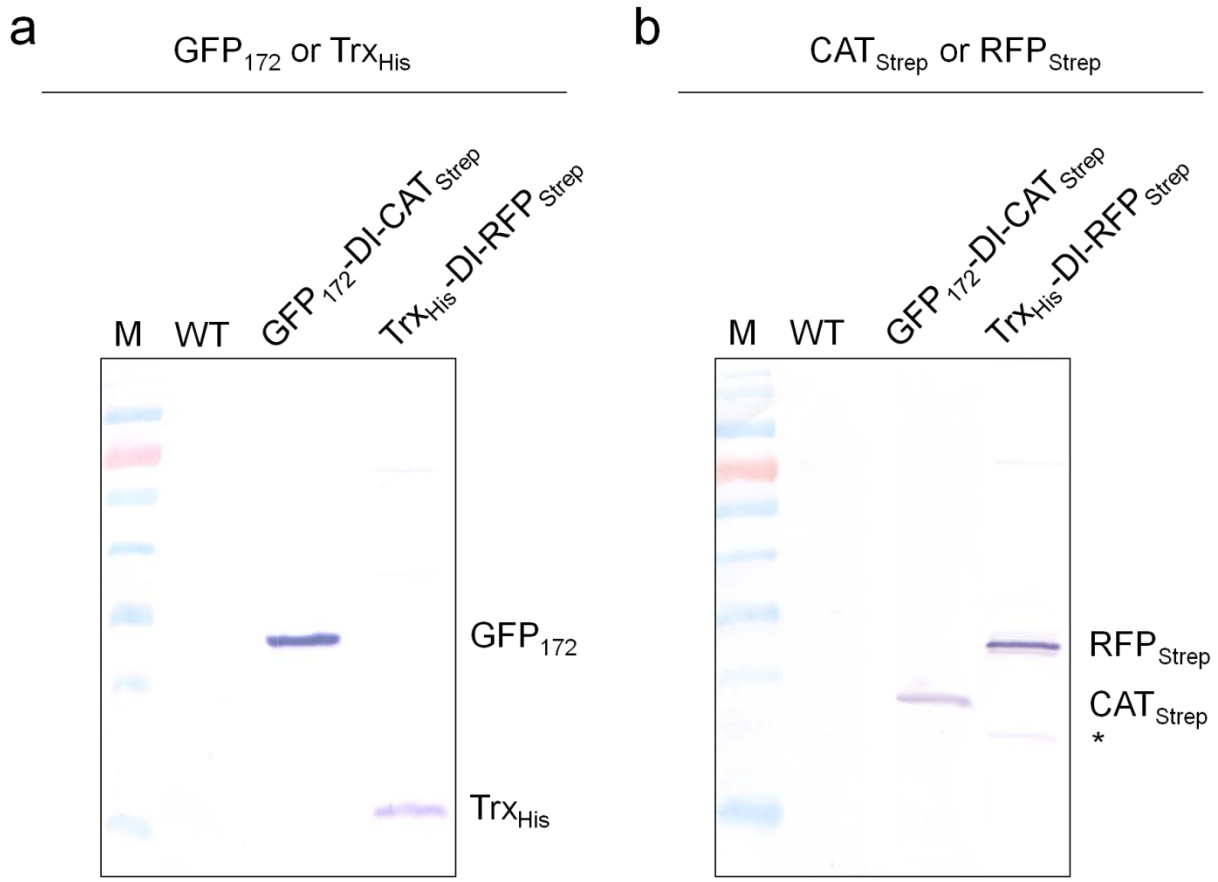


Figure 3.9. DI polyprotein system is also applicable to co-expressing proteins other than fluorescent reporters. Efficient release of protein constituents from GFP_{172}-DI-CAT_{Strep} and Trx_{His6}-DI-RFP_{Strep} polyproteins was detected using (a) anti-His tag, and (b) anti-Strep tag western blots.

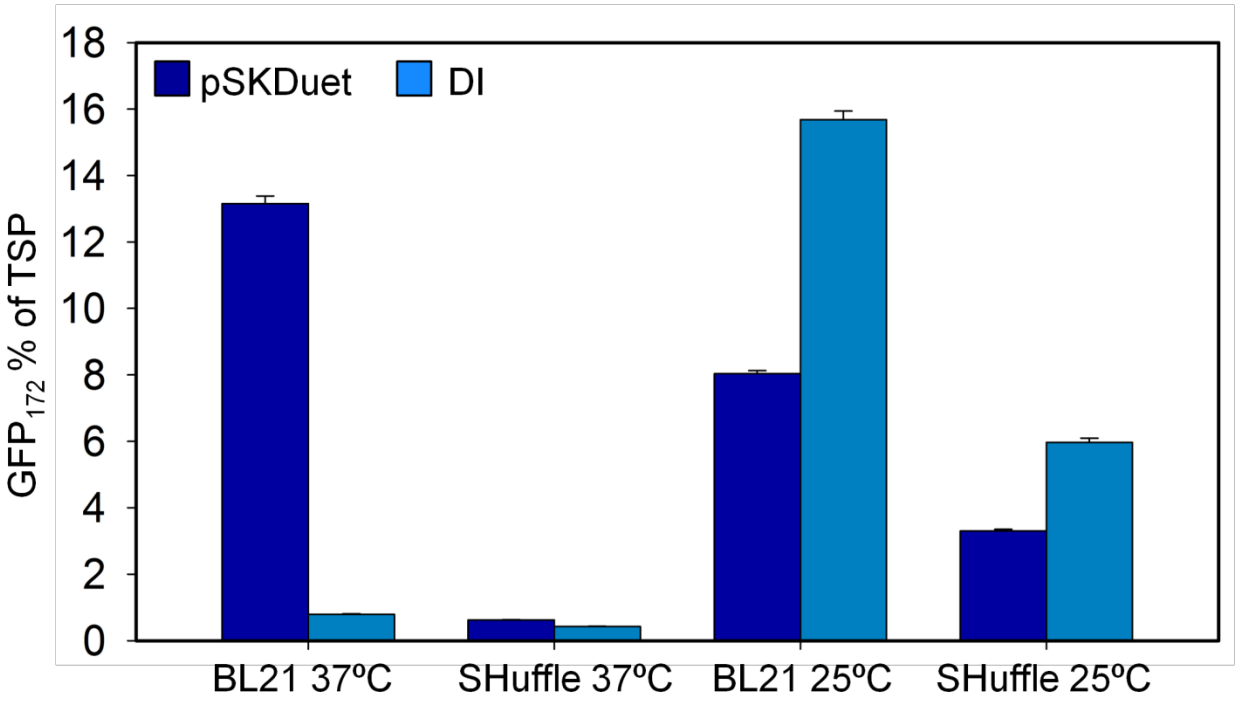


Figure 3.10. Comparison of protein expression levels using the DI based polyprotein vector vs. the duet vector in *E. coli* BL21(DE3) or SHuffle strain under the control of T7 promoter. Protein expression was induced by 0.1 mM IPTG at 37°C vs. 25°C. Protein level after overnight expression was determined based on the percentage of GFP_{172} in total soluble proteins, and normalized by considering the highest expression sample (i.e., expression with DI in BL21 at 25°C) as 100%. Amount of GFP172 was converted from the GFP fluorescence intensity of total soluble protein extracts using calibration curve established by purified protein standard. Total soluble proteins in cell extracts was measured based on Bradford protein assay.

Figure 3.11. DI domain without the N-terminal linker extension preserves high autocleavage activity (cf. Figure 2 for DI containing the N-terminal linker extension sequence LEGGSKFAND). Cell extract of *E. coli* expressing the GFP_{172}-DI-RFP_{Strep} polyprotein containing a DI domain lacking the N-terminal linker extension between GFP_{172} and *Ssp* DnaE intein was analyzed using western blots probed with anti-GFP (left) and anti-Strep tag (right) antibodies, respectively.

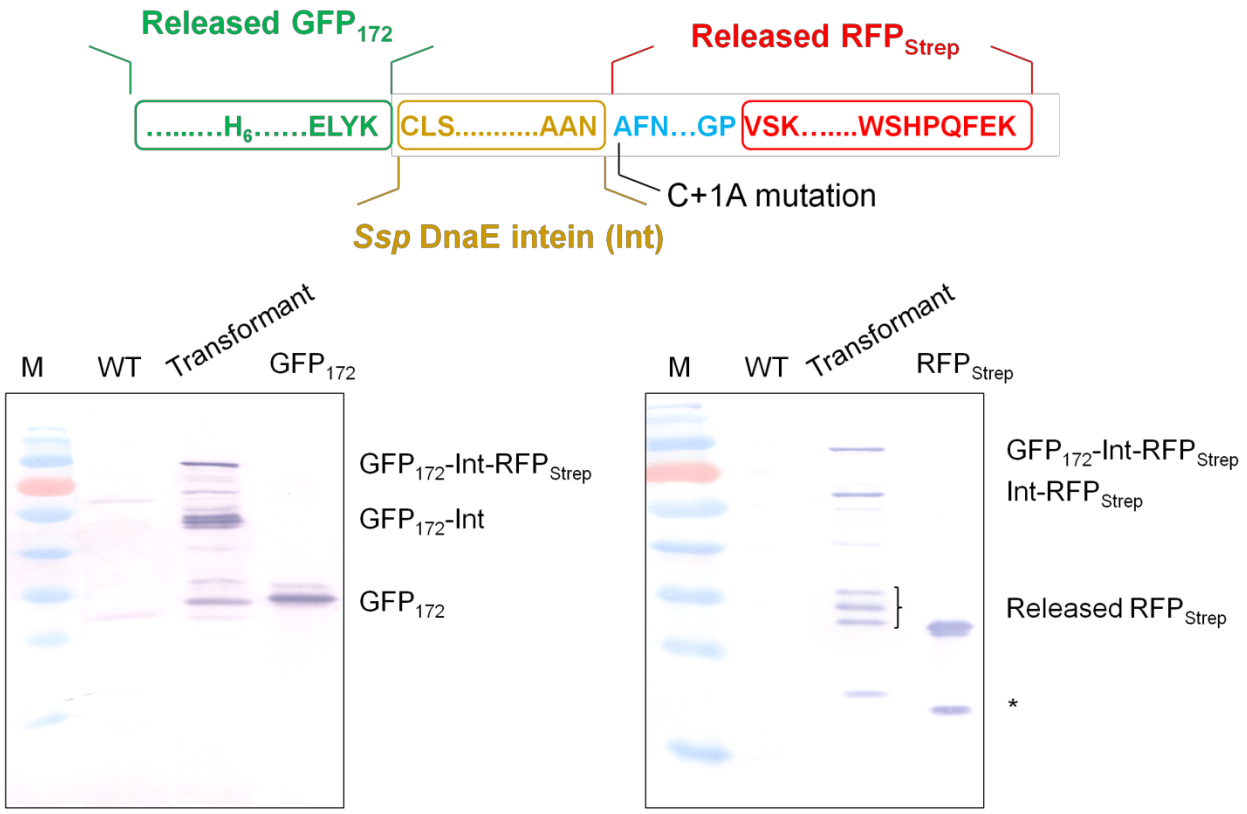


Figure 3.12. Inefficiency of single intein mediated polyprotein processing in *E. coli* (cf. Figure 3.11). A polyprotein vector containing a single Ssp DnaE intein domain was created and transformed into *E. coli* BL21(DE3). The Ssp DnaE intein domain has Asn at its C-terminus (as opposed to Ala in the DI domain), and the C+1 Cys is mutated to Ala. Protein extract of *E. coli* expressing the polyprotein coding sequence was analyzed using anti-GFP and anti-Strep tag western blots to detect release of the upstream and downstream proteins, respectively.

CHAPTER 4

IMPROVING THE SELF-CLEAVING POLYPROTEIN EXPRESSION SYSTEM BY HARNESSING THE SYNERGY BETWEEN AN INTEIN AND A FMDV 2A SEQUENCE

Abstract

Stoichiometric co-expression of multiple proteins in plants has ample applications in crop improvement and protection. However, currently available strategies for multi-protein co-expression have significant limitations. In this study, we developed a novel "IntF2A" autoprocessing protein unit that comprises an N-terminal cleavage *Ssp* DnaE intein mutant linked to a C-terminal cleavage 2A peptide from the foot-and-mouth disease virus (F2A) to direct co-expression of multiple proteins from a polyprotein vector in plants. We successfully demonstrated efficient processing of a GFP-IntF2A-RFP polyprotein precursor into physically separated GFP and RFP in tobacco NT1 cells. Efficient co-expression is also attained in transgenic tobacco plants, as well as agroinfiltrated leaf tissues of *Nicotiana benthamiana* and *Lactuca sativa L. var. longifolia*, and maize black Mexican sweet (BMS) cells. Utility of the IntF2A based polyprotein system was further extended to co-expression of other proteins of interest (POIs) including a bacterial enzyme (chloramphenicol acetyltransferase) and a human

cytokine (granulocyte-macrophage colony-stimulating factor). The processed POIs were confirmed to be functional. Independent cellular targeting of the released POIs can be achieved. By optimizing the linker sequence flanking the intein C-terminus, POI released from the N-terminus of the IntF2A domain can preserve its native C-terminal residues. Together, our results indicate the usefulness of the new IntF2A based polyprotein system to facilitate coordinated expression of multiple proteins in plants.

4.1. Introduction

Coordinate expression of multiple proteins in plants is an essential tool to unravel fundamental mechanisms as well as to enable development of next-generation crops with improved traits. To date, by co-expressing multiple proteins, crops have been successfully modified to acquire enhanced abiotic tolerance, improved pathogenic resistance, as well as enriched nutritional contents.[58,153-154] By introducing heterologous or synthetic metabolic pathways encoded by multiple genes, transgenic plants can be used for remediation of environmental pollutants and manufacture of products that exhibit high industrial or medicinal values.[73,155-157]

Currently, co-transformation of multiple monocistronic expression cassettes is the most prevailing approach for creating genetically modified crops with multiple stacked transgenes.[158] However, this gene stacking approach requires cumbersome and costly screening. Moreover, coordinate expression of the resulting protein products often calls for extensive tuning of the promoter and regulatory elements used. To this end, polyprotein expression systems offer unique opportunities to attain stoichiometric co-expression of multiple proteins.[4-5] With the polyprotein

expression system, multiple constituent POIs are connected by intervening protein processing elements, forming the polyprotein precursor which is subsequently processed into physically separated POIs and the processing elements. In early studies, endogenous protease activity was exploited for multiple protein expression in plants. In those cases, constituent proteins were connected by the putative substrate which can be subsequently digested by endogenous plant proteases inside the cells.[61-62,65,136] However, processing of polyprotein can only occur in a certain cellular compartment within particular hosts where the specific proteases are located. This limitation can be potentially resolved by co-expressing an exogenous protease flanked with its substrate, together with the constituent proteins. A widely reported example in this class of approach is based on the nuclear inclusion (NIa) proteins of plant potyviruses. The feasibility of this NIa based polyprotein approach for protein co-expression was demonstrated in several independent researches. For example, co-expressing and processing of coat proteins for tobacco mosaic tobamovirus (TMV) and soybean mosaic potyvirus (SMV) from the TEV-NIa based polyprotein successfully conferred pathogen-derived protection on transgenic tobacco plant.[63,79] Subcellular localization signal can also be effectively recognized in the context of the polyprotein with NIa protease originated from tobacco vein mottling virus (TVMV).[64] However, proteins expressed from this NIa based polyprotein were generally low yielded and imbalanced co-expressed, which is consequent on the intrinsic sequence of NIa protease. The translated polyprotein could be targeted to nucleus and subsequent for degradation due to the highly effective nuclear targeting potency of NIa protease.[63,159] Although, increased protein expression level could be obtained after deletion of the internal nuclear targeting signal within the NIa protease sequence, residual amino acids from the substrate sequence may destabilize the attached

protein upon releasing from polyprotein precursor. Moreover, cellular targeting of proteins expressed from the NIa based polyprotein is influenced by the location of NIa protease within the polyprotein context, due to the slow proteolytic action compared to the protein transportation.[64]

Since the discovery of the unique ribosomal skipping mechanism of 2A peptide within the native FMDV polyprotein,[51] this small and versatile peptide has been widely applied in multi-protein co-expression in a wide range of eukaryotic hosts.[77,160] The residual 2A amino acids that remain on the carboxyl terminus of the processed proteins could however hinder protein activity and cellular targeting.[56-57] The potential adverse effects of C-terminal 2A extension can be eliminated by incorporating a cleavage linker peptide immediately upstream of 2A sequence. When the LP4 linker peptide within the native polyprotein from the seeds of *Impatiens balsamina* was applied, cleavage of the LP4 linker was not occurred in cytosol of transgenic tobacco.[58] The same LP4/2A hybrid linker used for co-expressing two secretory plant defensins caused mistargeting of processed upstream protein to vacuole, which is probably due to the potential vacuole localization of the putative proteases responsible for the LP4 cleavage.[57] While successful removal of the C-terminal 2A extension has been reported in mammalian system by incorporating a furin substrate site preceding the 2A sequence, this strategy is only applicable for expressing secretory proteins due to the *trans*-Golgi localization of endogenous furin protease.[59-60,118]

Recently, we reported an autocleavage domain, called the "dual-intein" of "DI" domain, which comprises an N-terminal cleavage intein mutant and a C-terminal cleavage intein mutant connected in tandem, for synchronized co-expression of multiple proteins.[161] The DI domain

based co-expression is applicable to both prokaryotes and eukaryotes. In plants, while the DI based polyprotein system allows highly efficient release of upstream POI, release of the POI downstream of the DI domain was less efficient due to reduced C-terminal cleavage activity of the intein mutant at lower temperature (i.e., room temperate suited for plant growth). As a result, equimolar expression of the POIs was not attained. The partially cleaved DI-downstream POI fragment may be degraded in plants, leading to less accumulation of the downstream protein compared to its upstream counterpart.

In this study, we hypothesized that by replacing the downstream intein in the DI domain with the F2A sequence to create the IntF2A domain; this new domain could endow improved polyprotein autoprocessing. This is based on the hyperactive N-terminal autocleavage we observed with the DI domain in plants, and the ample data in the literature with the F2A system. This IntF2A domain mediates release of flanking proteins via a two-stage co-translational/posttranslational autocleavage process. Here we investigated the *in vivo* processing of the IntF2A-based polyproteins in tobacco cells and several other plant species. We also optimized the IntF2A domain to preserve native C-terminal residues of the released POI.

4.2. Materials and Method

Construction of the IntF2A constructs.

Multiple IntF2A polyprotein cassettes co-expressing different POIs were assembled using routine molecular cloning techniques. Within these constructs (Fig. 4.1), upstream POI is comprised of a GFP variant (GFP_{His} or GFP_{172})[129] or a N-terminal His tagged human cytokine ($_{His}$hGMCSF), whereas the downstream POI is comprised of a C-terminal Strep tagged RFP

variant (RFP_{Strep}) or chloramphenicol acetyltransferase (CAT_{Strep}), forming GFP_{172}-IntF2A-RFP_{Strep} (pND-1), $_{His}$hGMCSF-IntF2A-RFP_{Strep} (pND-2), and GFP_{172}-IntF2A-CAT_{Strep} (pND-3), respectively. To assemble pND-1, the coding sequence of the IntF2A domain containing an *Ssp* DnaE mini-intein (N159A) mutant with five N-extein residues (KFAND) previously reported to accelerate N-terminal cleavage activity,[128] along with three native C-extein residues (CFN), connected to the N-terminal of 59 aa F2A sequence by flexible GS linker was synthesized (Genscript, Piscataway, NJ) and ligated into pUC57 vector between *Xho*I and *Apa*I, to yield pUC-Int(ND/CFN)F2A. The coding sequence of GFP_{172} was amplified from pGEM5z-GFP172 vector using primers *Sal*I-mGFP5 and GFP-*Xho*I-R, to generate *Sal*I-GFP_{172}-*Xho*I fragment. The coding sequence of RFP_{Strep} from pETMD[130] vector was amplified by two-step PCR with primers (mCherry-2A-F, mCherry-STREP-R and GKZ-R), to generate *Apa*I-RFP_{Strep}-*Sac*I fragment. The *Sal*I-GFP_{172}-*Xho*I fragment and *Apa*I-RFP_{Strep}-*Sac*I fragments were successively ligated into pUC-Int(ND/CFN)F2A vector digested with corresponding restriction enzymes, to yield pUC-ND-1. To assemble pND-2, $_{His}$hGMCSF fragment was amplified from pE1804-GKG vector[136] using the forward primer (*Sal*I-His-GMCSF-F) with a 5'-*Sal*I site and the reverse primer (GMCSF-*Xho*I-R) with a 3'-*Xho*I site. The resulting *Sal*I-$_{His}$hGMCSF-*Xho*I fragment was ligated into *Sal*I-*Xho*I digested pUC-ND-1 vector to replace the upstream GFP_{172} fragment, to yield pUC-ND-2. To replace upstream protein with GFP_{His}, coding sequence of GFP_{His} with the N-terminal *Arabidopsis thaliana* basic chitinase signal peptide (SP1) was first synthesized by Genscript and placed between *Sal*I and *Xho*I sites of pUC57 vector to yield pUC-SP1GFP_{His}. *Sal*I-GFP_{His}-*Xho*I fragment was excised from pUC-SP1GFP_{His} vector and ligated into pUC-nND-1 vector after *Sal*I-*Xho*I double enzyme digestion, to yield pUC-GFP_{His}-Int(ND/CFN)F2A-

RFP_{Strep}. Coding sequence of CAT_{Strep} amplified by two-step PCR from the template pET-CAT using primers (CAT-2A-F1, CAT-STREP-R and GKZ-R) that incorporate an *Apa*I site and a *Sac*I sites at the 5'- and 3'-ends, respectively. The resulting *Apa*I-CAT_{Strep}-*Sac*I fragment was subsequently ligated into pUC-GFP_{His}-IntF2A-RFP_{Strep} vector digested with *Apa*I and *Sac*I restriction enzymes, to replace the downstream POI with C-terminal Strep tagged CAT protein (pUC-ND-3). The feasibility of IntF2A based polyprotein for co-expression of three proteins was investigated using construct harboring three fluorescent proteins, GFP_{172}, $mKO1_{FLAG}$ and RFP_{Strep} connected by IntF2A autoprocessing domain. To assemble this IntF2A mediated three protein co-expression vector, *Apa*I-RFP_{Strep}-*Sac*I fragment in pUC-ND-1 vector was replaced with *Apa*I-$mKO1_{FLAG}$-*Sac*I fragment amplified by two-step PCR from pET-His6-mKO1 using primers mKO1-*Apa*I-F, mKO1-R and FLAG-*Sac*I-R1. The resulting pUC-GFP_{172}-Int(ND/CFN)F2A-$mKO1_{FLAG}$ vector was digested with *SacI* enzyme and ligated with *Sac*I-RFP_{Strep}-*Sac*I amplified using primers G11c-*Sac*I-mChe-F and GKZ-R, to yield pUC-ND-4. Orientation of the inserted RFP_{Strep} fragment was screened by colony PCR using primers M13F and G11c-*Sac*I-mChe-F.

To preserve the native C-terminus of processed upstream protein, GFP_{172} and DnaE intein direct fusion were created by utilizing the in-frame *Bst*BI site near the C-terminal region of GFP_{172} and the *Xba*I site at 3'-end of Int(EY/CSC) fragment. This was carried out by assembling the *Bst*BI site along with the remaining C-terminal region of GFP_{172} onto the N-terminus of DnaE intein (EY/CSC) mutant by overlapping PCR using primers (*Bst*BI-G172C-F, G172-DnaE-R, G172-DnaE-F and *Xba*I-SC-R). The resulting *Bst*BI-GFP_{172}C-Int(EY/CSC)-*Xba*I fragment was ligated into pUC-ND-1 vector digested with corresponding restriction enzymes

(*Bst*BI and *Xba*I), to yield the fluorescence based IntF2A polyprotein construct without N-extein overhang (pUC-SC-1).

To evaluate different cellular targeting of POIs processed from the IntF2A based polyprotein precursor, the secretory signal peptide was incorporated into the N-terminus of the either the upstream protein (pND-5) or both upstream and downstream proteins (pND-6). In the former case, SP1 was appended onto the N-terminus of GFP_{172} by ligating the *Eco*RI-GFP_{172}-*Xho*I fragment amplified from pGEM5z-GFP172 vector with primers mGFPS2E and GFP-*Xho*I-R into pUC-SP1-GFP_{His} vector, to generate pUC-SP1-GFP_{172}. The *Sal*I-SP1-GFP_{172}-*Xho*I fragment was then excised from pUC-SP1-GFP_{172} and ligated into pUC-ND-1 vector, to generate pUC-ND-5. To introduce ER targeting signals to the N-terminus of both upstream (GFP_{172}) and downstream (RFP_{Strep}) proteins, rice α-amylase signal peptide (SP2) was first introduced to the N-terminus of RFP_{Strep} by replacing the mGFP5 in pBluescript-SP1-GMCSF-2A-SP2-mGFP5 vector (created by M. Rapolu) with *Sph*I-RFPStrep-*Sal*I, which was amplified from pUC-ND-1 vector using primers mCherry-*Sph*I-F and GKZ-R. The *Apa*I-SP2-RFP_{Strep}-*Sac*I fragment was then excised and ligated into pUC-ND-5 vector, to generate pUC-ND-6.

N-terminal inactive Intein mutant was generated by site-direct mutagenesis using primers (DnaE-*Xho*I-F-2 and DnaE-*Xba*I-R-2) to introduce an alanine mutation at the first cysteine residue (C1A mutation).[142] This N-inactive intein fragment was used to replace the active DnaE intein between *Xho*I and *Xba*I sites in pUC-ND-1 vector, to generate pUC-N(-)-1.

To examine whether intein destabilized the IntF2A domain after polyprotein processing, intein fragment was removed by replacing the GFP_{172}-Int(CSC) fragment in pUC-SC-1 vector

with GFP_{172} fragment amplified with primers (*Sal*I-mGFP5-F and GFP-GS-*Xba*I-R) that anneal to the template pGEM5z-GFP172 vector, to create pUC-2A-1.

All of the assembled IntF2A based polyprotein cassettes were mobilized into binary vector pE1775 between *Sal*I and *Sac*I sites, in which protein expression in plants was driven by the mannopine/octopine synthase $(ocs)_3$/mas promoter.[87] The sequences of all the constructs were verified by DNA sequencing (performed by ASGPB, University of Hawaii at Manoa).

Plant transformation and protein extraction.

pE1775 vectors containing IntF2A based polyprotein sequences were transformed into *Agrobacterium tumefaciens* C58C1 via eletroporation. Stable transformation of tobacco NT1 cells and *N. tabacum* plants was performed using *agrobacterium* co-cultivation approach as described in (section2.2). Transformants were selected on MS medium[92] supplemented with hygromycin. Vacuum assist agroinfiltration for transient expression in leaf tissues of *N. benthamiana* and Romaine lettuce was performed as described in Section 2.2. Fluorescence and western blot were used to screen the highest expression lines which were selected for subsequently characterization. Total soluble intracellular proteins were extracted in boric acid extraction buffer following the procedures described in section 2.2. For collecting secreted proteins, apoplastic fluid or spent media of transgenic NT1 calli or suspension cultures, respectively, were filtered through a 10μm nylon gauge. The crude filtrates were clarified by filtering though a Whatman #1 filter paper. Total protein concentration was determined using Bradford protein assay (Bio-Rad, Hercules, CA)[162] and target proteins were tracked by fluorescence measurement.

SDS-PAGE and western blot.

Protein extracts were mixed with SDS-PAGE sample buffer and denatured at 95 ºC for 5 min. After brief centrifugation, denatured protein samples were subjected to 12% SDS-PAGE, followed by western blot analysis as described in section 2.2. Primary antibodies rabbit anti-GFP (Invitrogen, Grand Island, NY) was used to detect GFP_{172} and GFP_{His}. Rabbit anti-Strep tag (Genscript, Piscataway, NJ) was used to detect RFP_{Strep} and CAT_{Strep}. Rabbit anti-human GMCSF (BioVision, Milpitas, CA) was used to detect $_{His}$hGMCSF. Alkaline phosphatase conjugated goat anti-rabbit or goat anti-mouse (Southern Biotech, Birmingham, AL) were used as the secondary antibodies.

Fluorescence confocal microscopy.

Cellular destination of processed proteins was analyzed using an Olympus Fluoview FV-1000 confocal laser scanning microscope system mounted on an Olympus IX-81 inverted microscope (Nikon, Tokyo, Japan) (preformed at the Biological Electron Microscopy Facility at University of Hawaii at Manoa). Cell images were observed with a UPLSAPO 20X lens (numerical apertures (NA), 0.70; Nikon, Tokyo, Japan), along with 3X digital zooming. Cells were excited with laser beams at 488 nm and 543 nm, respectively, for detection of GFP and RFP (mCherry) fluorescence. Filters BA505-525 and BA560IF were used for collecting fluorescence emission of GFP and RFP (mCherry), respectively.

For wortmannin treatment, 2mM of wortmannin stock solution was dissolved in dimethyl sulfoxide (DMSO) and added to the suspension NT1 cell culture at the final concentration of 20 µM. Cells were harvested after 24 hour treatment at 28 ºC and examined by confocal fluorescence microscope under the condition as described above.

Purification of processed proteins for N-terminal amino acid sequencing and ESI-TOFMS analysis.

Processed GFP_{172} from ND-1 and SC-1 cells, as well as RFP_{Strep} from ND-1 cells were purified using the method described in Section 3.2 and 3.2, respectively. N-terminal amino acid sequencing of processed proteins was performed by the Protein Facility at Iowa State University as described in Section 2.2. ESI-TOFMS analysis of purified processed GFP_{172} protein was carried out by Z. Liang and Dr. P. Williams from Department of Chemistry, University of Hawaii at Manoa following the procedure described in Section 2.2.

Biological activity assay of processed $_{His}$hGMCSF and CAT_{Strep}.

GMCSF dependent human TF1 cell proliferation assay was performed as described in Section 2.2. Enzyme activity assay of processed CAT_{Strep} protein was analyzed as described in Section 3.2.

4.3. Results

Design and construction of the IntF2A based polyprotein system for coordinating multi-protein co-expression in plants.

IntF2A based polyprotein cassettes were assembled by connecting the upstream POI and the downstream POI with the IntF2A autoprocessing domain that enables self-excision at both terminal junctions (Fig. 4.2). As the mRNA encoding the polyprotein precursor is being translated, the downstream POI is released through the co-translational ribosomal recoding activity of the F2A peptide. To maximize the F2A activity, a 58 aa FMDV 2A sequence which has been reported to show enhanced C-terminal cleavage efficiency was used.[49] This F2A

sequence also includes a portion of the FMDV 1D capsid protein in addition to the core F2A sequence. After co-translational F2A cleavage, upstream POI can be liberated by the N-terminal cleavage activity of the *Ssp* DnaE mini-intein modified with N159A mutation. The N159A mutation is known to significantly attenuate splicing activity and promote cleavage activity at its N-terminal junction.[142] To further enhance N-terminal cleavage efficiency, dipeptide ND was inserted at the intein N-terminal junction for the pND constructs while the pSC construct incorporates the CSC sequence at the intein C-terminal junction (Fig. 4.1). Each of these modifications was previously shown to augment N-terminal cleavage efficiency of the Ssp DnaE mini-intein based on random extein library screening.[128] For the pND constructs, the peptide sequences "LEGGSKFAND" and "CFNGSGSR" are flanking the N- and C-termini of the intein domain, respectively. Within these linker sequences, amino acids LE and SR were resulted from the *Xho*I and *Xba*I cloning sites whereas amino acids KFA and CFN were derived from the native flanking extein sequences. A flexible GS linker (GGS or GSG) was inserted between the native flanking extein residues and the cloning sites. For the pSC construct (Fig. 4.1), the intein N-terminal linker extension is completely eliminated to avoid adding any extraneous residues to the C-terminus of POI1, i.e., GFP_{172} (GFP variant with an internal His tag inserted between amino acid 172 and 173). The native flanking C-extein residues CFN were used to replace CSC to augment the intein N-terminal cleavage efficiency. Fluorescent proteins GFP_{172} and RFP_{Strep} (RFP with a C-terminal Strep tag) were used in most polyprotein cassettes as the upstream and downstream POI, respectively. *Arabidopsis thaliana* basic chitinase signal peptide (SP1) and rice α-amylase signal peptide (SP2) were used to direct ER targeting of upstream and downstream POIs, respectively, for investigating the subcellular targeting of proteins expressed from the

IntF2A-based polyprotein. In addition to fluorescent proteins, upstream or downstream POI was replaced by GMCSF with an N-terminal His tag or CAT with a C-terminal Strep tag, respectively, in the pND vector system to test general utility of the IntF2A based polyprotein system for multi-protein co-expression. To further verify that the release of upstream POI is attributed to the N-terminal intein cleavage activity, N-terminal cleavage inactive version of IntF2A domain (N-) was created by introducing C1A mutation to the intein domain, while using the native flanking extein sequences (LEY and CFN, respectively) to minimize cleavage activity.

Processing of the IntF2A based polyprotein in plants.

Processing of the IntF2A based polyprotein in plants was initially characterized using western blot analysis of total protein extract of tobacco NT1 expressing pND-1. As shown in Figure 4.3, essentially complete release of both POIs, i.e., GFP_{172} and RFP_{Strep}, were observed on western blots probed with anti-GFP and anti-Strep tag antibodies, indicating efficient autocleavage at both N- and C- terminal junctions of the IntF2A domain. When the ND-1 polyprotein was expressed in *N. tabacum* cv. Xanthi plants, similar efficient release of both upstream GFP_{172} and downstream RFP_{Strep} could be detected in leaf, stem and root extracts (Fig. 4.4a & b) compared with that observed in undifferentiated tobacco NT1 cells. Highest expression was achieved in the root tissue, which is due to the unique characteristic of the $(ocs)_3$/mas promoter used to drive protein expression (Fig. 4.4c). Similar tissue specific expression using the same $(ocs)_3$/mas promoter was observed in our previous studies in transgenic tobacco, as well as in the studies reported by Lee *et. al.* in transgenic tobacco and maize plants.[132,136,161] In addition to tobacco, efficient processing of the ND-1 polyprotein was observed in the leaf tissue of different plant species including *N. benthamiana*, Romaine lettuce (*Lactuca sativa L. var.*

longifolia), as well as in maize BMS cells based on transient expression via agroinfiltration (Fig. 4.5). These results support the general utility of the IntF2A polyprotein system in a wide range of plant species, spanning from monocot to dicot, for efficient production of multiple proteins. Although further testing in stably transformed maize plants is necessary, demonstration of efficient processing of the IntF2A based polyprotein in maize is of particular significance as it is one of the most important biotech monocot crops with enormous commercial value globally.

When examined using fluorescence microscopy, tobacco NT1 cells expressing pND-1 displayed bright green and red fluorescence (Fig. 4.6a). Characteristic GFP and RFP spectra were detected in the protein extracts of both transgenic tobacco NT1 cells and tobacco plants (Fig. 4.6b). Since complete cleavage of the IntF2A based polyprotein precursor was confirmed by western blot analysis (cf. Fig. 4.3 & Fig. 4.4), detection of fluorescence confirmed that constituent proteins are functional upon release from the IntF2A based polyprotein precursor.

The IntF2A polyprotein system allows coordinated co-expression of more than two proteins in plants. This was illustrated by co-expressing three fluorescent reporter proteins: GFP_{172}, orange fluorescent protein mKO1 with a FLAG tag at the C-terminus ($mKO1_{FLAG}$), and an RFP_{Strep} (ND-4; Fig. 4.1) in tobacco NT1 cells. As shown in Fig. 4.7a-c, nearly complete separation of all three proteins was demonstrated by western blots probed with anti-GFP, anti-Strep tag and anti-mKO1 antibodies, respectively. As with the observation in the two-protein co-expression construct, i.e., pND-1, all processed fluorescent reporters are found functional as evident from the characteristic fluorescence spectra detected using fluorescence spectroscopy (Fig. 4.7d-f).

In the IntF2A based polyprotein system, multiple constituent proteins were translated from a single transcript, thus it is possible that equal molar protein co-expression can be achieved providing posttranslational events do not cause differential destabilization of the released POIs. Using fluorescence and western band densitometry quantification, processed constituent POIs from ND-1 and ND-4, respectively, were accumulated at similar levels inside the cells (Fig. 4.8). More importantly, no significant reduction in expression level was noted when three proteins were co-expressed using the IntF2A based polyprotein system (0.54% TSP) as compared to those co-expressing two proteins (0.24% TSP). These expression levels were comparable to those obtained using the single protein vectors where GFP_{172} or RFP_{Strep} alone was expressed from the same $(ocs)_3$/mas promoter (0.29% or 0.61%TSP, respectively) (Fig. 4.8).

Application of the IntF2A based polyprotein for co-expressing multiple proteins other than fluorescence reporters.

To demonstrate the general utility of the IntF2A based polyprotein expression system, pND-1 was modified to create constructs pND-2 and pND-3 (Fig. 4.1). In ND-2, $_{His}$hGMCSF (hGMCSF with N-terminal His tag) was used as the upstream POI while in ND-3, CAT_{Strep} was used as the downstream POI. Processing of polyprotein precursor was examined using western blots probed with anti-GFP, anti-Strep tag, or anti-GMCSF antibody for detection of GFP, CAT_{Strep}/RFP_{Strep} or $_{His}$hGMCSF, respectively. Similar to the fluorescent reporter based construct, i.e., ND-1, a single strong immunoreactive band corresponding to the processed proteins was detected when the ND-3 extracts were probed with anti-GFP and anti-Strep tag antibodies (Fig. 4.9a). For ND-2, while intense band of processed RFP_{Strep} was detected by western blot, only a faint immunoreactive band corresponding to a putative $_{His}$hGMCSF dimer was detectable in

concentrated ND-2 extract (Fig. 4.9b). We can rule out protein glycosylation as a plausible cause for the increased molecular size of the released hGMCSF because it was designed for cytosolic accumulation (Fig. 4.9b). Nevertheless, ND-2 extract can support the growth of hGMCSF-dependent TF1 cells in a dose dependent manner, activity of processed $_{His}$hGMCSF estimated based on RFP fluorescence was over 10-fold lower than the commercial *E. coli* expressed hGMCSF (spiked in 10 µg of NT1 WT extract) at the same concentration (Fig. 4.10). The reduced bioactivity in combination with the weak western band suggests degradation of the $_{His}$hGMCSF in plant cytosol after its release from the IntF2A polyprotein precursor.

Determination of cleavage sites of IntF2A based polyprotein.

To investigate whether the observed cellular processing of polyprotein precursors was indeed resulted from the autocleavage activity of the IntF2A fusion domain, N-terminal cleavage-inactive intein mutant was introduced by introducing the C1A mutation in the intein domain in pND-1 to create pN(-)-1 (Fig. 4.1). As shown in Fig. 4.11a & b, release of upstream POI was impaired with this mutation, while complete release of the downstream protein was still observed. This result not only confirms that the N-terminal cleavage activity of the *Ssp* DnaE intein domain was solely responsible for the release of the upstream POI, but also demonstrates that the intein and F2A function independently as inactivation of N-terminal cleaving intein domain does not interfere with the function of the F2A peptide.

The exact cleavage sites within the IntF2A polyprotein were determined by ESI-TOFMS and N-terminal sequencing of processed POIs purified from the extract of transgenic NT1 cell expressing ND-1. Processed downstream POI possesses a proline followed by its native N-terminal sequence (VSKGEE) according to N-terminal amino acid sequencing (Fig. 4.11c). This

result is consistent with the proposed mechanism of F2A-mediated C-terminal cleavage in which peptide bond formation between the last two amino acid residues, G and P, of the 2A sequence is disrupted during translation. On the other hand, N-terminus of the processed upstream GFP_{172} was found to be blocked as revealed by the N-terminal amino acid sequencing result. ESI-TOFMS data demonstrated that the molecular mass of processed upstream GFP_{172} from the ND-1 polyprotein precursor matches the properly processed GFP_{172} protein with an N-terminal acetylation modification (Fig. 4.11c). Collectively, our results support that the observed cellular processing of the IntF2A based polyprotein was indeed mediated by the specific actions of intein and F2A.

Cellular targeting of proteins processed from the IntF2A based polyprotein.

In order to examine the feasibility of the IntF2A based polyprotein for co-expressing secretory proteins, ER targeting signals were incorporated into the N-termini of both upstream and downstream proteins to direct protein secretion. When pND-6 was expressed in tobacco NT1 cells, both secreted GFP_{172} and RFP_{Strep} proteins were detected in the spent media (Fig. 4.12a). Cellular localization of the processed GFP_{172} and RFP_{Strep} expressed from pND-6 was further examined using confocal fluorescence microscopy. As shown in Fig. 4.12b, different from the nuclear localization of its cytosolic counterpart as in ND-1, secretory upstream GFP was accumulated around the cell envelope and perinuclear endoplasmic reticulum (ER), whereas RFP was observed in the secretory network and accumulated in the vacuole. When the cells were treated with wortmannin, an antibiotic that specifically inhibits vacuole targeting in plants,[163] increased amount of secretory RFP was detected in concentrated spent media, although

considerable RFP fluorescence remained in vacuole as revealed by the confocal fluorescence microscopy (Fig. 4.12a & b).

In an additional construct, pND-5, where the ER targeting signal was only incorporated to the N-terminus of the upstream protein, only secreted GFP was detected in concentrated apoplastic fluid, even though both distinctive GFP and RFP fluorescence was detected in intracellular extract (Fig. 4.12a). The differential targeting of processed proteins was also confirmed confocal fluorescence microscopy. As shown in Figure 4.12b, similar to that of the secretory construct, processed GFP was mainly accumulated in the apoplast, whereas RFP was mainly accumulated inside the cytoplasm as with the cytosolic construct. Collectively, our results clearly demonstrated the feasibility of the IntF2A based polyprotein system for producing proteins designating to different cellular compartments. This provides strong evidence that F2A cleavage happened co-translationally as the cytoplasm located downstream protein did not leak through the ER membrane when the ER targeting signal was incorporated to the N-terminus of the upstream POI.

Processing of the IntF2A based polyprotein without N-extein extension.

One limitation of the IntF2A based polyprotein system described above is that residual amino acids within the N-extein linker remain attached to the C-terminus of the released upstream POI. Although proper function of processed upstream GFP_{172} was observed even with the additional non-native sequence at its C-terminus as discussed above, other proteins might not tolerate such terminal modification. Previous study suggested that intein N-terminal autocleavage can be accelerated not only by modifying amino acid residues flanking the N-terminal intein junction, mutation of amino acids behind the conserved C+1 Cys (first C-extein residue flanking the C-

terminus of the intein domain), at C+2 and C+3 positions (the second and third C-extein residues downstream of intein C-terminus) (cf. Fig. 4.1) alone can be sufficient to confer high N-terminal cleavage efficiency of intein.[128] As the C+2 and C+3 residues reside within the linker region between the intein and F2A domains it can be released as a part of the IntF2A domain, native C-terminus of processed upstream POI can be preserved after polyprotein processing. To prove this hypothesis, pND-1 construct was modified by removing the N-extein linker extension sequence and incorporating two residues Ser and Cys, which were previously reported to enhance N-terminal intein cleavage activity, at the C+2 and C+3 positions, respectively. The processing of this modified IntF2A based polyprotein (SC-1; Fig. 4.1) in plants was evaluated in transgenic tobacco NT1 cells based on the western blot analysis. As shown in Figure 4.13a & b, essentially complete release of both upstream and downstream POIs demonstrate that the cleavage efficiency of the IntF2A domain was not attenuated by the elimination of N-extein accelerating linker. N-terminal amino acid sequencing together with ESI-TOFMS analysis of processed protein revealed that the processing of the SC-1 polyprotein occurred at the expected sites and processed upstream GFP_{172} preserves its native C-terminus (Fig. 4.13c).

4.4. Discussion

In this study, we have developed an IntF2A fusion protein domain with hyperactive autocleavage activity to enable coordinated co-expression of multiple proteins in plants from a single transcript. The processing of the IntF2A-based polyprotein is initiated during protein translation, in which peptide bond formation between the last two amino acids of the F2A peptide, i.e., G and P, are disrupted by the unique ribosomal recoding activity of the F2A

sequence.[51] The 2A activity could be influenced by its upstream protein sequence since the ribosomal recoding event requires interaction of ribosomal exit tunnel with the nascent peptide that forms probable helical conformation.[49,52,126,164-165] Incorporation of additional residues derived from the native viral sequence upstream to the commonly used 18-22 amino acid F2A sequence has been reported to improve 2A cleavage efficiency.[34, 35] However, this modification further increases the overall length of the 2A sequence which may worsen the effects of 2A overhang that attached to the C-terminus of processed upstream protein. This potential adverse effect caused by the C-terminal 2A overhang can be eliminated by virtue of the N-terminal autocleavage activity of the intein mutant within the IntF2A fusion domain. While N-terminal intein cleavage was considered as a post-translational process since amino acids downstream of the intein sequence are involved in the cleavage initiation step, prior kinetic studies revealed that N-terminal intein cleavage rate is close to the protein synthesis rate of *Lemna minor* cells under rapid growth condition.[128,166-168] By incorporating N-terminal cleavage accelerating extein flanking residues, e.g., ND at intein N-terminal junction or CSC bordering the intein C-terminus, cleavage rate enhanced by four folds compared to that with the native flanking extein residues.[128] This high N-terminal intein cleavage activity facilitates rapid removal of the IntF2A overhang from the C-terminus of the upstream POI before the adverse effect occurs inside the cell.

Using the autoprocessing IntF2A fusion domain, two or three POIs, including fluorescent reporters, bacterial enzyme, and human cytokine, were successfully co-expressed in undifferentiated tobacco cells. Efficient processing of the IntF2A based polyprotein was also demonstrated in different tissues of transgenic tobacco plants, as well as in three other plant species. These results demonstrated that the intein domain and F2A peptide functioned

synergistically as a fusion protein in plants to direct protein co-expression from a single transcript. N-terminal amino acid sequencing and ESI-TOF mass spectrometry analysis revealed that processing of the IntF2A based polyprotein is consistent with the expected cleavage mechanisms mediated by intein variant and F2A peptide. Authentic C-terminus of processed protein can be preserved by optimizing the amino acid residues flanking the intein C-terminal junction in the IntF2A domain. Released POIs from the polyprotein precursor display proper function and POIs with similar stability were shown to accumulate at similar levels. Conversely, when a single F2A was used without being fused to the intein domain, fluorescence of processed upstream protein (GFP_{172}) was approximately one fold lower than downstream protein (RFP_{Strep}) (Fig. 4.14a). The lower accumulation of processed upstream protein was confirmed by scanning densitometry of western bands. These results suggest that the C-terminal F2A overhang does not hamper the fluorescence function of GFP_{172}; instead, it might cause degradation of GFP_{172} by serving as a putative peptide signal for proteolysis. This hypothesis was evidenced by the observation of the slightly smaller processed upstream product compared to the purified GFP_{172} standard on the western blot probed with anti-GFP antibody (Fig. 4.14b). N-terminal cleavage activity of the intein moiety in IntF2A prevents the degradation of processed upstream protein by rapidly removing C-terminal F2A overhang, and therefore enabled stoichiometric accumulation of constituent POIs (Fig. 4.14a).

When a single ER targeting signal was appended onto the N-terminus of the polyprotein sequence, i.e., ND-5, downstream RFP_{Strep} co-translationally released from the polyprotein precursor before leaking through the ER translocation channel and resided in the cytoplasm.[126,169] On the other hand, processed upstream GFP_{172} released from the polyprotein

precursor entered the secretory pathway and was eventually detected in the spent media of suspension culture. The confocal fluorescence microscope image showed that cellular GFP fluorescence was correctly localized to the reticulated network and accumulated around the cytoplasmic membrane. Similar observation was also found in a previous study when the F2A peptide alone (without fusing to intein) was used to direct co-expression of a secretory version of GFP and a cytosolic phleomycin resistance gene in transgenic tobacco leaf.[54] In that study, the F2A sequence was appended to the C-terminus of the secretory GFP. However, results from other studies suggested that the protein anterior to the F2A peptide intended for secretion is prone to accumulate in vacuole after being released from the F2A based polyprotein precursor.[56-57] Here we observed a similar phenomenon, i.e., when a protein is flanked by an ER-targeting signal and a F2A sequence at its N- and C-termini, respectively, the C-terminal F2A peptide appeared to direct the protein to accumulate in vacuole. This is evidenced from the results based on the expression of the pN(-)-2 vector (Fig. 4.1) in which an ER targeting signal was used to direct secretion of the upstream POI GFP_{172}, but the intein was mutated to inactivate N-terminal autocleavage. In this case, no secreted GFP_{172}-IntF2A fragment was detected in the spent medium of suspension cells. It is plausible that the large size of the processed GFP_{172}-IntF2A fragment (~53 kDa) impeded its permeation through the plant cell wall.[170] However, there are reports showing that recombinant Fab antibodies with similar molecular size as the GFP_{172}-IntF2A fragment can be secreted into the culture medium of plant suspension cells.[171] Thus, the lack of detectable extracellular GFP_{172}-IntF2A fragment implies that protein mistargeting might have occurred along the secretory pathway. This hypothesis was substantiated by the observation of GFP fluorescence in the vacuole of transgenic tobacco NT1 cells under confocal fluorescence

microscopy examination (Fig. 4.15a). The vacuole targeted GFP_{172}-IntF2A fragment was partially degraded, probably due to the hydrolysis of vacuole resident proteinases, as multiple degradation bands were detected by western blot probed with anti-GFP antibody (Fig. 4.15b). Correct targeting of the upstream protein into extracellular space after processing from our IntF2A based polyprotein system indicates that the unwanted vacuole sorting effect of the F2A sequence was successfully resolved by the N-terminal cleavage of intein mutant in the IntF2A domain. This finding also agrees with our previous observation (Section 3.3) that the intein cleavage activity was insensitive to the cellular oxidative environment, since the lumen of reticulate network is more oxidative comparing to the cytoplasm.[161]

When the ER targeting signal was incorporated into the N-terminus of both upstream and downstream proteins, i.e., pND-6, cellular targeting of the processed upstream protein remained the same as that of the single signal peptide construct (pND-5). Although processed downstream RFP_{Strep} can be detected in the culture medium (Fig. 4.12a), a substantial amount of processed RFP_{Strep} was accumulated inside the vacuole (Fig. 4.12b). This vacuole accumulation indicates that the processed RFP_{Strep} protein successfully passed through the ER membrane as directed by the secretory signal peptide before encountering the vacuole sorting receptors located late in the secretory pathway. Previous study also demonstrated that residual Pro resulted from F2A cleavage would not interfere with the function of the secretory signal peptide of the downstream protein.[57] Upon removal of secretory signal sequence in ER lumen, downstream RFP_{Strep} protein was free of any residual amino acid residues originated from the IntF2A domain. Although originally modified from the naturally cytoplasmic protein, the same RFP protein was able to be exclusively accumulated in the apoplastic space of *N. benthamiana* leaf when an N-terminal

secretory signal peptide was incorporated.[56] Therefore, the only exogenous sequence, the C-terminal Strep tag, is suspected to be a putative vacuole sorting determinant as many nonstructured C-terminal linker sequences were reported to be sufficient for vacuole targeting in plants.[172] When protein processing and secretion of the ND-1, ND-5 and ND-6 cells were analyzed using western blots, processed upstream GFP_{172} in both intracellular and extracellular portions showed expected molecular size (Fig. 4.12c). However, the released secretory RFP_{Strep} protein exhibited a smaller molecular size in both intracellular and extracellular portions on the western blot probed with anti-RFP antibody (Fig. 4.12d). This smaller band cannot be detected using anti-Strep tag antibody, which indicates partial or complete degradation of the Strep tag sequence in tobacco secretory pathway. This degradation was thought to take place within the vacuole trafficking pathway as the Strep tag sequence could be detected after wortmannin treatment for 24 hours that blocks the vacuole targeting.[163,173] Detection of the secreted RFP in extracellular fluid suggests that the Strep tag sequence was removed from a portion of processed RFP_{Strep} protein before reaching the vacuole sorting receptor.

In the IntF2A based polyprotein expression system reported here, the intervening IntF2A domain is found to be susceptible to cellular degradation after being liberated from the polyprotein precursor. This is evidenced from the lack of detection of the IntF2A fragment by the anti-2A antibody. Because the 2A antibody recognizes the C-terminal conserved sequence in 2A, the lack of detection of the IntF2A fragment may be due in part to partial digestion of the 2A C-terminal residues. With the pN(-)-1 extract we could detect a faint band corresponding to the GFP_{172}-IntF2A fragment by the anti-2A antibody (Fig. 4.16), although the amount is also very low. At this point we cannot rule out that inability to detect IntF2A was due completely to the

low presence of the fragment, since it is possible that partial proteolytic digestion of the 2A peptide did occur, at least to a good portion of the released fragments. More studies are needed to unequivocally determine the fate of the released IntF2A fragment.

In summary, the IntF2A based polyprotein system enables simultaneous co-expression of multiple proteins in plant cells and whole plants. While only up to three proteins were tested in this study, co-expression of more proteins can be conceivable by using this IntF2A based polyprotein system since no significant yield reduction was observed as the number of constituent proteins increased from two to three (Fig. 4.8). Thus, engineering of more sophisticated cellular metabolic pathway involving multiple enzymes or transcription factors should greatly benefit from using the IntF2A based polyprotein system. Moreover, precise subcellular targeting of constituent proteins can be achieved by incorporating proper targeting signals. These beneficial properties make the IntF2A based polyprotein system a highly useful molecular tool for plant sciences and biotechnology.

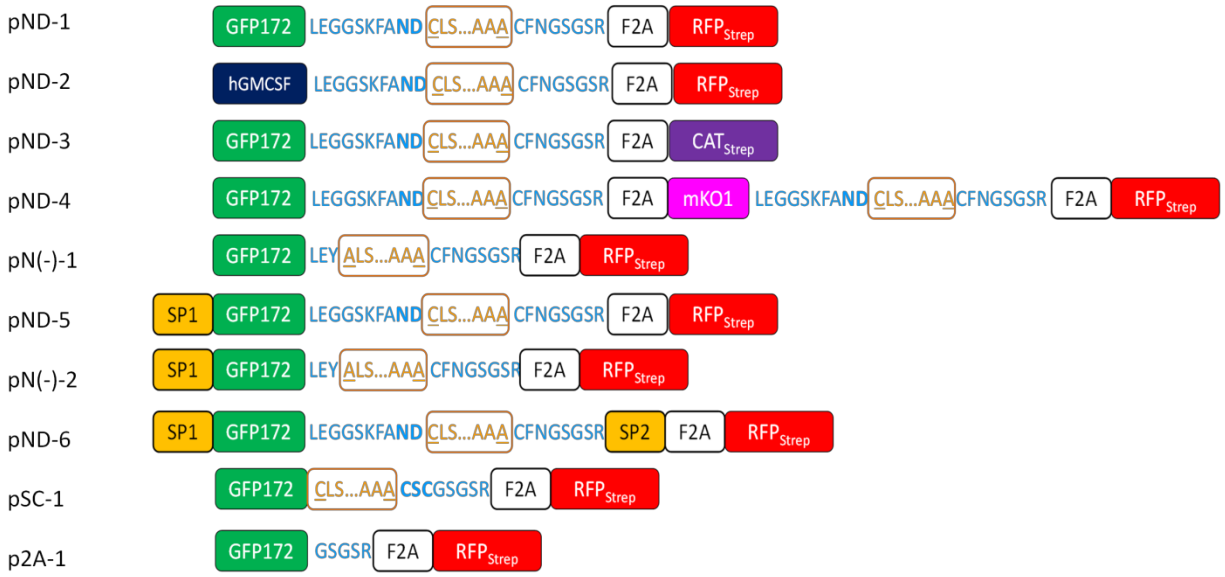


Figure 4.1. Organization of the IntF2A based polyprotein constructs used in this study. The intein domain is framed within the orange box. Residues within intein mutants that are critical for N-terminal cleavage activity are underlined. Linker sequences are labeled in blue wherein N-terminal cleavage accelerating bordering extein residues are in bold letters. F2A: 2A peptide from foot-and-mouth disease virus; SP1: *A. thaliana* basic chitinase signal peptide; SP2: rice α-amylase signal peptide.

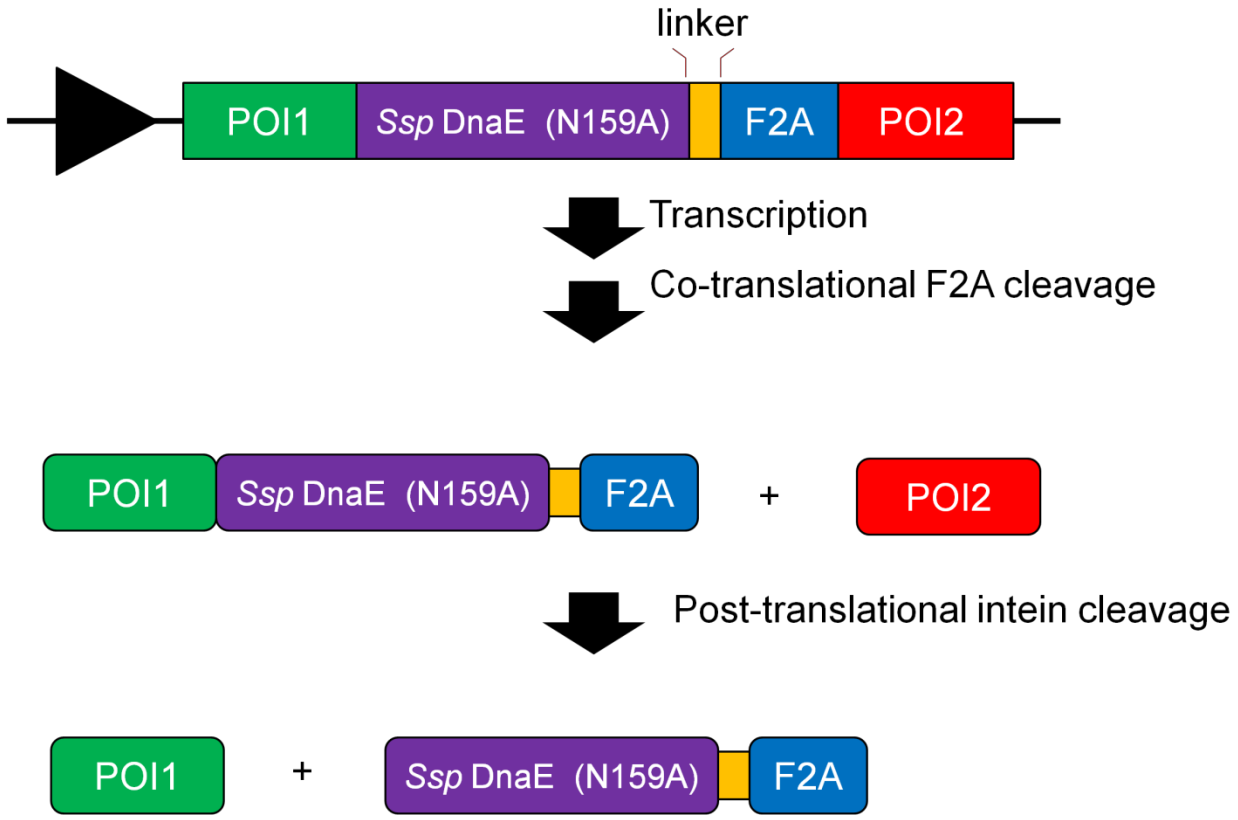


Figure 4.2. Organization and cellular processing of the IntF2A based polyprotein. POI: protein of interest

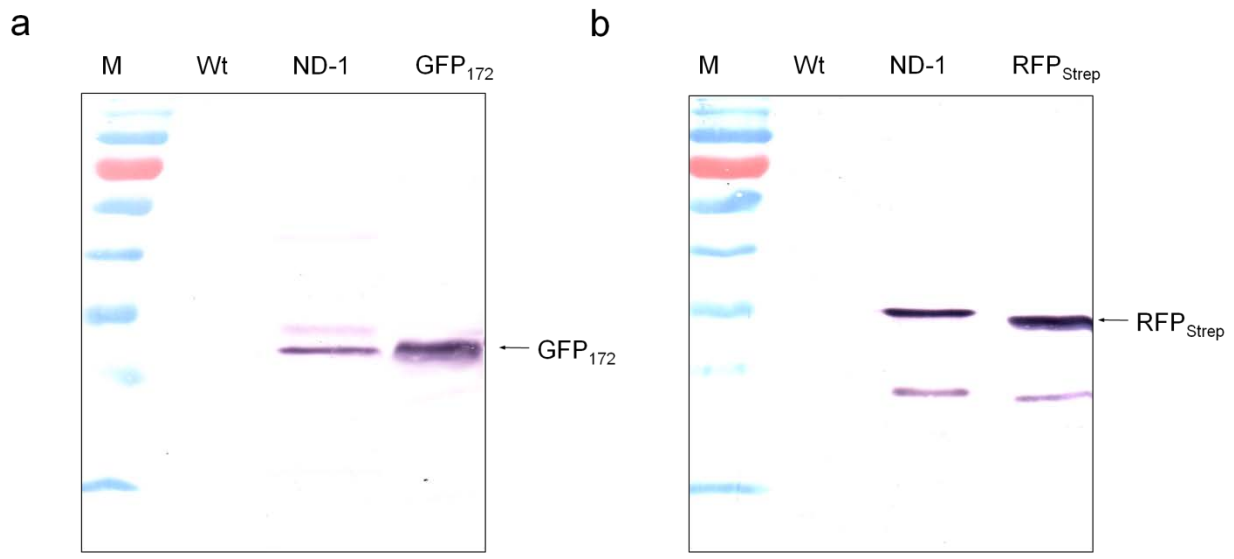


Figure 4.3. *In vivo* processing of the IntF2A based polyprotein expressing two fluorescent proteins from pND-1 in tobacco NT1 cells. Total soluble proteins were extracted and analyzed using western blots probed with anti-GFP (a) and anti-Strep tag (b) antibodies for detection of released upstream and downstream POIs, respectively. Hereinafter, “M” & “WT” denote molecular marker and non-transformed wild-type control, respectively.

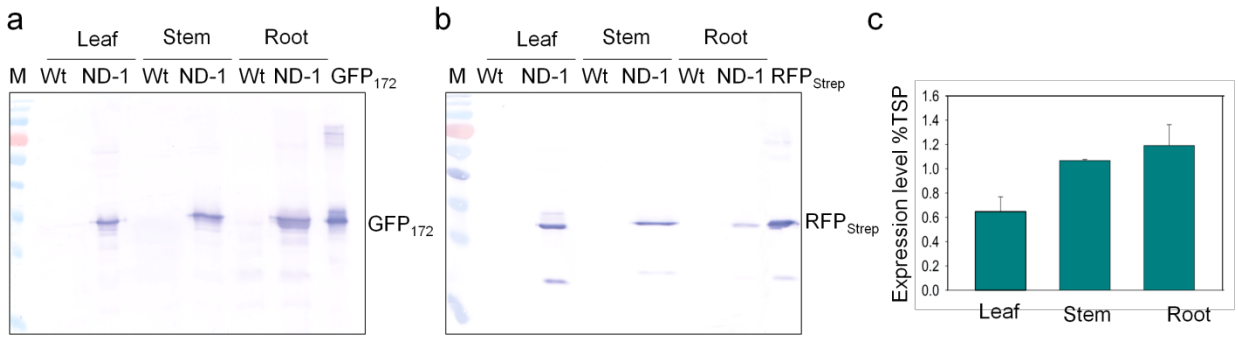


Figure 4.4. *In vivo* processing of IntF2A based polyprotein (ND-1) in transgenic *N. tabacum* cv. Xanthi plants. Total soluble proteins were extracted from leaf, stem and root tissues of transgenic tobacco plants. Polyprotein processing was analyzed using western blots probed with anti-GFP (a) and anti-Strep tag (b) antibodies for detection of released upstream and downstream POIs, respectively. Protein expression in different tissues was measured based on the percentage of processed GFP_{172} in total soluble proteins (c).

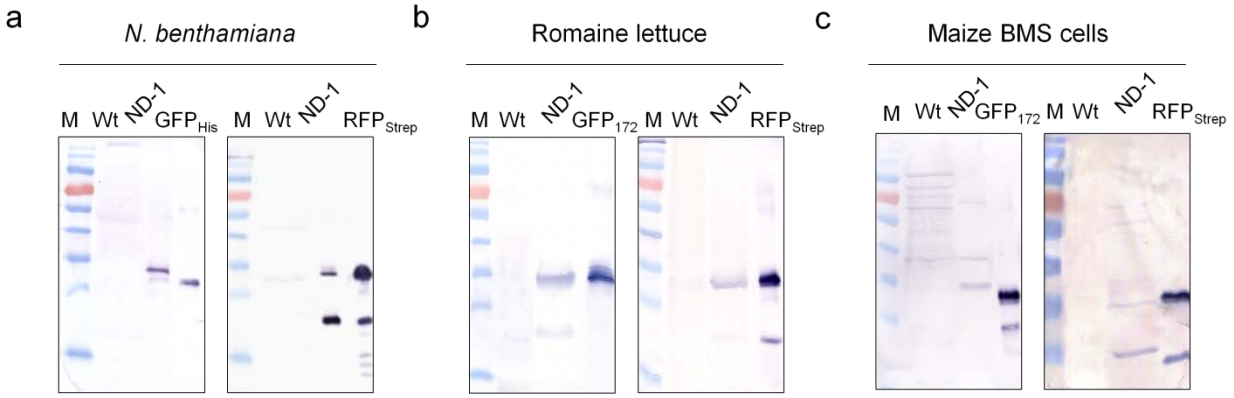


Figure 4.5. *In vivo* processing of IntF2A based polyprotein in different plant species. ND-1 was transiently expressed in leaf tissue of *N. benthamiana* (a) and Romaine lettuce (b), as well as in maize BMS cells (c) via agroinfiltration. Polyprotein processing was analyzed using western blots of total soluble protein extracts. Anti-GFP (left) and anti-Strep tag (right) antibodies were used for detection of released upstream and downstream proteins, respectively.

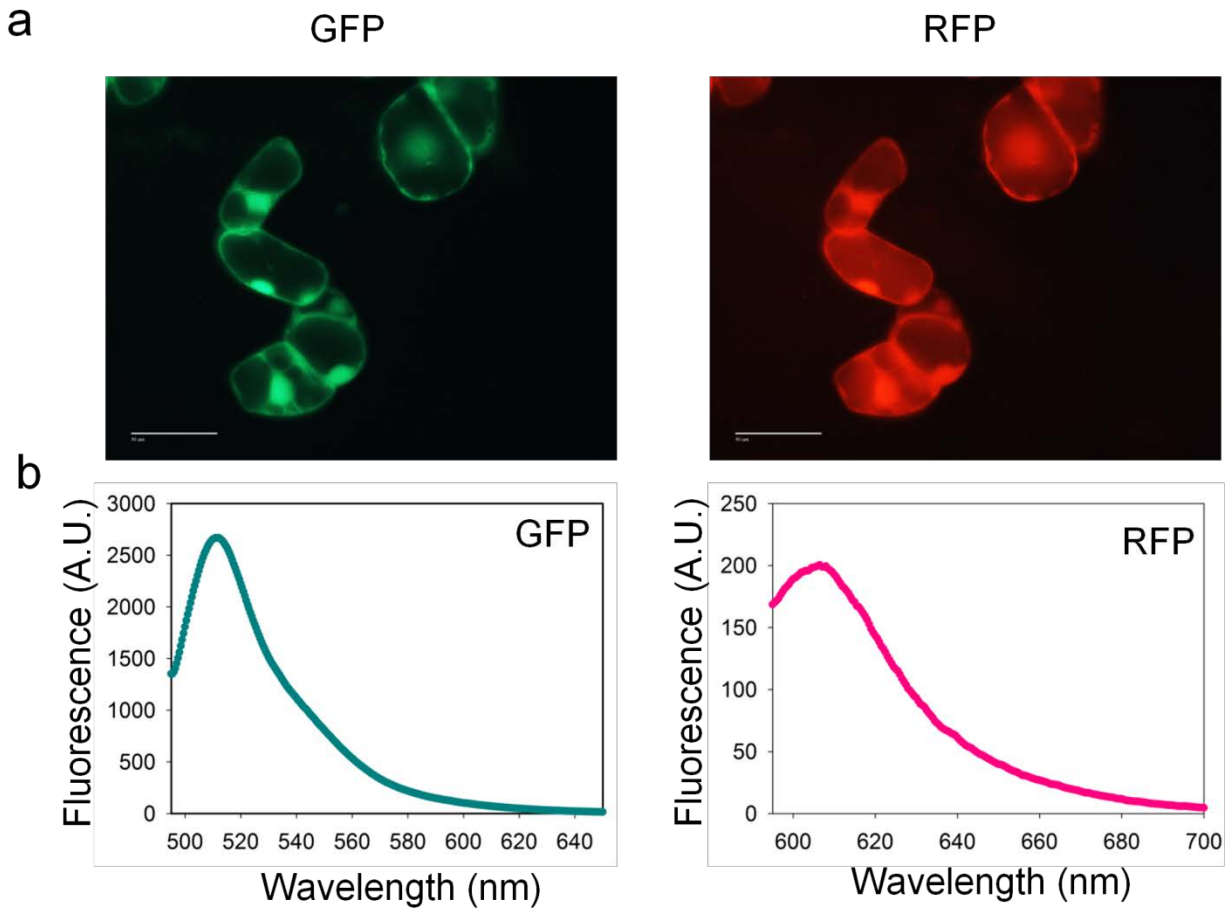


Figure 4.6. Proper fluorescence property was detected from POIs processed from the ND-1 polyprotein precursor. (a) Tobacco NT1 cells expressing pND-1 display both green (right) and red (left) fluorescence. Scale bar: 50 µm. (b) Characteristic GFP (left) and mCherry (right) fluorescence spectra were detected when cell extracts were excited with GFP and mCherry excitation wavelengths at 470 and 570 nm, respectively.

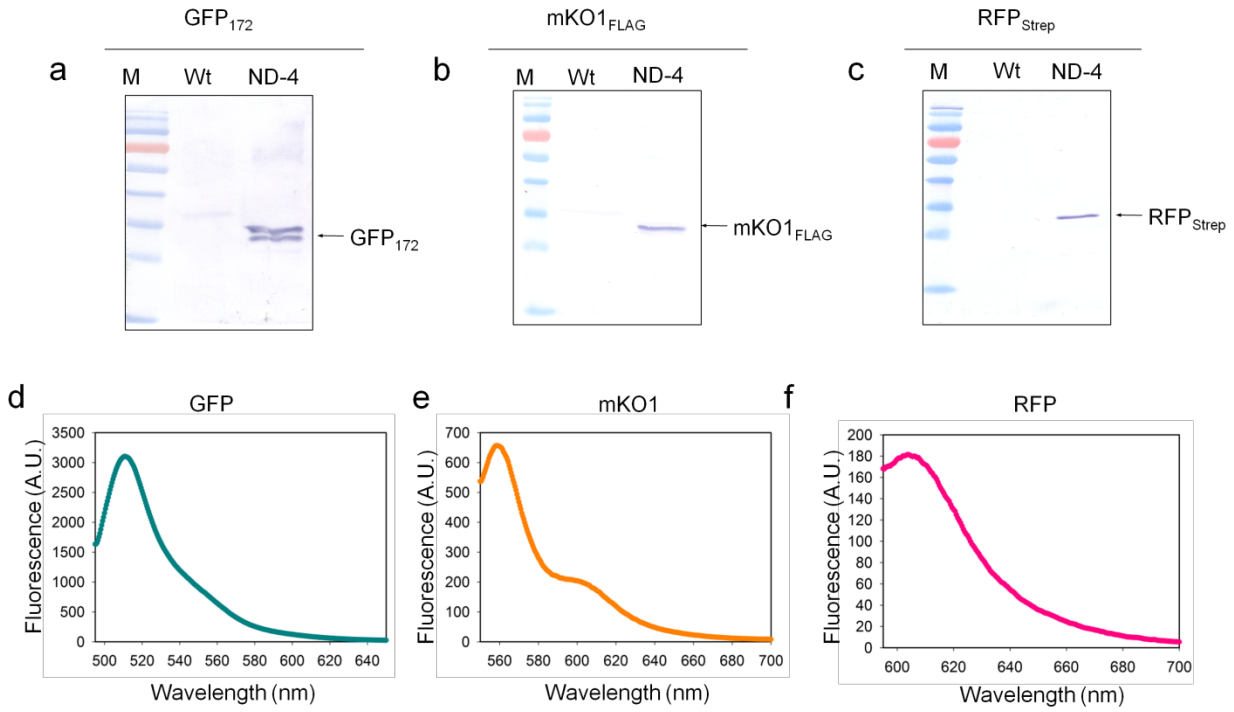


Figure 4.7. Co-expression of three POIs using the IntF2A based polyprotein. Efficient *in vivo* processing of three POIs was confirmed by western blots of transgenic NT1 cell extracts expressing pND-4 using anti-GFP (a), anti-mKO1 (b) and anti-Strep tag (c) antibodies, respectively. Processed POIs show characteristic fluorescence spectra (d-f).

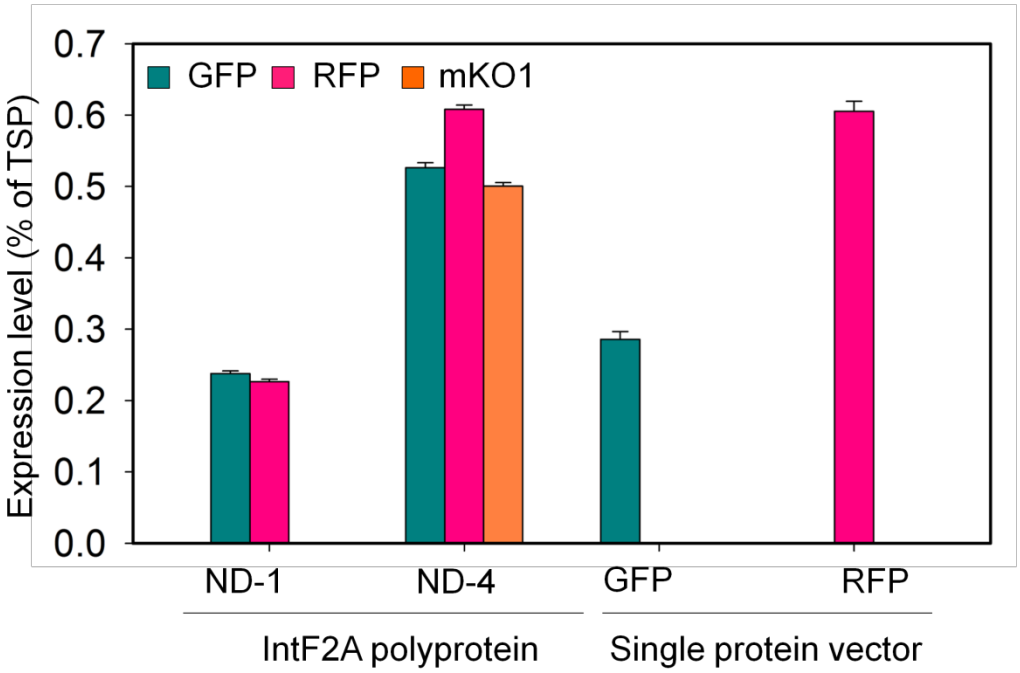


Figure 4.8. Expression level of POIs (GFP & RFP) processed from the IntF2A based polyprotein compared to that from the single protein vector. Amount of processed POIs was measured based on the fluorescence calibration curves established by purified protein standards. Expression level was calculated based on the percentage of the constituent proteins in total soluble proteins (TSP) of transgenic tobacco NT1 cell extracts.

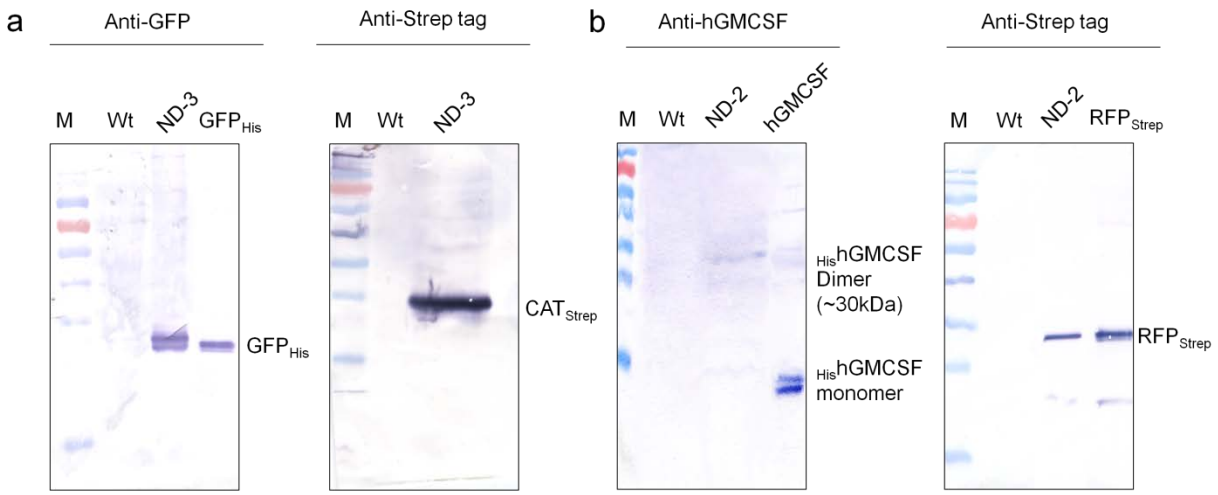


Figure 4.9. Application of IntF2A based polyprotein for co-expressing multiple proteins other than fluorescent reporters in plants. (a) *In vivo* processing of IntF2A based polyprotein expressing an upstream GFP_{His} and a downstream CAT_{Strep} (ND-3) based on western blot analysis. (b) *In vivo* processing of IntF2A based polyprotein expressing an upstream $_{His}$hGMCSF and a downstream RFP_{Strep} (ND-2) based on western blot analysis. Antibodies against GFP, Strep tag and hGMCSF were used to detect processed GFP_{His}, RFP_{Strep}/CAT_{Strep} and $_{His}$hGMCSF, respectively.

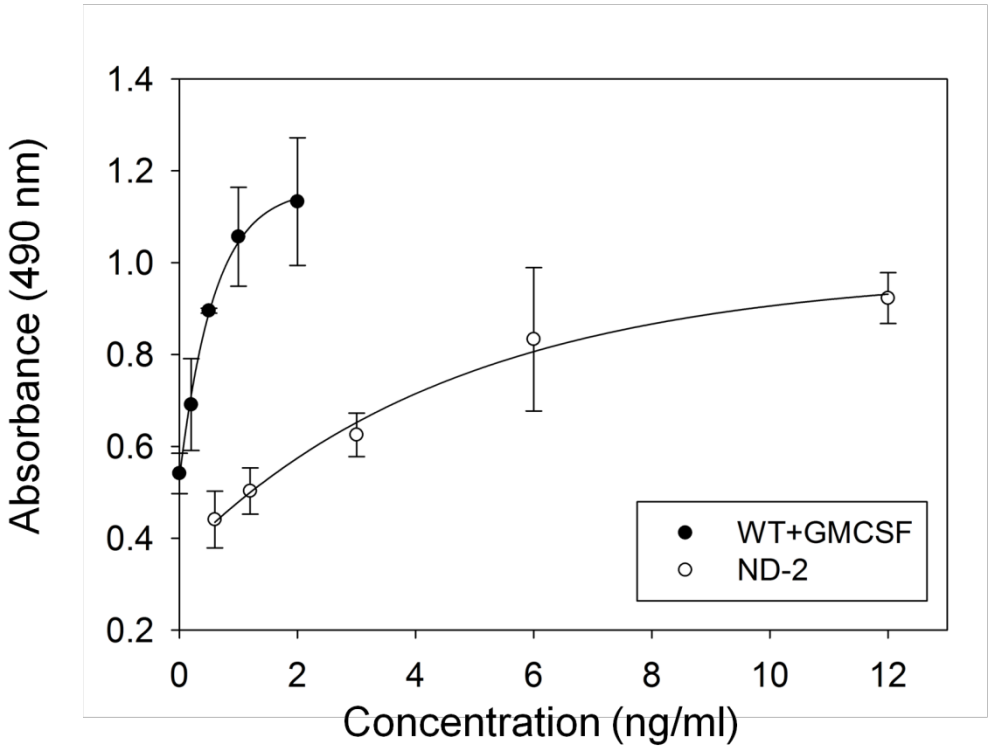


Figure 4.10. Bioactivity of $_{His}$hGMCSF processed from the IntF2A based polyprotein precursor. GMCSF dependent human TF1 cells were cultured with different amounts of protein extracts of transgenic tobacco NT1 cells expressing ND-2. The cellular amount of $_{His}$hGMCSF was estimated based on fluorescence measurement of its co-expression partner, RFP_{Strep}. To minimize any effect from plant extract on $_{His}$hGMCSF activity, protein extract was partially purified by ammonium sulfate precipitation and desalted before adding to the cell culture. Increased of absorbance at the wavelength of 490 nm indicates the increase of cell proliferation.

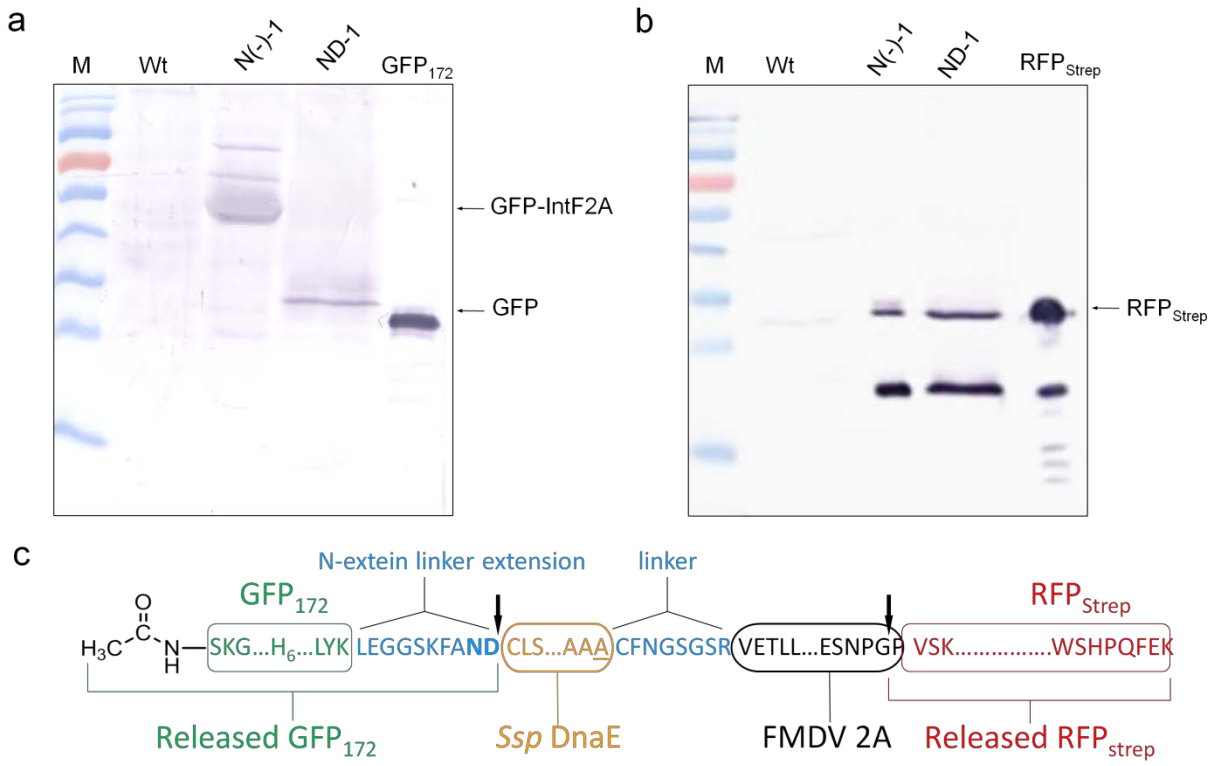


Figure 4.11. Analysis of cellular cleavage sites within the IntF2A polyprotein precursor ND-1. (a) N-terminal cleavage inactivated intein mutant abolished processing of the upstream POI. (b) Inactive N-terminal intein cleavage does not impair C-terminal cleavage of F2A in the IntF2A hybrid autoprocessing domain. (c) *In vivo* cleavage of the ND-1 polyprotein precursor occurred at the expected sites, as confirmed using N-terminal amino acid sequencing and ESI-TOFMS analysis of released POIs.

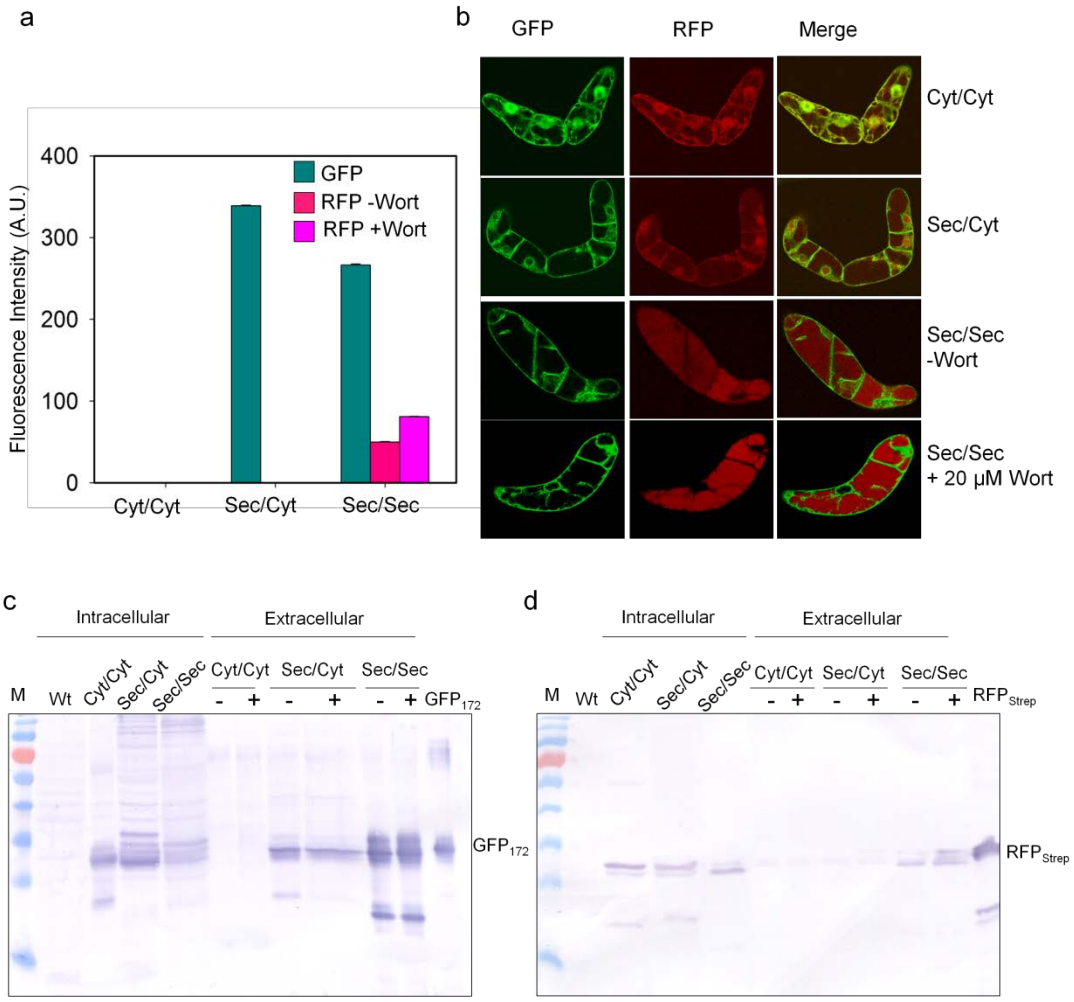


Figure 4.12. Subcellular targeting of constituent POIs using the IntF2A based polyprotein expression system in plants. (a) Secreted GFP and RFP proteins were detected in concentrated apoplastic fluid of tobacco NT1 cells when the N-terminal ER targeting signal peptide was incorporated. (b) Confocal fluorescence microscopy confirmed the proper cellular targeting of processed POIs inside tobacco NT1 cells. Cellular extracts and concentrated spent media of transgenic tobacco NT1 suspension cells with (+) or without (-) wortmannin treatment were analyzed by western blots. (c) Processed upstream protein was detected by anti-GFP antibody. (d) Processed downstream protein was detected by anti-mCherry antibody.

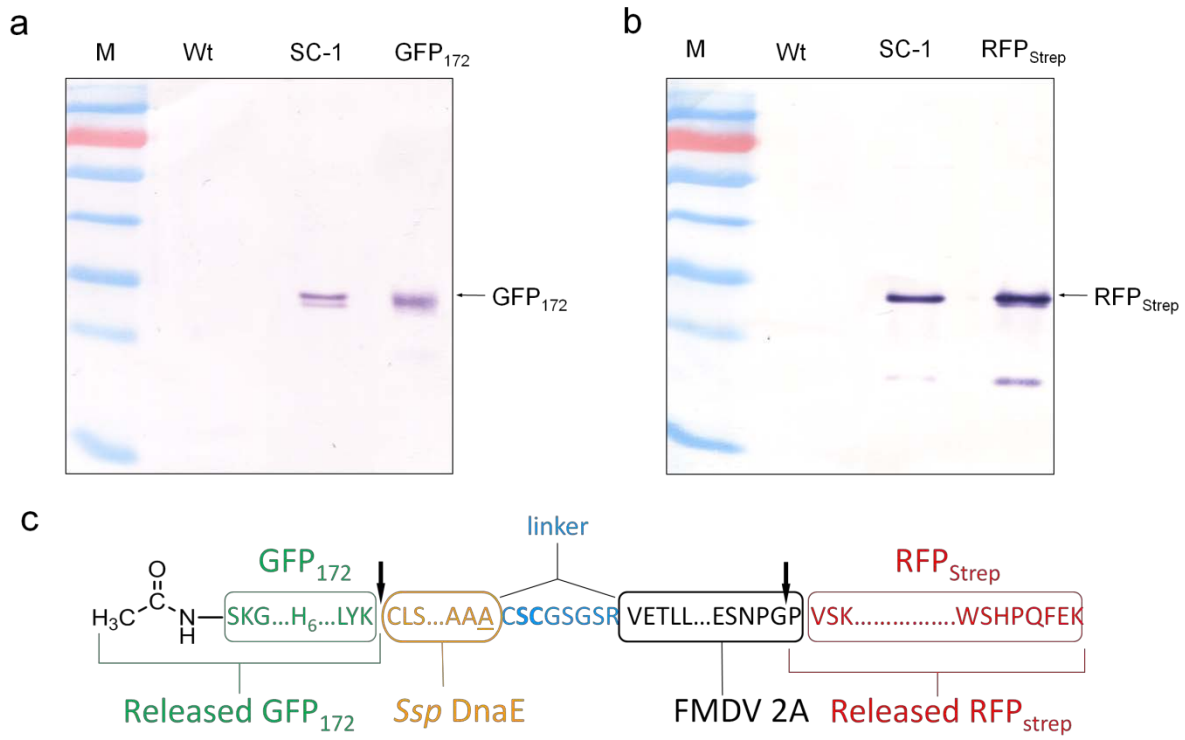


Figure 4.13. Incorporation of CSC flanking C-terminus of the intein domain is sufficient to promote efficient released of upstream POI without adding additional residues to its C-terminus. (a-b) Processing of SC-1 polyprotein without N-extein linker extension (cf. ND-1) in NT1 cells analyzed using western blots. Processed upstream GFP_{172} and downstream RFP_{Strep} proteins were detected by anti-GFP (a) and anti-Strep tag (b) antibodies, respectively. (c) N-terminal amino acid sequencing and ESI-TOFMS analysis confirmed that *in vivo* cleavage of the IntF2A polyprotein without N-extein linker extension occurred at the expected sites.

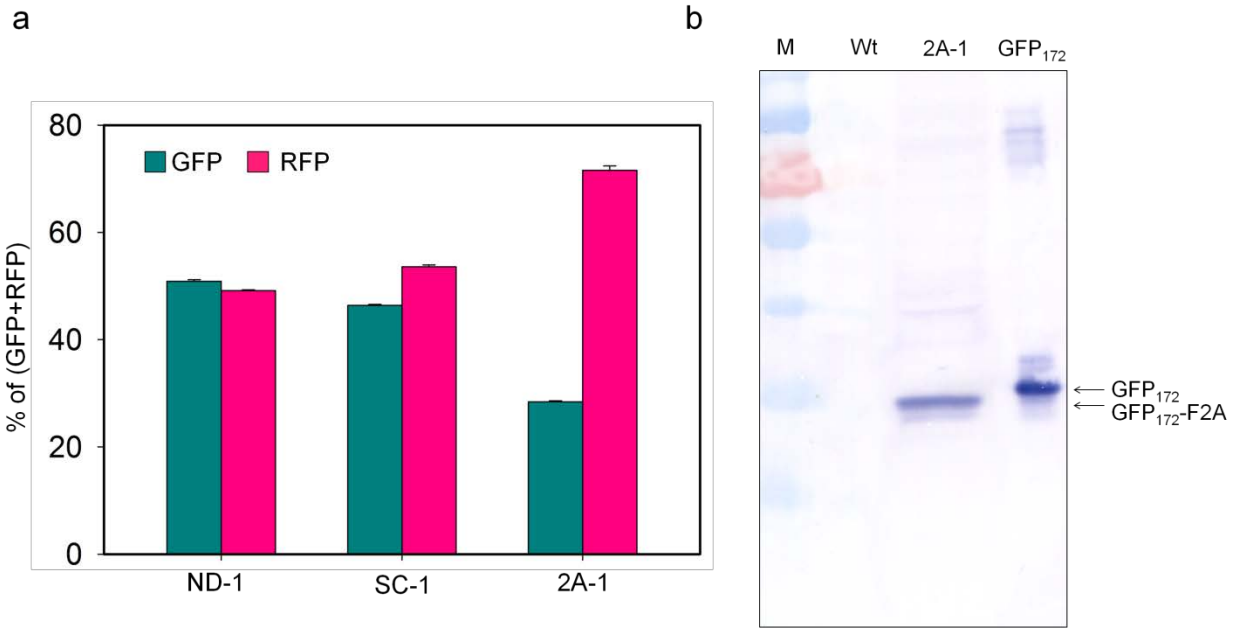


Figure 4.14. Proteins with F2A sequence at their C-terminus may be unstable. (a) Uneven accumulation of POIs processed from the polyprotein by F2A sequence alone (without the intein domain) was noted, while balanced co-expression of POIs was noted from IntF2A polyproteins. (b) The GFP_{172}-F2A fragment released from the 2A-1 polyprotein appeared smaller than GFP_{172} on the western blot probed with anti-GFP antibody, suggesting potential proteolytic digestion prompted by the F2A overhang.

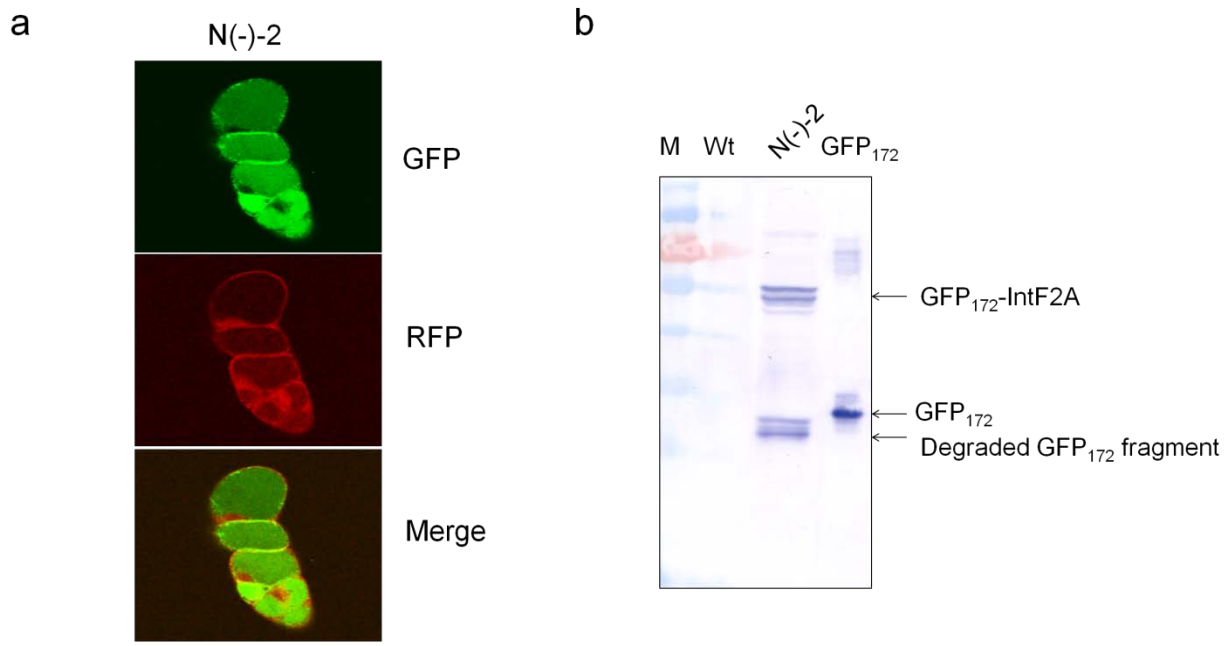


Figure 4.15. C-terminal F2A extension causes mistargeting of POI to vacuole when a ER targeting signal is included on the N-terminus of the POI. (a) Confocal fluorescence microscopy images of NT1 cells expressing pN(-)-2. (b) Partial degradation of the GFP_{172}-Int(N-)F2A fragment was detected using western blot probed with anti-GFP antibody.

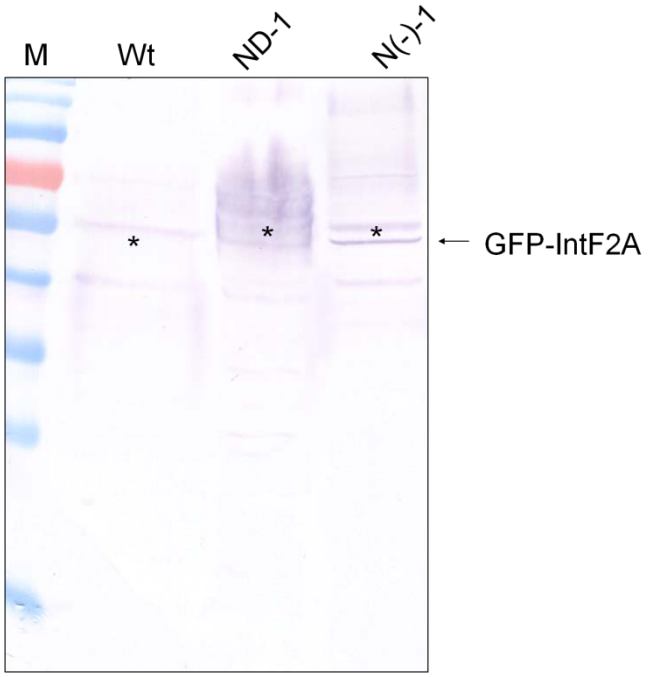


Figure 4.16. Determination of cellular destination of released IntF2A fragment by western blot analysis. Anti-2A antibody was used to detect residual F2A fragment in the total soluble proteins of NT1 cells expressing ND-1 and N(-)-1, respectively. * indicates non-specific bands.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS FOR FUTURE STUDIES

Simultaneous stacking of multiple transgenes in plants lays the foundation of abundant applications in both commercial practice and basic research that require coordinate action of multiple proteins.[5] This study focused on the development and characterization of three novel polyprotein expression systems for coordinated co-expressing multiple proteins in plants, to overcome the limitations of traditional gene stacking approaches. All three systems are capable to correctly release constituent proteins of interests (POIs) from the polyprotein precursor in transgenic tobacco. Moreover, these systems are found to be applicable to additional plant species including *N. benthamiana* and romaine lettuce, based on transient expression results. The proteins of interest expressed using the polyprotein systems exhibited proper functions and accumulated to nearly stoichiometric amounts inside the cells.

The first polyprotein system developed in this dissertation research utilizes endogenous kex2p protease like activity found in a broad range of plant species[68-69] for facilitating co-expression of multiple secretory proteins connected by a putative kex2p substrate sequence. Similar to the yeast and mammalian analogs, the plant kex2p like proteases are found in the *trans*-Golgi network.[83] Processing of the polyprotein precursor was also confirmed to occur beyond *cis*-Golgi in secretory pathway based on our result of Brefeldin A treatment.[174] Thus, this kex2p based polyprotein system provides an ideal tool for production of secretory proteins such

as pathogen defensins and cell wall hydrolases. Furthermore, targeting recombinant proteins into culture medium can simplify the protein purification process. However, this system is not suited for producing non-secretory proteins, or proteins that require authentic terminal sequences to function.

The limitation of the kex2p based polyprotein system prompted us to develop a more universally applicable system, in which a novel dual-intein fusion domain with high autocatalytic cleavage activity was applied to enable efficient release of the constituent proteins from the polyprotein precursor.[128,131] As the intein cleavage activity is solely dependent on the nucleophilic displacement of the amino acid residues at the intein terminal junctions,[142] this dual-intein fusion domain can in principle mediate cleavage of polyproteins in all cell types. Our results had successfully demonstrated the application of this dual-intein based polyprotein system for mediating co-expression of multiple proteins in the cytosol of bacterial, mammalian and plant systems. When expressed in *E. coli* at 37 ºC, low expression level of proteins processed from the dual-intein based polyprotein was noted, which was probably due to the poor folding of the dual-intein domain at rapid expression condition. Reducing induction temperature (i.e., 25 ºC) facilitates the protein folding that leads to the improvement of protein expression level. However, lower expression temperature in *E. coli* compromises the C-terminal intein cleavage efficiency by reducing the Asn cyclization rate.[142-143] In plants, the DI-downstream protein fragment resulted from the attenuated C-terminal intein cleavage might be prone to degradation inside the cells, which directly reduces the accumulation of downstream protein. The predicament of expression level and C-terminal intein cleavage efficiency can be resolved by incubating the *E. coli* cells at 37 ºC after overnight protein expression at lower temperature. However, plants that

are preferably grown at room temperature might not tolerate this high temperature treatment. Therefore, improvement of C-terminal cleavage efficiency is necessary to allow more effective application of this intein based polyprotein expression system in plants.

In the third approach, the downstream intein in the dual-intein autoprocessing domain was replaced with the FMDV 2A (F2A) sequence,[42] aimed to promote more efficient release of the protein of interest located downstream of the autoprocessing domain. With this design, not only the release of the downstream POI from the IntF2A autoprocessing domain was improved, potential adverse effect of residual 2A overhang on the upstream POI was significantly reduced by highly effective N-terminal intein self-cleavage.[54-57] As the C-terminal cleavage activity of F2A occurs in ribosome at the translational stage,[51] cellular targeting of the constituent POIs would not be affected using this IntF2A based polyprotein system. Our results successfully demonstrated differential cellular targeting of processed POIs from the IntF2A based polyprotein precursor when different signal peptides were incorporated. Interestingly, processed downstream POI (RFP) intended for secretion was eventually accumulated in the vacuole when a Strep tag sequence was appended to its C-terminus. Because correctly targeting of the same RFP protein (but without the Strap tag) for secretion was previously reported in *N. benthamiana*,[56] we suspect that the C-terminal Strep tag sequence might serve as a vacuole targeting signal in tobacco cells. Since there is no specific rules for the plant C-terminal vacuole sorting sequences,[172] particular cautions should be taken when the affinity sequence is designed for detection and purification of recombinant secretory proteins in plants for future applications.

To further improve the autoprocessing domains, smaller intein variants that fold more efficiently may be sought to replace the intein variants currently used in the dual-intein domain

to increase protein expression level in bacterial system at high induction temperature. As the ribosomal recoding activity of 2A sequence is only dependent on eukaryotic ribosomes,[48] application of IntF2A based polyprotein system can be extended to eukaryotic hosts other than plants (e.g., yeast, mammals, insects, invertebrates, etc.). In addition to co-expressing proteins as the intended products, cellular metabolism can be engineered using the polyprotein expression systems developed in this study to significantly broaden the applications of these systems.

Although the released IntF2A domain in transgenic tobacco cells was investigated using 2A western blot in this study, degradation of this autoprocessing domain is not conclusive. This is because proteolytic removal of just few residues from the C-terminal end of the 2A peptide can impede detection using the anti-2A antibody. To further study the fate of the intervening autoprocessing domain, a short epitope sequence can be inserted into the linker between the intein and the 2A peptide to facilitate western blot analysis or immuno-electron microscopy examination. Considering the potential food safety issue using the dual-intein or IntF2A based system in genetically modification of consumable crops, it is preferable to degrade the autoprocessing domain after completing protein processing function. Incorporating an internal domain or motif between N- and C-terminal autoprocessing domains to direct proteasome targeting can be considered as one strategy for removing the released autoprocessing domain inside the cells. With the better elucidation of protein degradation mechanism in plants, more degradation signals may be exploited for removing the autoprocessing domains *in vivo*.

APPENDICES

Appendix 1. Primers Used for Construct Cloning.

Name	Sequence
*Bst*BI-G172C-F	5’-CCCTTTCGAAAGATCCCAACGA-3’
CAT-2A-F1	5’-GAGACGTCGAGTCCAACCCTGGGCCCGAGAAAAAAATC ACTGGATATACCAC-3’
CAT-STREP-R	5’-CAAACTGAGGATGTGACCATCCAGAACCCGCCCCGCCCT GCCACTCATCGCAG-3’
DnaB-SRGP-R	5’-GCTCACGGGCCCGCGTGAGTTGTGTACAATGATGTCATTC GCGAC-3’
DnaB-*Xba*I-F	5’-TCTGGTTCTAGAGAGTCTGGAGCTATCTCTGGCGATAGTC TGATC-3’
DnaENAFN-*Xba*I-R	5’-AACTCTAGAACCAGAACCATTGAAAGCGTTTGCAGCAA TAGCACCGTTTGC-3’
DnaE-*Xba*I-R-2	5’-GTATTCTAGAACCACCAGCTGCAGCAATAGCACCGTTTGC GAG-3’
DnaE-*Xho*I-F-2	5’-ACTCCTCGAGTATGCATTGTCCTTCGGAACTGAGATACTTA C-3’
DsRed-R	5’-GCTCTCGAGCAGGAACAGGTGGTGGCGGCC-3’
DSRED2HDELX	5’-GCTCTCGAGCTCATCATGCAGGAACAGGTGGTGGCG-3’
DSRED2M1S	5’-GCAGCATGCCCTCCTCCGAGAACGTCATC-3’
FLAG-*Sac*I-R1	5’-TGCATGAGCTCCTTATCATCATCATCCTTATAATCTCCAGA AC-3’
G11c-*Sac*I-mChe-F	5’-GCTGCTGAGCTCGTGAGCAAGGGCGAGGAGGAT-3’
G172-DnaE-F	5’-ACTATACAAATGTTTGTCCTTCGGAACTGAG-3’

G172-DnaE-R	5’-AGGACAAACATTTGTATAGTTCATCCATGCC-3’
GFP-GS-*Xba*I-R	5’-AACTCTAGAACCAGAACCTTTGTATAGTTCATCCATGCC-3’
G172-*Kpn*I-R	5’-GTGGTACCTTTGTATAGTTCATCCATG-3’
GFP-*Kpn*I-F	5’-GACGGTACCCAAGGAGATATAACAATG-3’
GFP-R	5’-GCGTCTAGATTATTTGTATAGTTCATCCAT-3’
GFP-*Xho*I-R	5’-TGCCTCGAGGTGATGGTGATGGTGATGC-3’
GFS2	5’-GCACCGCGGGAAGTAAAGGAGAAGAACTTTTC-3’
GKZ-R	5’-CTGGAGCTCTCATTTTTCAAACTGAGGATGTGAC-3’
GMCSF-*Xho*I-R	5'-CCGCTCGAGCTCCTGGACTGGCTCCCAGCAG-3'
GRS	5’-CGCGAGCTCTTATTTGTATAGTTCATCCATGC-3’
IDnaB-*Apa*I-R	5’-ATGGGCCCACTTCCCGCGTGTACAATGATGTCATTCG-3’
IDnaE-*Kpn*I-F	5’-ACTCCTCGAGTATGCATTGTCCTTCGGAACTGAGATACTT AC-3’
*Kpn*I-DnaE-F	5’-CGGGGTACCAAGTTTGCAAATGATTGTTTGTCC-3’
LJ607R4X	5’-GCACTCGAGGCAGGAGGAATAGGCAAACGG-3’
M13F	5’-GTAAAACGACGGCCAG-3’
MBR	5’-CGCGGATCCCTTGTACAGCTCGTCCATGC-3’
mCherry-2A-F1	5’-GAGACGTCGAGTCCAACCCTGGGCCCGTGAGCAAGGGC GAGGAGGATAACATG-3’

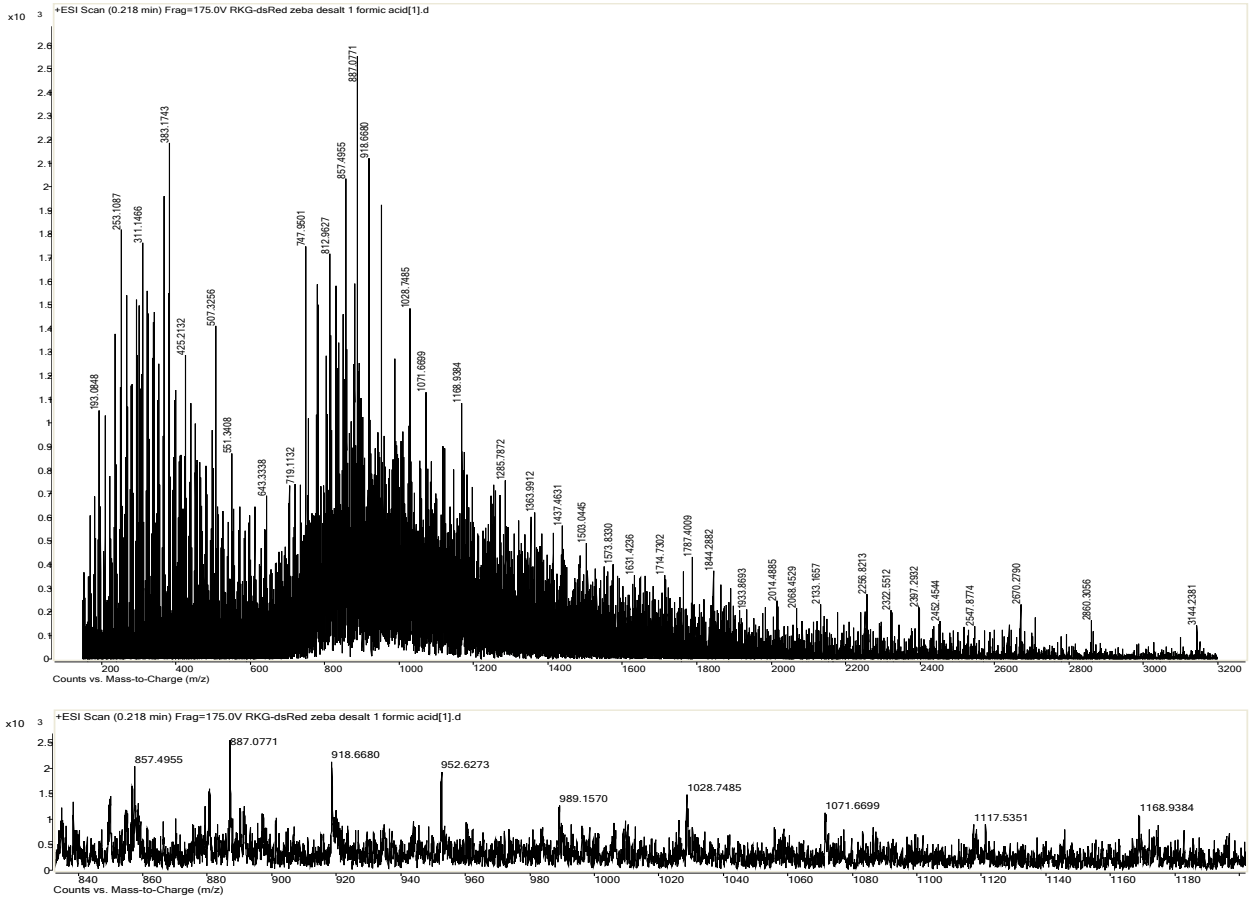
mCherry-*Sph*I-F	5’-CAGCCGGCATGCTGAGCAAGGGCGAGGAG-3’
mCherry-STREP-R	5’-CAAACTGAGGATGTGACCATCCAGAACCCTTGTACAGCTCGTCCATGCCGCCG-3’
mGFP5-His-*Xho*I-R	5’-AGTACTCGAGGTGATGGTGATGGTGATGACTTCCAATCCCAGCAGCTGTTACAAACTC-3’
mGFPL242AE	5’-GCGGAATTCCCTAGGTTAAAGCTCATCATGTTTGTATAG-3’
mGFPS2E	5’-GCCGAATTCAGTAAAGGAGAAGAACTTTTC-3’
mKO1-*Apa*I-F	5’-GCATGGGCCCGTGAGTGTGATTAAACCAGAGATGAAGA-3’
mKO1-R	5’-CATCATCCTTATAATCTCCAGAACCGGAATGAGCTACTGCATCTTCTACCAGC-3’
Nd172F	5’-ACCCAGATCATATGAAGCGGCAC-3’
Nd172R	5’-GTGCCGCTTCATATGATCTGGGT-3’
*Nde*I-G172-F	5’-GGAATTCCATATGAGTAAAGGAGAAGAACTTTT-3’
OsAASP-F	5’-GCAGTCGACTGTGTCTGCACCATGCAGGTG-3’
*Sac*I-*Nhe*-DI-F3	5’-TATATGAGCTCTCTGGCTAACTAGAGAACCCACTGCTTACTGGCTTATCGAAA-3’
*Sal*I-GMCSF-F	5'-CGCGTCGACGCCACCATGGCGACTAATCTT-3'
*Sal*I-His-GMCSF-F	5’-GCGTCGACCAAGGAGATATAACAATGCATCACCATCACCATCACGCACCCGCCCGCTCGC-3’
*Sal*I-mGFP5	5’-ACGCGTCGACCAAGGAGATATAACAATGAGTAAAGGAGAAGAACT-3’
Strep-*Not*I-R	5’-TTTTCTTTTGCGGCCGCTCATTTTTCAAACTGAGGATG-3’
XBF	5’-CCCTCTAGAAATAATTTTGTTTAAC-3’

*Xba*I-SC-R	5’-AGTAACTCTAGAACCAGAACCACAAGAACAAGCTGCAGC AATAGCACCGTTTGCG-3’
*Xho*I-Int-F	5’-CATCACCTCGAGGGAGGATCTAAGTTTGC-3’

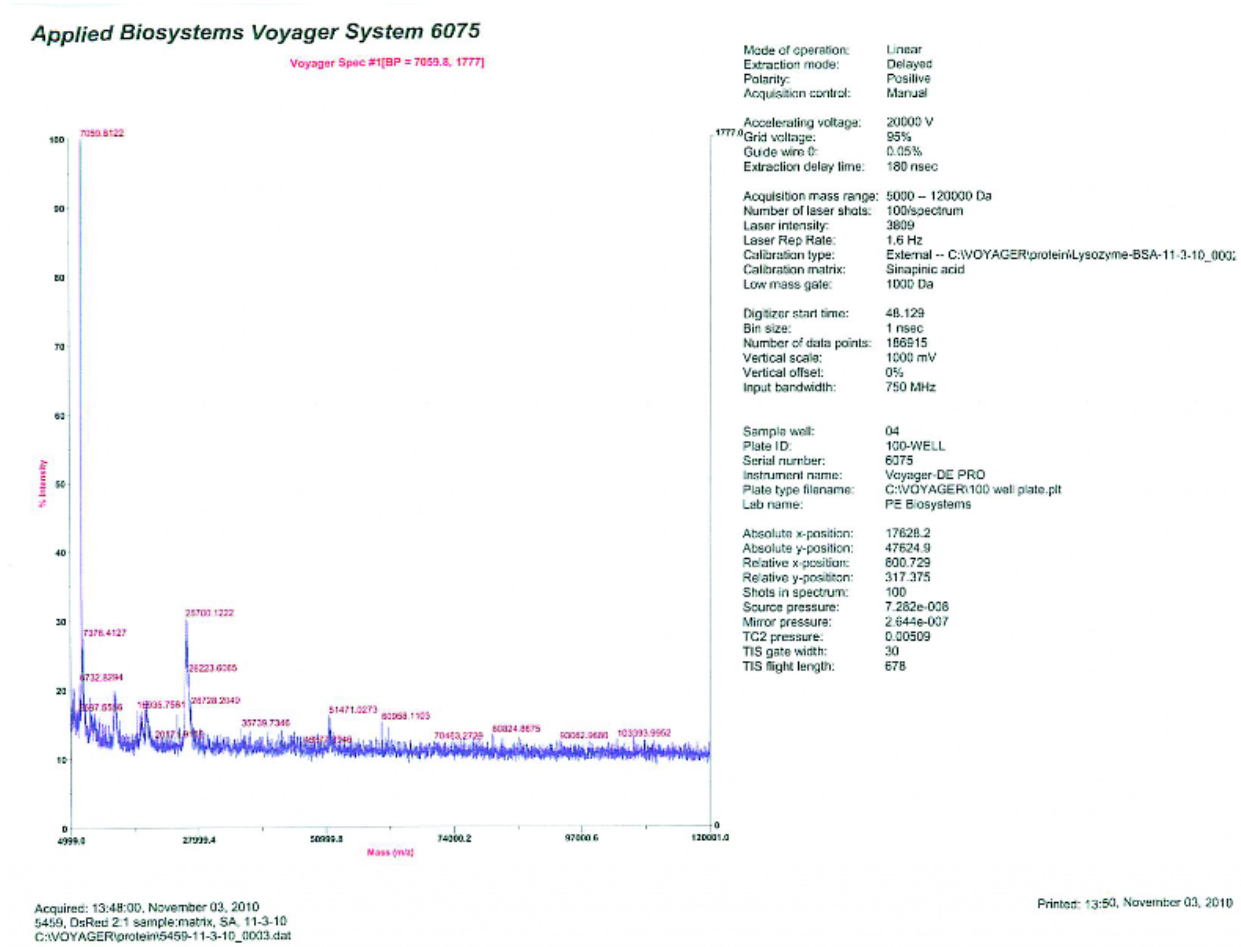
Appendix 2. Molecular analysis of DsRed and GFP products released from the RKG polyprotein precursor using MS analysis and N-terminal sequencing. SP1: rice α-amylase signal peptide; SP2: *Arabidopsis* basic chitinase signal peptide. The triangles indicate the amino-terminal residues as determined using Edman degradation, the kex2p linker sequence is boxed, and the protein domains of DsRed and GFP are double underlined; cf. Figure 2.1.

Construct	Protein processing
pE1226-RKG	MQ····AGMPSSENVIT·····HLFLLEAGG(IGKRGK)$_3$EFSKGEELFTGVVPILV····· SP1 released DsRed (calc. mass = 25,699 Da) released & secreted GFP
pBIN-GFP	MA····SAEFSKGEELFTGVVPILV····· SP2 GFP (secreted)

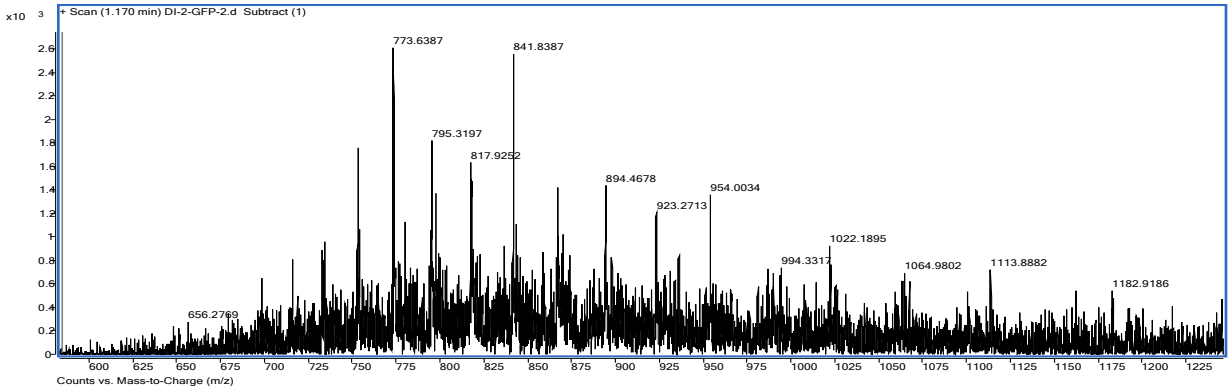
Appendix 3. ESI-TOF mass spectrum of purified DsRed released from RKG calli (Chapter 2).

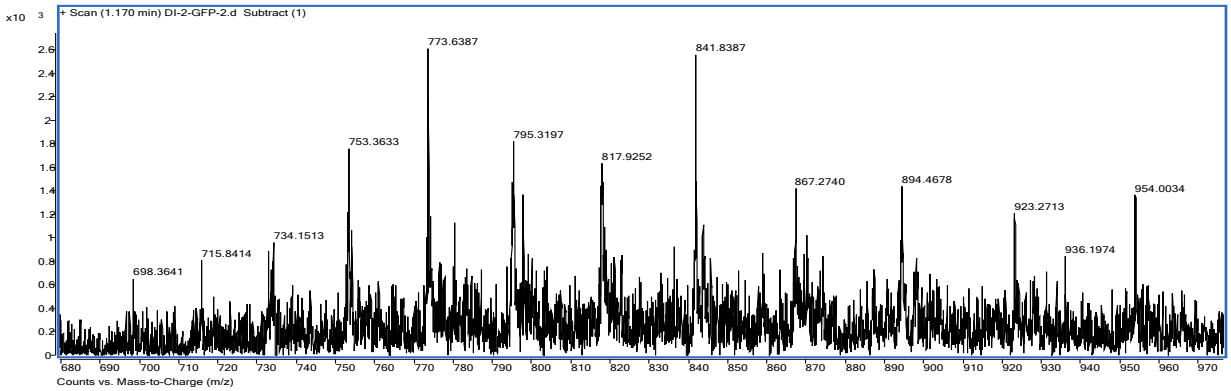


Appendix 4. MALDI-TOF mass spectrum of purified DsRed released from RKG calli (Chapter 2).

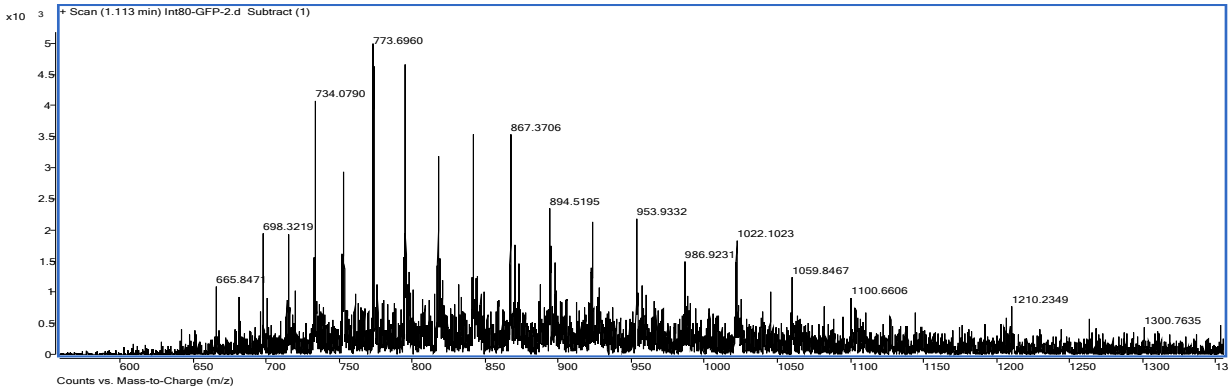


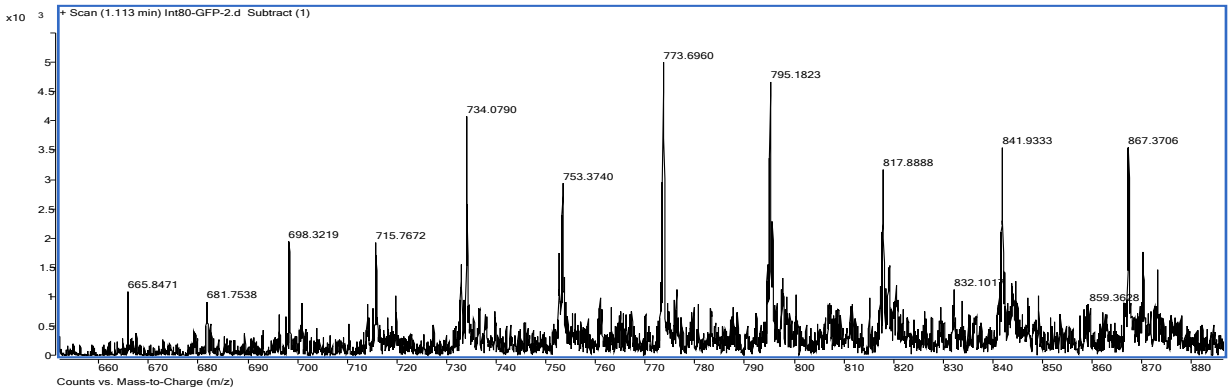
Appendix 5. ESI-TOF mass spectrum of purified GFP_{172} released from DI calli (Chapter 3).



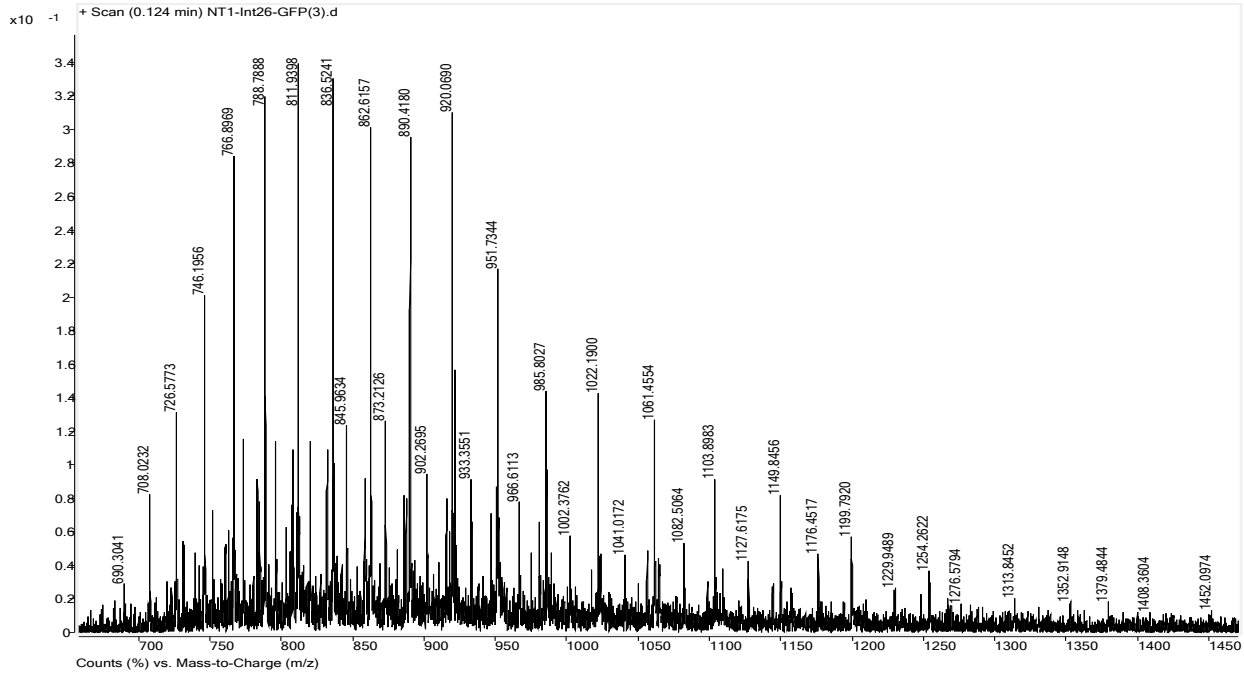


Appendix 6. ESI-TOF mass spectrum of purified GFP_{172} released from ND-1 calli (Chapter 4).





Appendix 7. ESI-TOF mass spectrum of purified GFP_{172} released from SC-1 calli (Chapter 4).



REFERENCES AND NOTES

(1) James, C. *ISAAA in Brief* **2012**, *44*.

(2) Barfoot, G. B. a. P. *AgBioForum* **2005**, *8*, 10.

(3) Moar, W. J.; Anilkumar, K. J. *Science* **2007**, *318*, 1561-1562.

(4) François, I. E. J. A.; Broekaert, W. F.; Cammue, B. P. A. *Plant Sci.* **2002**, *163*, 281-295.

(5) Halpin, C. *Plant Biotechnol. J.* **2005**, *3*, 141-155.

(6) Guo, J.; Dai, X.; Xu, W.; Ma, M. *Chemosphere* **2008**, *72*, 1020-1026.

(7) Guo, J.; Xu, W.; Ma, M. *J. Hazard. Mater.* **2012**, *199–200*, 309-313.

(8) Chan, Y.-L.; Lin, K.-H.; Sanjaya; Liao, L.-J.; Chen, W.-H.; Chan, M.-T. *Transgenic Res.* **2005**, *14*, 279-288.

(9) Senthilkumar, R.; Cheng, C.-P.; Yeh, K.-W. *Plant Biotechnol. J.* **2010**, *8*, 65-75.

(10) Rivero, M.; Furman, N.; Mencacci, N.; Picca, P.; Toum, L.; Lentz, E.; Bravo-Almonacid, F.; Mentaberry, A. *J. Biotechnol.* **2012**, *157*, 334-343.

(11) Roderick, H.; Tripathi, L.; Babirye, A.; Wang, D.; Tripathi, J.; Urwin, P. E.; Atkinson, H. J. *Mol. Plant Pathol.* **2012**, *13*, 842-851.

(12) Zhu, S.; Duwal, A.; Su, Q.; Vossen, J.; Visser, R. F.; Jacobsen, E. *Transgenic Res.* **2013**, *22*, 315-325.

(13) Khan, R.; Darwish, N.; Khattak, B.; Ntui, V.; Kong, K.; Shimomae, K.; Nakamura, I.; Mii, M. *Mol. Biotechnol.* **2014**, *56*, 814-823.

(14) Walawage, S.; Britton, M.; Leslie, C.; Uratsu, S.; Li, Y.; Dandekar, A. *BMC Genomics* **2013**, *14*, 668.

(15) Li, X.; van Loo, E. N.; Gruber, J.; Fan, J.; Guan, R.; Frentzen, M.; Stymne, S.; Zhu, L.-H. *Plant Biotechnol. J.* **2012**, *10*, 862-870.

(16) Dzivenu, O. K.; Park, H. H.; Wu, H. *Protein Expr. Purif.* **2004**, *38*, 1-8.

(17) Cingel, A.; Savić, J.; Ćosić, T.; Zdravković-Korać, S.; Momčilović, I.; Smigocki, A.; Ninković, S. *Euphytica* **2014**, *198*, 425-438.

(18) Chang, M.-M.; Culley, D.; Choi, J. J.; Hadwiger, L. A. *Plant Sci.* **2002**, *163*, 83-89.

(19) Morris, W. L.; Ducreux, L. J. M.; Fraser, P. D.; Millam, S.; Taylor, M. A. *Metab. Eng.* **2006**, *8*, 253-263.

(20) Romano, A.; Raemakers, K.; Bernardi, J.; Visser, R.; Mooibroek, H. *Transgenic Res.* **2003**, *12*, 461-473.

(21) Kebeish, R.; Niessen, M.; Thiruveedhi, K.; Bari, R.; Hirsch, H.-J.; Rosenkranz, R.; Stabler, N.; Schonfeld, B.; Kreuzaler, F.; Peterhansel, C. *Nat. Biotechnol.* **2007**, *25*, 593-599.

(22) Jonik, C.; Sonnewald, U.; Hajirezaei, M.-R.; Flügge, U.-I.; Ludewig, F. *Plant Biotechnol. J.* **2012**, *10*, 1088-1098.

(23) Rosellini, D. *In Vitro Cell. Dev. Biol. Plant* **2011**, *47*, 222-233.

(24) Righetti, L.; Djennane, S.; Berthelot, P.; Cournol, R.; Wilmot, N.; Loridon, K.; Vergne, E.; Chevreau, E. *Plant Cell Tiss. Org. Cult.* **2014**, *117*, 335-348.

(25) López-Noguera, S.; Petri, C.; Burgos, L. *Plant Cell Rep.* **2009**, *28*, 1781-1790.

(26) Bhatnagar, M.; Prasad, K.; Bhatnagar-Mathur, P.; Lakshmi Narasu, M.; Waliyar, F.; Sharma, K. *Plant Cell Rep.* **2010**, *29*, 495-502.

(27) Srivastava, V.; Gidoni, D. *In Vitro Cell. Dev. Biol. Plant* **2010**, *46*, 219-232.

(28) Lechtenberg, B.; Schubert, D.; Forsbach, A.; Gils, M.; Schmidt, R. *Plant J.* **2003**, *34*, 507-517.

(29) Nölke, G.; Houdelet, M.; Kreuzaler, F.; Peterhänsel, C.; Schillberg, S. *Plant Biotechnol. J.* **2014**, *12*, 734-742.

(30) Jang, S. K.; Kräusslich, H. G.; Nicklin, M. J.; Duke, G. M.; Palmenberg, A. C.; Wimmer, E. *J. Virol.* **1988**, *62*, 2636-2643.

(31) Pelletier, J.; Sonenberg, N. *Nature* **1988**, *334*, 320-325.

(32) Urwin, P.; Yi, L.; Martin, H.; Atkinson, H.; Gilmartin, P. M. *Plant J.* **2000**, *24*, 583-589.

(33) Urwin, P. E.; Zubko, E. I.; Atkinson, H. J. *Physiol. Mol. Plant Pathol.* **2002**, *61*, 103-108.

(34) Mizuguchi, H.; Xu, Z.; Ishii-Watabe, A.; Uchida, E.; Hayakawa, T. *Mol. Ther.* **2000**, *1*, 376-382.

(35) Martinez-Salas, E. *Curr. Opin. Biotechnol.* **1999**, *10*, 458-464.

(36) Ha, S.-H.; Liang, Y. S.; Jung, H.; Ahn, M.-J.; Suh, S.-C.; Kweon, S.-J.; Kim, D.-H.; Kim, Y.-M.; Kim, J.-K. *Plant Biotechnol. J.* **2010**, *8*, 928-938.

(37) Douin, V.; Bornes, S.; Creancier, L.; Rochaix, P.; Favre, G.; Prats, A.-C.; Couderc, B. *BMC Biotechnol.* **2004**, *4*, 16.

(38) Ramos, R.; Martinez-Salas, E. *RNA* **1999**, *5*, 1374-1383.

(39) Stewart, S. R.; Semler, B. L. *Semin. Virol.* **1997**, *8*, 242-255.

(40) Martinez-Salas, E.; Ramos, R.; Lafuente, E.; Lopez de Quinto, S. *J. Gen. Virol.* **2001**, *82*, 973-984.

(41) Baird, S. D.; Turcotte, M.; Korneluk, R. G.; Holcik, M. *RNA* **2006**, *12*, 1755-1785.

(42) Ryan, M. D.; King, A. M. Q.; Thomas, G. P. *J. Gen. Virol.* **1991**, *72*, 2727-2732.

(43) Sommergruber, W.; Zorn, M.; Blaas, D.; Fessl, F.; Volkmann, P.; Maurer-Fogy, I.; Pallai, P.; Merluzzi, V.; Matteo, M.; Skern, T.; Kuechler, E. *Virology* **1989**, *169*, 68-77.

(44) Bazan, J. F.; Fletterick, R. J. *Proc. Natl. Acad. Sci. USA* **1988**, *85*, 7872-7876.

(45) Jackson, R. J. *Virology* **1986**, *149*, 114-127.

(46) Robertson, B. H.; Grubman, M. J.; Weddell, G. N.; Moore, D. M.; Welsh, J. D.; Fischer, T.; Dowbenko, D. J.; Yansura, D. G.; Small, B.; Kleid, D. G. *J. Virol.* **1985**, *54*, 651-660.

(47) Ryan, M. D.; Drew, J. *EMBO J.* **1994**, *13*, 928-933.

(48) Donnelly, M. L.; Gani, D.; Flint, M.; Monaghan, S.; Ryan, M. D. *J. Gen. Virol.* **1997**, *78*, 13-21.

(49) Donnelly, M. L. L.; Hughes, L. E.; Luke, G.; Mendoza, H.; ten Dam, E.; Gani, D.; Ryan, M. D. *J. Gen. Virol.* **2001**, *82*, 1027-1041.

(50) Ryan, M. D.; Donnelly, M.; Lewis, A.; Mehrotra, A. P.; Wilkie, J.; Gani, D. *Bioorg. Chem.* **1999**, *27*, 55-79.

(51) Donnelly, M. L. L.; Luke, G.; Mehrotra, A.; Li, X.; Hughes, L. E.; Gani, D.; Ryan, M. D. *J. Gen. Virol.* **2001**, *82*, 1013-1025.

(52) Minskaia, E.; Ryan, M. D. *BioMed Res. Int.* **2013**, *2013*, 12.

(53) Nissen, P.; Hansen, J.; Ban, N.; Moore, P. B.; Steitz, T. A. *Science* **2000**, *289*, 920-930.

(54) El Amrani, A.; Barakate, A.; Askari, B. M.; Li, X.; Roberts, A. G.; Ryan, M. D.; Halpin, C. *Plant Physiol.* **2004**, *135*, 16-24.

(55) Randall, J.; Sutton, D.; Ghoshroy, S.; Bagga, S.; Kemp, J. D. *Plant Sci.* **2004**, *167*, 367-372.

(56) Samalova, M.; Fricker, M.; Moore, I. *Traffic* **2006**, *7*, 1701-1723.

(57) François, I. E. J. A.; Van Hemelrijck, W.; Aerts, A. M.; Wouters, P. F. J.; Proost, P.; Broekaert, W. F.; Cammue, B. P. A. *Plant Sci.* **2004**, *166*, 113-121.

(58) Sun, H.; Lang, Z.; Zhu, L.; Huang, D. *Plant Cell Rep.* **2012**, *31*, 1877-1887.

(59) Yang, S.; Cohen, C. J.; Peng, P. D.; Zhao, Y.; Cassard, L.; Yu, Z.; Zheng, Z.; Jones, S.; Restifo, N. P.; Rosenberg, S. A.; Morgan, R. A. *Gene Ther.* **2008**, *15*, 1411-1423.

(60) Camper, N.; Byrne, T.; Burden, R. E.; Lowry, J.; Gray, B.; Johnston, J. A.; Migaud, M. E.; Olwill, S. A.; Buick, R. J.; Scott, C. J. *J. Immunol. Methods* **2011**, *372*, 30-41.

(61) Urwin, P. E.; McPherson, M. J.; Atkinson, H. J. *Planta* **1998**, *204*, 472-479.

(62) François, I. E. J. A.; Bolle, M. F. C. D.; Dwyer, G.; Goderis, I. J. W. M.; Woutors, P. F. J.; Verhaert, P. D.; Proost, P.; Schaaper, W. M. M.; Cammue, B. P. A.; Broekaert, W. F. *Plant Physiol.* **2002**, *128*, 1346-1358.

(63) Marcos, J. F.; Beachy, R. N. *J. Gen. Virol.* **1997**, *78*, 1771-1778.

(64) Dasgupta, S.; Collins, G. B.; Hunt, A. G. *Plant J.* **1998**, *16*, 107-116.

(65) Walker, J. M.; Vierstra, R. D. *Plant Biotechnol. J.* **2007**, *5*, 413-421.

(66) Kinal, H.; Park, C. M.; Berry, J. O.; Koltin, Y.; Bruenn, J. A. *Plant Cell* **1995**, *7*, 677-688.

(67) Jiang, L.; Rogers, J. C. *Plant J.* **1999**, *18*, 23-32.

(68) Janzik, I.; Macheroux, P.; Amrhein, N.; Schaller, A. *J. Biol. Chem.* **2000**, *275*, 5193-5199.

(69) Liu, J.-X.; Srivastava, R.; Howell, S. *Planta* **2009**, *230*, 687-697.

(70) Shmuel, P. *Trends Genet.* **1996**, *12*, 287-288.

(71) Mathys, S.; Evans Jr, T. C.; Chute, I. C.; Wu, H.; Chong, S.; Benner, J.; Liu, X.-Q.; Xu, M.-Q. *Gene* **1999**, *231*, 1-13.

(72) Evans, T. C.; Martin, D.; Kolly, R.; Panne, D.; Sun, L.; Ghosh, I.; Chen, L.; Benner, J.; Liu, X.-Q.; Xu, M.-Q. *J. Biol. Chem.* **2000**, *275*, 9091-9094.

(73) Nawrath, C.; Poirier, Y.; Somerville, C. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 12760-12764.

(74) Potrykus, I. *Plant Physiol.* **2001**, *125*, 1157-1161.

(75) Ma, J. K.; Hein, M. B. *Annals NY Academy of Sciences* **1996**, *792*, 72-81.

(76) Halpin, C.; Barakate, A.; Askari, B. M.; Abbott, J. C.; Ryan, M. D. *Plant Mol. Biol.* **2001**, *47*, 295-310.

(77) de Felipe, P.; Luke, G. A.; Hughes, L. E.; Gani, D.; Halpin, C.; Ryan, M. D. *Trends Biotechnol.* **2006**, *24*, 68-75.

(78) Hunt, A.; Maiti, I. *In Vitro Cell. Dev. Biol. Plant* **2001**, *37*, 313-320.

(79) Marcos, J. F.; Beachy, R. N. *Plant Mol. Biol.* **1994**, *24*, 495-503.

(80) Ceriani, M. F.; Marcos, J. F.; Hopp, H. E.; Beachy, R. N. *Plant Mol. Biol.* **1998**, *36*, 239-248.

(81) François, I. E. J. A.; Dwyer, G. I.; De Bolle, M. F. C.; Goderis, I. J. W. M.; Van Hemelrijck, W.; Proost, P.; Wouters, P.; Broekaert, W. F.; Cammue, B. P. A. *Plant Physiol. Biochem.* **2002**, *40*, 871-879.

(82) François, I.; Broekaert, W.; Cammue, B. *Plant Sci.* **2002**, *163* 281-295.

(83) Wilcox, C. A.; Fuller, R. S. *J. Cell Biol.* **1991**, *115*, 297-307.

(84) Su, W. W. *Microb. Cell Fact.* **2005**, *4*, 12.

(85) Haseloff, J.; Siemering, K. R.; Prasher, D. C.; Hodge, S. *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 2122-2127.

(86) Goodin, M. M.; Dietzgen, R. G.; Schichnes, D.; Ruzin, S.; Jackson, A. O. *Plant J.* **2002**, *31*, 375-383.

(87) Ni, M.; Cui, D.; Einstein, J.; Narasimhulu, S.; Vergara, C. E.; Gelvin, S. B. *Plant J.* **1995**, *7*, 661-676.

(88) Becerra-Arteaga, A.; Mason, H. S.; Shuler, M. L. *Biotechnol. Prog.* **2006**, *22*, 1643-1649.

(89) Datla, R.; Bekkaoui, F.; Hammerlindl, J.; Pilate, G.; Dunstan, D.; Crosby, W. *Plant Sci.* **1993**, *94*, 139-149.

(90) Peckham, G. D.; Bugos, R. C.; Su, W. W. *Protein Expr. Purif.* **2006**, *49*, 183-189.

(91) Su, W. W.; Guan, P.; Bugos, R. C. *Biotechnol. Bioeng.* **2004**, *85*, 610-619.

(92) Murashige, T.; Skoog, F. *Physiol. Plant.* **1962**, *15*, 473-497.

(93) Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11984-11989.

(94) Litowski, J. R.; Hodges, R. S. *J. Biol. Chem.* **2002**, *277*, 37272-37279.

(95) Rahimi, Y.; Shrestha, S.; Deo, S. K. *Chromatographia* **2007**, *65*, 429-433.

(96) Bader, O.; Krauke, Y.; Hube, B. *BMC Micobiology* **2008**, *8*, 116.

(97) Joh, L. D.; Wroblewski, T.; Ewing, N. N.; VanderGheynst, J. S. *Biotechnol. Bioeng.* **2005**, *91*, 861-871.

(98) Kitamura, T.; Tange, T.; Terasawa, T.; Chiba, S.; Kuwaki, T.; Miyagawa, K.; Piao, Y. F.; Miyazono, K.; Urabe, A.; Takaku, F. *J. Cell Physiol.* **1989**, *140*, 323-334.

(99) Jiang, L.; Rogers, J. C. *J. Cell Biol.* **1998**, *143*, 1183-1199.

(100) Gross, L. A.; Baird, G. S.; Hoffman, R. C.; Baldridge, K. K.; Tsien, R. Y. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11990-11995.

(101) Sacchetti, A.; Subramaniam, V.; Jovin, T. M.; Alberti, S. *FEBS Lett.* **2002**, *525*, 13-19.

(102) Sardana, R.; Dudani, A. K.; Tackaberry, E.; Alli, Z.; Porter, S.; Rowlandson, K.; Ganz, P.; Altosaar, I. *Transgenic Res.* **2007**, *16*, 713-721.

(103) James, E. A.; Wang, C.; Wang, Z.; Reeves, R.; Shin, J. H.; Magnuson, N. S.; Lee, J. M. *Protein Expr. Purif.* **2000**, *19*, 131-138.

(104) Shin, Y. J.; Hong, S. Y.; Kwon, T. H.; Jang, Y. S.; Yang, M. S. *Biotechnol. Bioeng.* **2003**, *82*, 778-783.

(105) Ni, M.; Cui, D.; Einstein, J.; Narasimhulu, S.; Vergara, C.; Gelvin, S. *Plant J.* **1995**, *7*, 661-676.

(106) D'Aoust, M. A.; Lavoie, P. O.; Belles-Isles, J.; Bechtold, N.; Martel, M.; Vezina, L. P. *Methods Mol. Biol.* **2009**, *483*, 41-50.

(107) van Esse, H. P.; Thomma, B. P.; van 't Klooster, J. W.; de Wit, P. J. *J. Exp. Bot.* **2006**, *57*, 599-608.

(108) Pizzuti, F.; Daroda, L. *Environ. Biosaf. Res.* **2008**, *7*, 219-226.

(109) Broekaert, W. F.; Francois, I. E. J. A.; De Bolle, M. F. C.; Ian, J.; Ray, J. A. Genetic method for the expression of polyproteins in plants. WO 2000011175A1 2000.

(110) Jiang, L.; Sun, S. S. *Trends Biotechnol.* **2002**, *20*, 99-102.

(111) Faye, L.; Boulaflous, A.; Benchabane, M.; Gomord, V.; Michaud, D. *Vaccine* **2005**, *23*, 1770-1778.

(112) James, E.; Lee, J. M. *Adv. Biochem. Eng. Biotechnol.* **2001**, *72*, 127-156.

(113) Lee, J. H.; Kim, N. S.; Kwon, T. H.; Jang, Y. S.; Yang, M. S. *J. Biotechnol.* **2002**, *96*, 205-211.

(114) Shen, B.; Sun, X.; Zuo, X.; Shilling, T.; Apgar, J.; Ross, M.; Bougri, O.; Samoylov, V.; Parker, M.; Hancock, E.; Lucero, H.; Gray, B.; Ekborg, N. A.; Zhang, D.; Johnson, J. C.; Lazar, G.; Raab, R. M. *Nat. Biotechnol.* **2012**, *30*, 1131-1136.

(115) Chong, S.; Montello, G. E.; Zhang, A.; Cantor, E. J.; Liao, W.; Xu, M.-Q.; Benner, J. *Nucleic Acids Res.* **1998**, *26*, 5109-5115.

(116) Muir, T. W.; Sondhi, D.; Cole, P. A. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 6705-6710.

(117) Williams, N. K. *J. Biol. Chem.* **2001**, *277*, 7790-7798.

(118) Fang, J.; Qian, J.-J.; Yi, S.; Harding, T. C.; Tu, G. H.; VanRoey, M.; Jooss, K. *Nat. Biotechnol.* **2005**, *23*, 584-590.

(119) Hamorsky, K. T.; Grooms-Williams, T. W.; Husk, A. S.; Bennett, L. J.; Palmer, K. E.; Matoba, N. *Antimicrob. Agents Chemother.* **2013**, *57*, 2076-2086.

(120) Szymczak, A. L.; Workman, G. J.; Wang, Y.; Vignali, K. M.; Dilioglou, S.; Vanin, E. F.; Vignali, D. A. A. *Nat. Biotechnol.* **2004**, *22*, 589-594.

(121) Kerrigan, J. J.; Xie, Q.; Ames, R. S.; Lu, Q. *Protein Expr. Purif.* **2011**, *75*, 1-14.

(122) Brazier-Hicks, M.; Edwards, R. *Metab. Eng.* **2013**, *16*, 11-20.

(123) Vranová, E.; Coman, D.; Gruissem, W. *Annu. Rev. Plant Biol.* **2013**, *64*, 665-700.

(124) Bieniossek, C.; Imasaki, T.; Takagi, Y.; Berger, I. *Trends Biochem. Sci.* **2012**, *37*, 49-57.

(125) Halpin, C.; Ryan, M. D. In *Metabolic Engineering in the Post-Genomic Era*, Kholodenko, B.; Westerhoff, H., Eds. Horizon Scientific Press: Norfolk, England, 2004; pp 377-408.

(126) de Felipe, P.; Luke, G. A.; Brown, J. D.; Ryan, M. D. *Biotechnol. J.* **2010**, *5*, 213-223.

(127) Sharma, P.; Yan, F.; Doronina, V. A.; Escuin-Ordinas, H.; Ryan, M. D.; Brown, J. D. *Nucleic Acids Res.* **2012**, *40*, 3143-3151.

(128) Amitai, G.; Callahan, B. P.; Stanger, M. J.; Belfort, G.; Belfort, M. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 11005-11010.

(129) Paramban, R. I.; Bugos, R. C.; Su, W. W. *Biotechnol. Bioeng.* **2004**, *86*, 687-697.

(130) Han, Z.; Zhang, B.; Wang, Y. E.; Zuo, Y. Y.; Su, W. W. *Appl. Environ. Microbiol.* **2012**, *78*, 3249-3255.

(131) Mathys, S.; Evans, T. C.; Chute, I. C.; Wu, H.; Chong, S.; Benner, J.; Liu, X. Q.; Xu, M. Q. *Gene* **1999**, *231*, 1-13.

(132) Lee, L. Y.; Kononov, M. E.; Bassuner, B.; Frame, B. R.; Wang, K.; Gelvin, S. B. *Plant Physiol.* **2007**, *145*, 1294-1300.

(133) Lobstein, J.; Emrich, C.; Jeans, C.; Faulkner, M.; Riggs, P.; Berkmen, M. *Microb. Cell Fact.* **2012**, *11*, 56.

(134) Mayo, K. J.; Gonzales, B. J.; Mason, H. S. *Nat. Protoc.* **2006**, *1*, 1105-1111.

(135) Fisher, D.; Guiltinan, M. *Plant Mol. Biol. Rep.* **1995**, *13*, 278-289.

(136) Zhang, B.; Rapolu, M.; Huang, L.; Su, W. W. *Plant Biotechnol. J.* **2011**, *9*, 970-981.

(137) Xu, M. Q.; Evans, T. C., Jr. *Curr. Opin. Biotechnol.* **2005**, *16*, 440-446.

(138) Campbell, R. E.; Tour, O.; Palmer, A. E.; Steinbach, P. A.; Baird, G. S.; Zacharias, D. A.; Tsien, R. Y. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7877-7882.

(139) Olinger, G. G., Jr.; Pettitt, J.; Kim, D.; Working, C.; Bohorov, O.; Bratcher, B.; Hiatt, E.; Hume, S. D.; Johnson, A. K.; Morton, J.; Pauly, M.; Whaley, K. J.; Lear, C. M.; Biggins, J. E.; Scully, C.; Hensley, L.; Zeitlin, L. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 18030-18035.

(140) Qiu, X.; Wong, G.; Audet, J.; Bello, A.; Fernando, L.; Alimonti, J. B.; Fausther-Bovendo, H.; Wei, H.; Aviles, J.; Hiatt, E.; Johnson, A.; Morton, J.; Swope, K.; Bohorov, O.; Bohorova, N.; Goodman, C.; Kim, D.; Pauly, M. H.; Velasco, J.; Pettitt, J.; Olinger, G. G.; Whaley, K.; Xu, B.; Strong, J. E.; Zeitlin, L.; Kobinger, G. P. *Nature* **2014**, *514*, 47-53.

(141) Paulus, H. *Chem. Soc. Rev.* **1998**, *27*, 375-386.

(142) Xu, M. Q.; Perler, F. B. *EMBO J.* **1996**, *15*, 5146-5153.

(143) Chong, S.; Williams, K. S.; Wotkowicz, C.; Xu, M.-Q. *J. Biol. Chem.* **1998**, *273*, 10567-10577.

(144) Shah, N. H.; Muir, T. W. *Chem. Sci. J.* **2014**, *5*, 446-461.

(145) Frutos, S.; Goger, M.; Giovani, B.; Cowburn, D.; Muir, T. W. *Nat. Chem. Biol.* **2010**, *6*, 527-533.

(146) Ramirez, M.; Valdes, N.; Guan, D.; Chen, Z. *Protein Eng. Des. Sel.* **2013**, *26*, 215-223.

(147) von Arnim, A. G.; Deng, X. W.; Stacey, M. G. *Gene* **1998**, *221*, 35-43.

(148) Seibel, N. M.; Eljouni, J.; Nalaskowski, M. M.; Hampe, W. *Anal. Biochem.* **2007**, *368*, 95-99.

(149) Beetz, C.; Brodhun, M.; Moutzouris, K.; Kiehntopf, M.; Berndt, A.; Lehnert, D.; Deufel, T.; Bastmeyer, M.; Schickel, J. *Biochem. Biophys. Res. Commun.* **2004**, *318*, 1079-1084.

(150) Chen, J.; Song, J.-L.; Zhang, S.; Wang, Y.; Cui, D.-F.; Wang, C.-C. *J. Biol. Chem.* **1999**, *274*, 19601-19605.

(151) Martinez-Alonso, M.; Garcia-Fruitos, E.; Ferrer-Miralles, N.; Rinas, U.; Villaverde, A. *Microb. Cell Fact.* **2010**, *9*, 64.

(152) Pedersen, J.; Lauritzen, C.; Madsen, M. T.; Weis Dahl, S. *Protein Expr. Purif.* **1999**, *15*, 389-400.

(153) Arvinth, S.; Arun, S.; Selvakesavan, R. K.; Srikanth, J.; Mukunthan, N.; Ananda Kumar, P.; Premachandran, M. N.; Subramonian, N. *Plant Cell Rep.* **2010**, *29*, 383-395.

(154) Chen, L.; Marmey, P.; Taylor, N. J.; Brizard, J.-P.; Espinoza, C.; D'Cruz, P.; Huet, H.; Zhang, S.; Kochko, A. d.; Beachy, R. N.; Fauquet, C. M. *Nat. Biotechnol.* **1998**, *16*, 1060-1064.

(155) Suresh, B.; Ravishankar, G. A. *Crit. Rev. Biotechnol.* **2004**, *24*, 97-124.

(156) Lee, D.-S.; Lee, K.-H.; Jung, S.; Jo, E.-J.; Han, K.-H.; Bae, H.-J. *J. Exp. Bot.* **2012**, *63*, 4797-4810.

(157) Vaquero, C.; Sack, M.; Chandler, J.; Drossard, J.; Schuster, F.; Monecke, M.; Schillberg, S.; Fischer, R. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 11128-11133.

(158) James, C. *GM Crops* **2010**, *1*, 8-12.

(159) Ceriani, M. F.; Marcos, J.; Esteban Hopp, H.; Beachy, R. *Plant Mol. Biol.* **1998**, *36*, 239-248.

(160) de Felipe, P. *Genet. Vaccines Ther.* **2004**, *2*, 13.

(161) Zhang, B.; Rapolu, M.; Liang, Z.; Han, Z.; Williams, P. G.; Su, W. W. *Sci. Rep.* **2015**, *5*, 8541.

(162) Bradford, M. M. *Anal. Biochem.* **1976**, *72*, 248-254.

(163) Pimpl, P. *Plant Cell* **2006**, *18*, 198-211.

(164) Minskaia, E.; Nicholson, J.; Ryan, M. *BMC Biotechnol.* **2013**, *13*, 1-11.

(165) Yan, F.; Doronina, V. A.; Sharma, P.; Brown, J. D. *Biochem. Soc. Trans.* **2010**, *38*, 1576-1580.

(166) Martin, D. D.; Xu, M.-Q.; Evans, T. C. *Biochemistry* **2001**, *40*, 1393-1402.

(167) Saleh, L.; Southworth, M. W.; Considine, N.; O'Neill, C.; Benner, J.; Bollinger, J. M.; Perler, F. B. *Biochemistry* **2011**, *50*, 10576-10589.

(168) Trewavas, A. *Plant Physiol.* **1972**, *49*, 40-46.

(169) de Felipe, P.; Ryan, M. *Traffic* **2004**, *5*, 616-626.

(170) Carpita, N.; Sabularse, D.; Montezinos, D.; Delmer, D. P. *Science* **1979**, *205*, 1144-1147.

(171) Schillberg, S.; Fischer, R.; Emans, N. *CMLS, Cell. Mol. Life Sci.* **2003**, *60*, 433-445.

(172) Xiang, L.; Etxeberria, E.; Van den Ende, W. *FEBS J.* **2013**, *280*, 979-993.

(173) Davidson, H. W. *J. Cell Biol.* **1995**, *130*, 797-805.

(174) Fujiwara, T.; Oda, K.; Yokota, S.; Takatsuki, A.; Ikehara, Y. *J. Biol. Chem.* **1988**, *263*, 18545-18552.