Electronic Thesis and Dissertation Repository

February 2019

Uncovering a Mystery of the Isoflavonoid Metabolon: Identification and Characterization of the Arogenate Dehydratase Gene Family in Soybean

Kelsey Pannunzio
The University of Western Ontario

Supervisor
Dr. Sangeeta Dhaubhadel
Agriculture and Agri-Foods Canada, London Research and Development Centre Co-Supervisor
Dr. David Smith
The University of Western Ontario

Graduate Program in Biology
A thesis submitted in partial fulfillment of the requirements for the degree in Master of Science
© Kelsey Pannunzio 2019

Follow this and additional works at: https://ir.lib.uwo.ca/etd

Part of the Molecular Biology Commons

Abstract

Soybean (*Glycine max*) is a vastly important, multi-billion dollar global commodity; but this crop's yields are under threat from the pathogen *Phytophthora sojae,* which causes extensive stem and root rot in soybean crops. Isoflavonoids, a metabolite class unique to legumes, are a promising research target to combat *P. sojae.* Isoflavonoids are released as phytoalexins in response to stress, and also facilitate interactions with nitrogen-fixing rhizobial bacteria through nodule formation. An isoflavonoid biosynthesis metabolon was discovered in soybean through co-immunoprecipitation, anchored to the endoplasmic reticulum by isoflavone synthase (IFS) and Cinnamate 4-hydroxylase, two cytochrome P450 enzymes. One of the IFS-interacting partners discovered was arogenate dehydratase (ADT), which synthesizes phenylalanine, the precursor to the phenylpropanoid pathway. Generally ADTs are localized to the chloroplast, making the interaction with IFS seemingly impossible. In the current study, 9 *GmADTs* were identified, and the GmADT-GmIFS interaction was confirmed. These findings broaden the current knowledge of the isoflavonoid metabolon.

Acknowledgements

I owe a tremendous amount of gratitude to my supervisor, Dr. Dhaubhadel for her guidance and vision and for giving me this opportunity. I owe the completion of my thesis to her; had it not been for her keeping me on track I could very well still have been writing to this day! Thank you to my committee members, Dr. Smith, Dr. Menassa, and Dr. Kohalmi – your suggestions for my experiments and writing were invaluable and helped me greatly. Thank you to Arzie, Carol, and Diane, you are the glue that holds the department together. I also want to immensely thank Ling Chen. Not once when I asked for assistance did she ever say "I'm too busy"; she was ALWAYS willing to drop everything to offer assistance that ultimately helped me with so many technical issues I was facing with my research. Also her constant words of encouragement really brightened my spirits on days when I felt like giving up.

Arun, Nishat, Smarth, Kevin, Arjun, Gary, Jaeju, Ramtin, Jordan and Simran, my wonderful labmates, I could not have gotten through the last two years without all of you. To Shrikaar, Reza, Adam, Craig, Scott, Coby, Sanjay, Sommer, Ondre, Emily (1 and 2), Mandana, Kaitlyn, Carlie, Jackie, Bonnie, Brad, Jordan and Lisa, thank you so much for the friendships and comradery over the last two years, from coffee breaks to the fun times out of the lab, you made the struggles of graduate school seem easier.

Lastly, this thesis is dedicated to my parents, Pat and Kelly, and my Grammy. Without their support, there is no way I would have been able to finish my thesis, let alone my undergraduate degree. They have been incredibly supportive all these years; when I believed I wasn't as smart as others, or not capable, they always told me otherwise and always gave me a revitalized sense of determination. Thank you so much for everything.

List of Abbreviations

*Standard SI units not listed

2HID	2-hydroxyisoflavanone dehydratase
4CL	4–coumarate:CoA ligase
AbMV MP	*Abutilon* mosaic virus movement protein
ACT	<u>a</u>spartokinase, <u>c</u>horismate mutase, and <u>T</u>yrA
ADT	arogenate dehydratase
AgNO₃	silver nitrate
Arg	arginine
Asn	asparagine
AtADT	*Arabidopsis thaliana* ADT
att	attachment
BiFC	bimolecular fluorescence complementation
bp	base pair
C4H	cinnamate 4-hydroxylase
cDNA	complementary DNA
CDS	coding DNA sequence
CDT	cyclohexadienyl dehydratases
CHI	chalcone isomerase
CHR	chalcone reductase
CHS	chalcone synthase
CM	chorismate mutase

Co-IP	co-immunoprecipitation
cpHSC70-1	chloroplast heat shock protein 70-1
DAHP	3-Deoxy-D-arabinoheptulosonate 7-phosphate
ER	endoplasmic reticulum
F6H	flavone 6-hydroxylase
FPKM	fragments per kilobase per million mapped reads
Gln	glutamine
Glu	glutamic acid
GmADT	*Glycine max* arogenate dehydratase
GmPT	*Glycine max* prenyltransferases
IFS	isoflavone synthase
IOMT	O–methyltransferase
ispG	4-hydroxy-3-methylbut-2-enyl diphosphate synthase
LB	lysogeny broth
MEP	methylerythritol 4-phosphate
NLS	nuclear localization signal
NRIP1	nuclear receptor interacting protein 1
PAC	PDT conferring domain
PAL	phenylalanine ammonia lyase
PDAT1	phospholipid-diacylglycerol acyltransferase 1
PDT	prephenate dehydratase
pEarleyGate	(pEG)

PhADT	*Petunia hybrida* ADT
Phe	phenyalanine
PLACE	plant cis-acting regulatory DNA elements
PpADT	*Pinus pinaster* ADT
PPY-AT	phenylpyruvate aminotransferase
RNAseq	RNA sequencing
ROS	reactive oxygen species
RT-PCR	reverse transcription-polymerase chain reaction
Ser	serine
SRA	sequence read archive
TBE	tris/borate/EDTA
Thr	threonine
TP	transit peptides
Tyr	tyrosine
WGD	whole-genome duplications
YFP	yellow fluorescent protein

Table of Contents

List of Figures

List of Tables

List of Appendices

Chapter 1: Introduction

1.1 Soybean has extreme global importance

Soybean (*Glycine max)* is a globally invaluable commodity, with production worth hundreds of billions of dollars globally. Soybean seeds have very high protein and oil content, making them key players for a wide array of industrial uses. The high protein content makes soybean seeds useful for food processing, both for human consumption and animal feed production. However, soybean yields are threatened by a variety of pests and pathogens. The most notable of these pests is *Phytopthora sojae,* which on its own causes $2 billion in global crop losses annually (Tyler, 2007). Also threatening soybean crop yield are greater weather extremes, like drought, due to the changing global climate. As such, research targeted towards engineering more resistant soybeans is of utmost importance.

1.2 Isoflavonoids as a research target

Isoflavonoids are a family of specialized metabolites unique to legume plants (such as soybeans, alfalfa, lentils, etc). They make excellent targets for studying soybean defense against pathogens, since they are phytoalexins released as part of the basal immune response. Phytoalexins are low-molecular weight antimicrobial compounds synthesized upon pathogenic attack. They alter pathogenic growth and activity by interfering with enzymatic activity or preventing proper metabolism in the invading pathogens (Kaplan *et al.*, 1980, Boydston *et al.*, 1982, John L. Giannini *et al.*, 1988). Because of their role as phytoalexins, it is of great interest to select soybean cultivars that

can accumulate greater concentrations of isoflavonoids, in order to strengthen their basal immune system and reduce the need for chemical pesticides. In fact, upregulating both stilbenes, which are non-native to soybean, and isoflavonoid glyceollins, which are common soybean phytoalexins, caused a greater resistance against a pathogenic fungi, illustrating the importance of phytoalexins (Zernova *et al.*, 2014). They have also been shown to be released in response to abiotic stresses, indicating their importance to all aspects of plant stress-response (Ayers *et al.*, 1976).

Isoflavonoids are also imperative to nodule formation. Nitrogen in the atmosphere is largely inaccessible to most plants, leaving them dependent on nitrogen present in the soil. Root nodules are an adaptation of legumes that allows for a symbiotic relationship with nitrogen-fixing bacteria in the soil, allowing leguminous plants to harness nitrogen directly from the atmosphere. Isoflavonoids are key signalling components of this process, as they direct the formation of these necessary nodules, by helping to induce the *nod* gene expression in nitrogen-fixing bacteria, with secondary roles in modulating auxin concentrations in the nodules (Subramanian *et al.*, 2006). In doing so, the upregulation of isoflavonoid biosynthesis has the potential to reduce the need for artificial nitrogen fertilizers, greatly impacting the environmental impact and large costs of these fertilizers.

Furthermore, isoflavonoids offer many benefits to the human diet. Isoflavonoid intake confers a reduction in the risk of many chronic diseases, such as cardiovascular disease, hormone-dependent cancers, and osteoporosis, as well as potential benefits to women suffering from menopausal symptoms (reviewed by Messina, 2010, Dastmalchi

and Dhaubhadel, 2014). Furthermore, isoflavonoids were shown to arrest gastrointestinal cancer cell growth, demonstrating their immense health benefits (Yanagihara *et al.*, 1993). Therefore, isoflavonoids are tremendously important not only for plant immunity and communication with the external environment, but also provide a key pharmacological research target.

1.3 Isoflavonoids are synthesized through a metabolon

Plant metabolic synthesis is vast and complicated, with countless overlapping pathways, all being individually controlled both temporally – during stage of development and in response to external stimuli- and spatially – in different plant tissues and subcellular organelles (Vogt, 2010). Another layer of control exists at a molecular level-the "metabolon". The metabolon involves enzymes and regulatory or structural elements, involved in the same pathway, congregating together in order to maximize metabolic efficiency by allowing pathway intermediates to accumulate near each other in relatively high concentrations, which reduces diffusion and excess energy being spent on metabolite transport (Ralston *et al.*, 2005). This is especially relevant when considering isoflavonoid synthesis (Figure 1.1), which is a branch off the highly redundant and overlapping phenylpropanoid set of reactions.
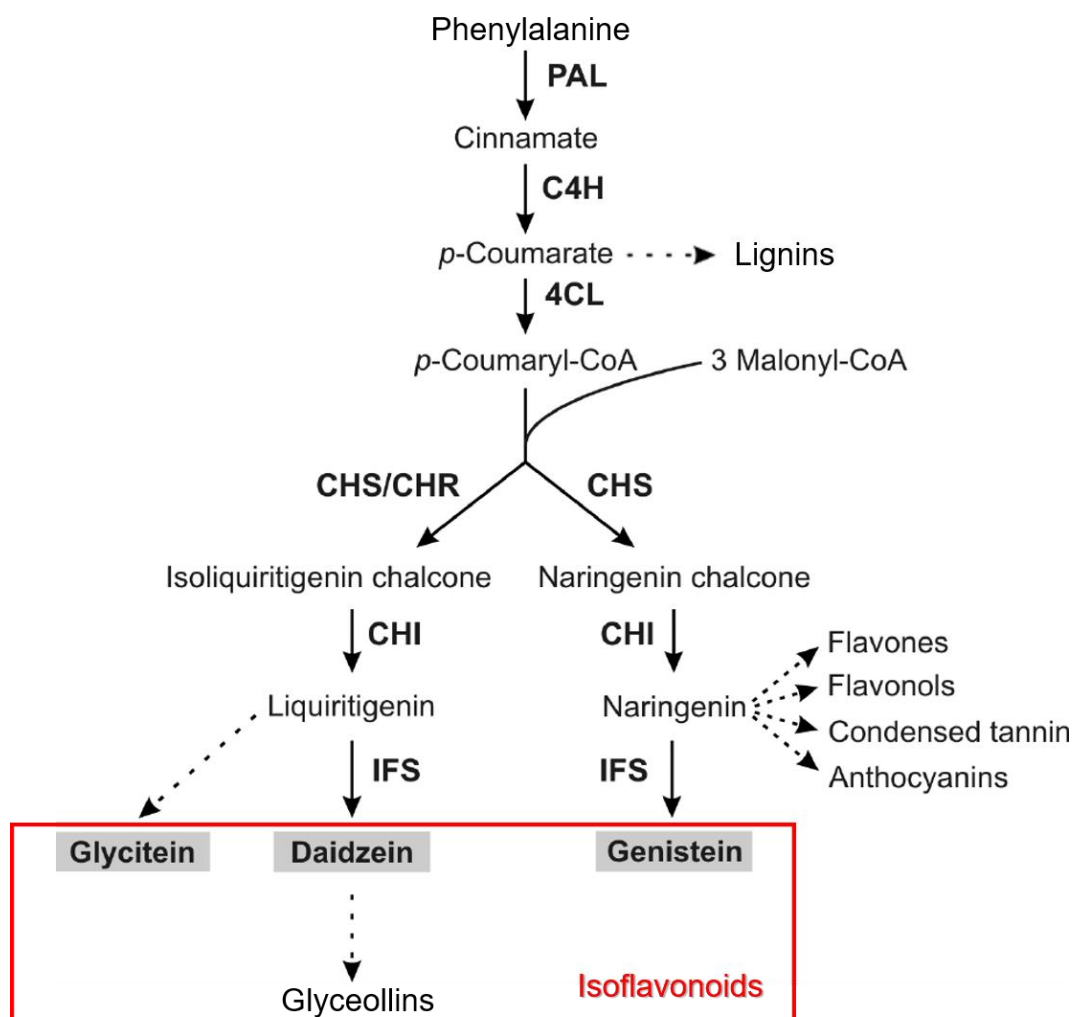
Figure 1.1| Overview of the entire phenylpropanoid biosynthetic pathway. Shown is the the phenylpropanoid pathway resulting in the synthesis of a variety of metabolites including lignins, flavones, flavonols, condensed tannins, anthocyanins and isoflavonoids. Dashed arrows indicate multiple steps involved. Abbreviations: PAL, phenylalanine ammonia lyase; C4H, cinnamate 4–hydroxylase; 4CL, 4–coumarate:CoA ligase; CHS, chalcone synthase; CHR, chalcone reductase; CHI, chalcone isomerase; IFS, isoflavone synthase (modified from Dastmalchi and Dhaubhadel, 2004).

These aggregations of enzymes and other elements require an anchoring point. Much evidence has been discovered that points to the ER as the prime scaffolding location. Cytochrome P450 proteins are integral membrane proteins that are anchored in the ER membrane, with their catalytic sites exposed to the cytosol (Neve and Ingelman-Sundberg, 2008). Because of this structure, there is growing evidence suggesting that these P450s act as anchors to soluble enzymes in pathways. It was recently confirmed that there is a metabolon for isoflavonoid synthesis in soybean (Dastmalchi *et al.*, 2016). Through a co-immunoprecipitation (co-IP) experiment, using isoflavone synthase (IFS) - which is a cytochrome P450 - as a bait protein, interacting partners were elucidated (Figure 1.2). One of the interactors was a second cytochrome P450, cinnamate 4–hydroxylase (C4H), which catalyzes the conversion of cinnamic acid to *p*-coumaric acid early on in the phenylpropanoid pathway (Figure 1.1). This interaction was confirmed with bimolecular fluorescence complementation assay (BiFC), demonstrating a strong interaction between the two anchoring enzymes (Figure 1.2) (Dastmalchi *et al.*, 2016). Other IFS-interacting partners were also pulled down in the co-IP experiment, including chalcone isomerase (CHI), chalcone synthase (CHS) and chalcone reductase (CHR), which are all enzymes in the isoflavonoid pathway (Figure 1.1 and Figure 1.2). These interactions with IFS, as well as interactions between these soluble enzymes and a CHS-C4H interaction, were also all confirmed with BiFC (Figure 1.2). These findings strongly suggest the existence of an isoflavonoid metabolon, anchored to the ER membrane by C4H and IFS P450s (Dastmalchi *et al.*, 2016).
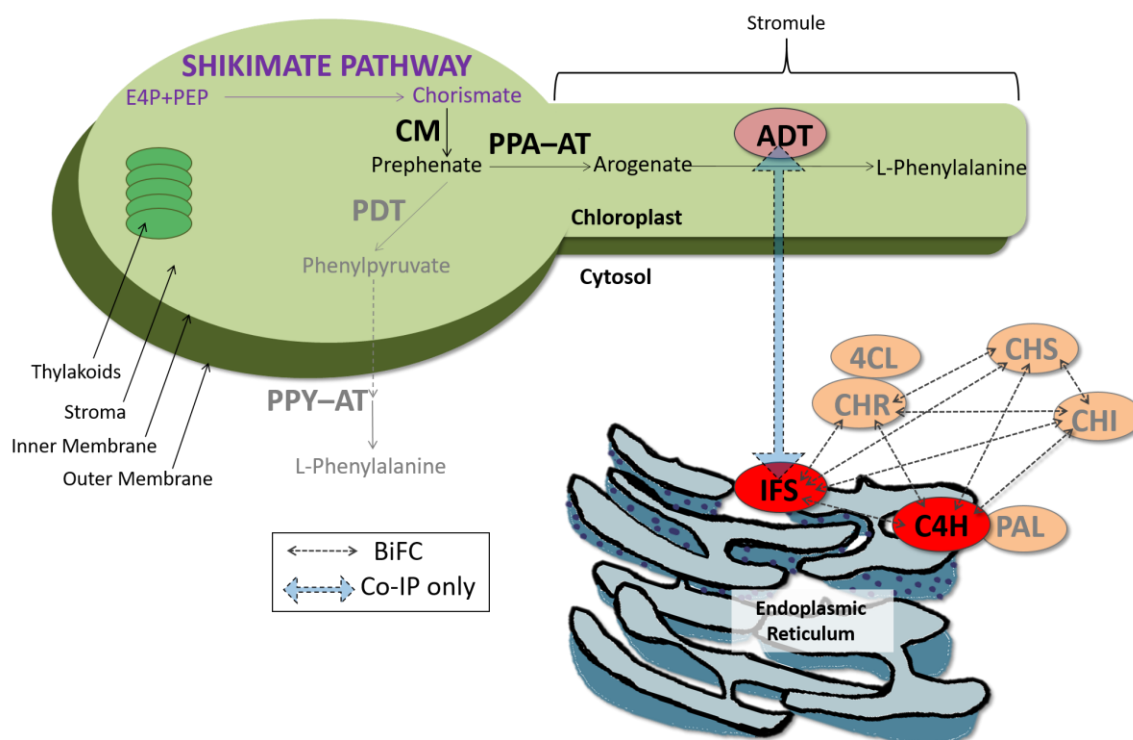
Figure 1.2| A model showing ADT interaction with the isoflavonoid metabolon. Two cytochrome P450 enzymes, IFS and C4H (⬤), anchor the cytosolic components of the isoflavonoid biosynthetic metabolon (⬤) to the ER. The interactions with the cytosolic enzymes were discovered using IFS as a bait protein in a co-immunoprecipitation experiment, and were confirmed via BiFC (⬍). ADT was also pulled-down in this co-IP experiment (⬍). Shown in purple is the shikimate pathway, in black is the ADT pathway of phenylalanine synthesis, and in grey is the PDT pathway of phenylalanine synthesis. Abbreviations: E4P, Erythrose 4-phosphate; PEP, Phosphoenol pyruvate; CM, chorismate mutase; PPA-AT, prephenate amino transferase; ADT, arogenate dehydratase; PDT, prephenate dehydratase; PPY-AT, phenylpyruvate aminotransferase; PAL, phenylalanine ammonia lyase; C4H, cinnamate 4–hydroxylase; 4CL, 4–coumarate:CoA ligase; CHS, chalcone synthase; CHR, chalcone reductase; CHI, chalcone isomerase; IFS, isoflavone synthase

1.4 ADTs may be involved in the isoflavonoid metabolon

In the study to identify the isoflavonoid metabolon (as explained earlier in Section 1.3), Dastmalchi *et al* (2016) also pulled-down two arogenate dehydratases (ADTs) in the Co-IP assay, Glyma.12G181800 and Glyma.13G319000. Generally ADT localizes in the chloroplast, where it catalyzes the conversion of arogenate into phenylalanine (phe), which is then transported into the cytosol where it is channeled into either primary metabolism such as protein synthesis or into the synthesis of specialized metabolites in the phenylpropanoid pathway such as lignins, flavones, or isoflavones (Figures 1.1 and 1.2). The ADT-IFS interaction was an unexpected finding, as ADTs have been shown to localize within the chloroplasts in other plant species (Cho *et al.*, 2007, Rippert *et al.*, 2009, Maeda *et al.*, 2010, El-Azaz *et al.*, 2016). Chloroplasts are double membrane-bound organelles, and given the ER membrane localization of IFS, this interaction was seemingly impossible. However, in *Arabidopsis,* one of the ADTs, AtADT6, localizes to the cytoplasm (Bross *et al.*, 2017), offering a potential explanation for the ADT-IFS interaction observed for Glyma.12G181800 and Glyma.13G319000.

The ADTs in *Arabidopsis* were found to be localizing to specific regions of the chloroplasts, known as stromules (Bross *et al.*, 2017). Stromules are dynamic cytoskeletal projections of the main chloroplast body, with stroma surrounded by the double-membrane of chloroplasts (Hanson and Sattarzadeh, 2008). The movement of metabolites between plastids through stromules has been shown; these stromule connections could possibly reduce the distance of the diffusion of these materials between plastids and different organelles (Hanson and Sattarzadeh, 2008). Furthermore,

there has also been research done that suggests stromules are involved in transmitting signals from the plant's external environment, implementing them in cellular response mechanisms (Hanson and Sattarzadeh, 2008).

A high degree of interconnectedness between the chloroplast and ER membranes has been previously suggested, through a phenomenon known as a trans-organelle continuity. First proposed by Whatley *et al.* in 1990, it describes the process where the ER membrane becomes continuous with the chloroplast outer membrane, potentially allowing the movement of metabolites between the organelles (Whatley *et al.*, 1990). There is a growing body of evidence to support this theory; for instance, it has been well documented in the literature that the ER and chloroplast membranes form contact sites, where the contact sites of the two membranes are held together with very strong attraction forces (Andersson *et al.*, 2007).

To further substantiate the potential existence of the trans-organelle continuity, it was determined that infection by the AbMV *Geminiviridae* virus induced a network of stromules forming interconnections between chloroplasts, nuclei, and even the external cell wall (Krenz *et al.*, 2012). BiFC determined that a chloroplast-localized chaperone protein, cpHSC70-1, and the AbMV movement protein (MP) were interacting. The authors showed increased AbMV MP in the nuclei, further implying the ability of stromules to form connections with other organelles and traffic proteins between them (Krenz *et al.*, 2012).

1.5 ADT discovery

The synthesis of phe in bacteria and fungi was well established by the early 1970s, but in higher plants, much was left unknown. In most micro-organisms, the synthesis of phe occurs through prephenate dehydratases (PDTs), in which prephenate is dehydrated and decarboxylated to phenylpyruvate, which is subsequently transaminated into phe. In these organisms, tyrosine is also synthesized using prephenate as a precursor, whereby prephenate is dehydrogenated, and then the intermediate, OH-phenylpyruvate, is transaminated to form tyrosine. Some micro-organisms, mainly members of the gram-negative bacteria phylum, have evolved dual function proteins that possess both chorismate mutase (CM), the enzyme responsible for converting chorismate into prephenate, and PDT domains (Dopheide *et al.*, 1972, Zhang *et al.*, 1998). It was later discovered that in various species of algae, a second set of pathways to phe and tyrosine synthesis was present (Stenmark *et al.*, 1974). In these pathways, prephenate is first transaminated to form arogenate, which either undergoes decarboxylation/dehydration via ADT to form phe, or dehydrogenation to form tyrosine. Further work identified and purified distinct ADT enzymes in tobacco suspension cultures and spinach chloroplasts, and also in sorghum etiolated seedlings. It was also determined that PDT was not present in these purified products, pointing further evidence that this alternate pathway of phe and tyr synthesis was in fact the main pathway in plants (Jung *et al.*, 1986, Siehl and Conn, 1988). It was not until 2007 that the entire ADT family was identified and characterized in the plant species *Arabidopsis*. Using enzyme assays, it was determined that most of the ADTs possessed exclusive arogenate substrate affinity, while some had a secondary

smaller affinity for prephenate- verifying that higher plants preferentially use the arogenate pathway, with some ADTs retaining the secondary PDT activity (Figure 1.2) (Cho *et al.*, 2007). Another class of enzymes, known as cyclohexadienyl dehydratases, CDTs, exists in some bacteria, namely *Pseudomonas aeruginosa* and *Erwinia herbicola* (Xia *et al.*, 1991, Zhao *et al.*, 1991). This enzyme groups distinctly from sequences of ADTs and PDTs in phylogenetic analyses (Cho *et al.*, 2007), suggesting it is an artifact of evolution.

While the plastid-localized shikimate pathway is the predominant path in plants, evidence of co-existing plastidic and cytosolic shikimate pathway enzymes has been proposed (Hrazdina and Jensen, 1992). Furthermore PPY-AT, the enzyme that transaminates phenylpyruvate into phe through the PDT pathway, was determined to localize to the cytosol, requiring phenylpyruvate transport out of the chloroplast (Figure 1.2) (Yoo *et al.*, 2013). Therefore, co-existing cytosolic and plastid branches of phe synthesis are possible as well.

1.6 ADT domain architecture and conserved motifs

ADTs have three main domains: the loosely conserved transit peptide (TP) domain, a catalytic ADT/PDT domain, which contains some conserved residues and motifs across all phyla, and the regulatory ACT domain (Cho *et al.*, 2007, Bross *et al.*, 2011). The ACT domain was named after three of the original proteins discovered to contain the domain: aspartokinase, chorismate mutase, and TyrA (prephenate dehydrogenase) (Liberles *et al.*, 2005). The ACT domain is responsible for ligand binding during allosteric-induced feedback regulation, as well as for protein-protein interactions (Liberles *et al.*, 2005, Tan *et al.*, 2008). In bacterial PDT and P-proteins, it has been predicted that the actual active

form of PDTs is in the dimerized state (Tan *et al.*, 2008). In ADTs, PDTs, and P-proteins, the ACT domain binds phe, which alters the dimerization pattern, preventing activity (Zhang *et al.*, 1998, Pohnert *et al.*, 1999). In another study, the first to characterize plant ADT dimerization, it was determined that all AtADTs form homo and heterodimers (Styranko, 2011).

There have been multiple studies conducted, using the *E. coli* P-protein and the PDT of the gram-positive bacterium *Corynebacterium glutamicum,* to determine critical residues and/or motifs associated with catalytic function and regulation (Zhang *et al.*, 2000, Hsu *et al.*, 2004). In the catalytic PDT domain, multiple residues were found that when mutated resulted in significantly reduced catalytic activity (or loss of activity altogether). These include a glutamate-asparagine pair, a serine residue, and also a threonine-arginine-phenyalalanine (TRF) motif (Zhang *et al.*, 2000, Hsu *et al.*, 2004). The same threonine residue was also found to be conserved in cyclohexadienyl dehydratase (CDT), an ancestral dehydratase enzyme found in some gram-negative bacteria that allows the synthesis of phe through both the prephenate and arogenate routes (Zhao *et al.*, 1991). This high degree of conservation suggests that these residues are important for both ADT and PDT activity, potentially suggesting that the mechanisms of ADT and PDT catalysis are similar.

There have also been motifs in the ACT domain that have been determined to be critical for allosteric regulation of the *E. coli* P-protein. Using fluorescence assays to measure phe binding with different portions of the *E.coli* P-protein, it was determined that the phe binding was reserved to a C-terminal area of the protein (Zhang *et al.*, 1998).

Through more fluorometric and microcalorimetric assays targeted towards specific mutants designed via site-directed mutagenesis, Pohnert *et al.* (1999) showed that mutants in the ESRP and GALV hydrophobic motifs either completely lost all sensitivity to feedback inhibition to phe, or had a marked reduction in feedback regulation. This was further demonstrated when a rice mutant encoding an ADT2-like protein with mutations in the ESRP motif displayed increased levels of phe and tyrosine (Yamada *et al.*, 2008). In the GALV hydrophobic region, specifically mutants at the alanine or leucine residues resulted in loss of regulation (Pohnert *et al.*, 1999).

There have also been studies done with the goal of determining critical residues and motifs necessary to allow certain ADTs the ability to have secondary PDT activity. *Arabidopsis* ADTs were studied in a complementation assay to determine regions of importance for possessing secondary PDT activity. Using a yeast *pha2* mutant system, (Bross *et al.*, 2011), in which the yeast strain used is a knockout for the *PHA2* gene that encodes the PDT, a critical residue substitution was discovered at Phe341Leu (Smith, 2014). When this phe residue was substituted to a leucine in AtADT5, the normally ADT-only isoform was able to restore yeast growth in the *pha2* system, implementing a leucine residue in this position as being important for PDT activity (Smith, 2014). This residue position corresponds to a leucine in AtADT2, an ADT/PDT isoform. When AtADT2 was modelled using the known crystal structure of bacterial PDT as a template, it was determined that this leucine residue sits in the alpha-helix of the ACT domain, the site of ADT-dimerization and phe binding for allosteric regulation (Smith, 2014). Furthermore, a region known as the PAC or 'PDT conferring domain', that elicits PDT activity by

significantly increasing the active site's affinity for prephenate, was discovered in the ACT domain (El-Azaz *et al.*, 2016).

1.7 ADTs act at a crossroads

Phe is a primary metabolite product produced from the shikimate pathway. As shown in Figure 1.2, the shikimate pathway is responsible for the production of chorismate, which sits at a crossroads between synthesis of all three aromatic amino acids and the secondary metabolites produced from them. Phe leads to the production of primary metabolites as well as phenylpropanoid compounds. Therefore, given ADT's function in the synthesis of phe, all metabolic flux into the phenylpropanoid pathway is dependent on ADT activity (Maeda and Dudareva, 2012). The phenylpropanoid pathway is a complex metabolic web, producing a vast array of secondary metabolites, all with different roles in the cell, including lignins, tannins, flavonoids, anthocyanins among others (Figure 1.1). Phe is also important for its role in protein synthesis, meaning ADTs are positioned at a critical junction for plant function (Maeda and Dudareva, 2012). ADT activity is regulated through a feedback loop, in which its product, phe, allosterically inhibits ADT from binding its substrate arogenate. Arogenate levels then increase, which trigger a second feed-back mechanism that inhibits 3-Deoxy-D-arabinoheptulosonate 7-phosphate (DAHP)-synthase, an upstream shikimate pathway enzyme. This essentially halts phe synthesis, rendering ADTs essential to the control of phe levels in the cell, such that the needs of individual cell types can be met depending on if they need more protein or phenylpropanoid metabolites (Maeda and Dudareva, 2012).

1.8 Gene duplications and the benefits of large gene families

Whole-genome duplications (WGD) have occurred several times throughout the history of plant evolution, and are thought to have been a prominent force in plant speciation (reviewed in Airoldi and Davies, 2012). WGDs are fundamental for speciation and evolution because they increase the genome size, providing more genetic material that can then undergo further evolution. The duplicated genes can either be lost, gain a new function (neofunctionalization) or the function of the gene that was duplicated can be essentially divided up amongst the new duplicates (subfunctionalization) (Airoldi and Davies, 2012). Most duplicated genes are lost- however, some gene families have higher selection pressures to be retained, as their redundancy is important (Edger and Pires, 2009). For instance, genes encoding enzymes involved in highly interconnected networks are often retained after WGD, as they are usually dosage-dependent (Edger and Pires, 2009). In addition, gene-dosage is important when considering enzymes that act in complexes, as a missing component of a complex jeopardizes the integrity of the entire complex (Edger and Pires, 2009).

In gene families with multiple isoforms, different isoforms have often evolved unique properties and/or roles in the organism. Duplicate gene families provide an evolutionary advantage over singleton genes, as they allow for redundancy in the event of a null mutation (Hanada *et al.*, 2009). With more isoforms present across various locations in the genome, these family members tend to evolve different roles across the cell. For instance, in *Arabidopsis,* it has been shown that ADT family members have unique expression patterns, and different ADTs participate differently in the flux of phenylalanine

towards the synthesis of various phenylpropanoid metabolites (Corea *et al.*, 2012a). *AtADT4* and *AtADT5* are both upregulated in lignin synthesis (Corea *et al.*, 2012b), whereas *AtADT2* has the highest contribution to anthocyanin production (Chen *et al.*, 2016a).

Soybean is a paleopolyploid organism that has undergone two WGD and a variety of tandem duplication events (Schmutz *et al.*, 2010). Many of the phenylpropanoid genes in soybean are present in large families where members of the gene families have different tissue-specific expression patterns, including *GmPT*, *GmCHR*, *GmCHI*, *GmCHS* and *GmIFS* (Todd and Vodkin, 1996, Subramanian *et al.*, 2005, Dastmalchi and Dhaubhadel, 2015, Sepiol *et al.*, 2017, Sukumaran *et al.*, 2018). *ADTs* are also present in large gene families in other plant species studied thus far as well (Cho *et al.* 2007, Maeda *et al.* 2010, El-Azaz, *et al.* 2016).

This pattern of large gene families underlies the importance of the phenylpropanoid metabolites' role in plant response to biotic and abiotic stresses, as the multiple copies suggest a functional compensation role (discussed previously). Because ADT lies in close proximity to the phenylpropanoid pathway, and given the large ADT family sizes present in other organisms studied, it is likely that soybean also has multiple members that may have undergone their own unique neofunctionalizations.

1.9 Hypothesis

Given the paleopolyploidy of soybean, and the tendency of phenylpropanoid-related gene families to have many members, I hypothesize that the soybean genome

will have multiple *ADT* genes, and the *ADT* gene products will interact with GmIFS.

1.10 Objectives

The overall objective of my research is to identify and characterize all *ADT* gene-family members present in the soybean genome, and to validate the IFS-ADT interaction *in-planta.*

1. Identify all *ADT* family members in soybean (*GmADTs)*

- To determine if soybean *ADTs* have a large family size

- To compare the soybean *ADT* gene structures to those present in other organisms

- To study their phylogeny, so predictions about the evolutionary history of *GmADTs* can be inferred

2. Determine the subcellular localization of GmADTs

- To confirm if GmADTs have similar localizations to ADTs in other organisms

- To determine if GmADT-GmIFS interaction is possible through GmADT potentially localizing to the cytosol

3. Confirm the GmADT-GmIFS interaction

- BiFC confirms if protein interactions discovered via other assays are occurring *in-planta* and it shows where the interaction is occurring

4. Investigate *GmADT* expression patterns: tissue-specific and in response to stresses in order to determine if *GmADT* expression is induced by similar conditions as isoflavonoid biosynthesis genes.

Chapter 2: Materials and Methods

2.1 Plant materials and growth conditions

Nicotiana benthamiana were grown on PRO-MIX® BX MYCORRHIZAETM soil (Rivière-du-Loup, Canada) in a growth chamber. The growth chamber was set to a 16 hour light cycle at 24°C, and 8 hour dark cycle, with 60% relative humidity and a light intensity of 57-78 µmol m^{-2}s^{-1}. *N. benthamiana* seeds were cast onto the surface of wet soil and sprayed with water daily for a week to two weeks until germination. At this stage, the seedlings were transferred to sterilized pots and grown in the same PRO-MIX® BX MYCORRHIZAETM soil at the aforementioned conditions (with regular watering). Fertilization was done along with watering, using a nitrogen-phosphorous-potassium (20-8-20) fertilizer at a concentration of 0.5 g/L.

2.2 Bacterial strains

Eschericia coli strain DH5α was used for cloning *GmADTs* into the Gateway vectors. *Agrobacterium tumefaciens* strain GV3101 was used to transiently express GmADT-YFP fusion proteins in *N. benthamiana*.

2.3 *In silico* and phylogenetic analysis

Candidate *GmADTs* were identified by mining the *Glycine max* genome in Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax). To begin, keyword searches using the following entries were used to search for putative *ADTs*: arogenate dehydratase, prephenate dehydratase, ADT, and PDT (as the sequence similarity between ADTs and PDTs could have caused some ADTs to be mislabeled). Then

the two previously identified soybean ADT genes (Dastmalchi *et al.*, 2016), (*Glyma.13G319000.1* and *Glyma.12G181800.1*) were used in BLAST searches of the soybean genome, to narrow down the original list compiled from the keyword search. Each unique gene identified from the two initial input sequences was used in BLAST searches to look for all possible *ADTs.* Additionally, the protein sequences of the candidate *ADTs* were aligned using ClustalOmega (https://www.ebi.ac.uk/Tools/msa/clustalo/) to look for areas of high conservation, and a pairwise comparison of both the coding sequences and amino acid sequences to determine percent identity was performed on the aligned sequences. Each of the protein sequences were input into the TargetP subcellular localization prediction software, selecting for "plant" sequences. TargetP uses a cut-off scoring matrix of 0.5, where sequences with scores of 0.5≤ are assigned a chloroplast localization. (Emanuelsson *et al.*, 2007).

Before constructing a phylogenetic tree, the transit peptides of the plant ADTs were removed according to alignments of the catalytic domains of the *E. coli* and *Salmonella enterica* P-proteins. The chorismate mutase domains of the *E. coli* and *Salmonella enterica* P-proteins were removed according to Zhang *et al* 1998**.** A neighbour-joining tree was constructed using MEGA7, using default settings (Kumar *et al.*, 2016).

2.4 Promoter sequence analysis

To identify promoter elements in the candidate genes, the region 1000 bp upstream of the translational start site of each candidate *GmADT* was acquired from the Phytozome(https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax)

database. The upstream promoter sequences (as shown in Appendix F) were input into the Plant cis-acting Regulatory DNA Elements (PLACE) (Higo *et al.*, 1999). The results were mined manually to analyze promoter element landscapes and sorted into categories.

2.5 RNA-seq analysis

2.5.1 Tissue-specific expression

Soybean RNA-seq data, from the NCBI database (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29163), was retrieved from Phytozome in the form of fragments per kilobase per million mapped reads (FPKM) (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax). These values were normalized and a heatmap was constructed using R to compare the transcript levels of *GmADTs* across different tissues.

2.5.2 Expression in response to abiotic stress

Previously published, publicly available RNAseq data (Accession: PRJNA324522, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA324522) was downloaded from a study that measured the gene expression changes in soybean leaf tissue treated with drought and flood stresses (Chen *et al.*, 2016b). The RNAseq data was downloaded as .sra, and was converted to .fasta files using the fastq-dump feature in the sequence read archive (SRA) toolkit. The soybean reference transcripts (in the form of cDNA) was downloaded from Phytozome. From here, Tophat (Kim *et al.*, 2013) was used to align the SRA data to the reference soybean CDS. The 'idxstats' feature of the Samtools (Li *et al.*, 2009) set of utilities was then used to generate the RNA read counts and respective contig (i.e. gene) lengths. Using the read counts and the gene lengths, FPKM was calculated

manually. A heatmap was constructed using 'R', in which the expression values are scaled

for analysis within each gene.

2.6 Reverse transcription-polymerase chain reaction (RT-PCR)

DNase1-treated RNA (1 µg) was used to synthesize cDNA from RNA that had

previously been extracted from soybean hypocotyls using the ThermoScript™ RT-PCR

System (Life Technologies). For the PCR, reactions were set-up using gene-specific primers

(Table 2.1). *CONS4* was used as a loading control.

2.7 Gene cloning

To isolate the *GmADTs,* primers were designed in the 5' and 3' untranslated

regions (UTR) of the genes, to prevent amplifying multiple genes, as there is a high level

of sequence conservation among members of the same gene family. The primers used

are shown in Table 2.1. After PCR amplification of the genes with gene-specific UTR

primers, the PCR products were run on a RedSafe (iNtROn Biotechnology) stained 1%

agarose gel in 0.5X TBE buffer and then visualized on a Bio-Rad Gel Doc imager. The

amplicons (corresponding to the expected sizes) were extracted using EZ-10 Spin Column

DNA Gel Extraction Kit (Bio Basic Inc). Purified products were then used as templates in

nested PCRs using primers designed to amplify the complete CDS without the stop codon.

These primers were designed to be Gateway cloning-compatible, with the *attB1* adaptor

sequence (5'-GGGGACAAGTTTGTACAAAAAAGCAGGCT-3') for the forward primer and the

attB2 adaptor sequence (5'-GGGG AC CAC TTT GTA CAA GAA AGC TGG GT-3') for reverse

primer. After these nested PCRs were completed, the same steps for purifying the

amplicons were taken as above, followed by DNA quantification using the NanoDrop 1000

spectrophotometer (ThermoScientific). The amplicons were then recombined into pDONRZeo using the Gateway BP Clonase® II Enzyme mix (Invitrogen). The BP reactions were then transformed into *E. coli* strain DH5α via electroporation and then plated on LB-agar plates containing zeocin (50 µg/mL) for selection. Transformation success was determined via colony PCR using the gene-specific gateway primers. These positive colonies were grown overnight in liquid media containing zeocin and the plasmid DNA was extracted using the EZ-10 Spin Column Plasmid DNA kit (Bio Basic Inc.), quantified using NanoDrop 1000 spectrophotometer (ThermoScientific), and their cloned sequence confirmed.

For subcellular localization studies, the pDONRZeo-GmADT plasmids were recombined into the pEarleyGate101 (pEG101) destination vector using Gateway LR Clonase® II Enzyme mix (Invitrogen), to form expression clones in which the *GmADTs* were translationally fused to the full YFP sequence. *GmADT13B* was also recombined into a modified pEG101 vector containing both YFP and the mCherry fluorescent protein, which adds an additional 28.8 kDa to the translational fusion protein. The LR reactions were then transformed again into *E. coli* strain DH5α via electroporation, and plated on LB-agar plates containing kanamycin (50 µg/mL) for selection. Transformation success was determined via colony PCR using the gene-specific primers. The plasmid DNA was again extracted using the EZ-10 Spin Column Plasmid DNA kit (Bio Basic Inc.) and quantified using NanoDrop 1000 spectrophotometer (ThermoScientific).

Table 2.1| List of primers used for Gateway-cloning of GmADTs

Gene Name	Primer Name	Sequence (5'-3')	Amplicon size (bp)	Purpose
GmADT11A	GmADT11AF	ATG CAG ACC CTC AAT CAA CCG	1284	External UTR primers for nested PCR
	GmADT11AR	TCGGCAATATGTTAATTTTGGCGCGG		
GmADT12A	GmADT12AF	CCA TAA TAT GCA GAC TCT TTC GCC	1347	
	GmADT12AR	TCCTCTCTTTTGGGAGGGAGAGATG		
GmADT12B	GmADT12BF	CAACTAAATTCCCCTTTCCAACC	1370	
	GmADT12BR	TCATGAAAGAAATGGAGGTGGATG		
GmADT12C	GmADT12CF	TTGAGAACCGTTGACCTCC	1279	
	GmADT12CR	TCTATTTGGACATGAAGGTAGCTGC		
GmADT12D	GmADT12DF	CCAAACACTGTCTCCGTCTTGATG	1001	
	GmADT12DR	TCACTCTGATCAGCCATTGATGTTTC		
GmADT13A	GmADT13AF	ATTCCTCTGTCAAGCCACTCG	1322	
	GmADT13AR	TCCAAGAGGGGAAAAAGACGATGC		
GmADT13B	GmADT13BF	GTACTTTTGGTCCAAGCGGTT	1225	
	GmADT13BR	TCAGATGAACTAATGGCACTGTCTAAAGGTC		
GmADT17A	GmADT17AF	TTCATTTTGATGGCTCTTAAGGCTG	1247	
	GmADT17AR	TCCAGCAAAATGAACAGCATGACT		
GmADTU4	GmADTU4F	AAACCCAAACACTGTCTCCGTCTGG	1204	
	GmADTU4R	TCCGATAATCTTCAAAAGTGGACTCCCG		

Table 2.1 continued| List of primers used for Gateway-cloning of GmADTs

Gene Name	Primer Name	Sequence (5'-3')	Amplicon size (bp)	Purpose
GmADT11A	GmGTADT11AF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGCAGACCCTCAATCAA	1287	Coding-sequence specific with Gateway attachment sites
	GmGTADT11AR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCATTTTGGCGCGGACAA		
GmADT12A	GmGTADT12AF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGCAGACTCTTTCGCC	1278	
	GmGTADT12AR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCTTTAAATTTATCTCCCCGGGAGG		
GmADT12B	GmGTADT12BF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGCAGACCCTCACCC	1287	
	GmGTADT12BR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCATTTTGGCGCGGAGAAGA		
GmADT12C	GmGTADT12CF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGGCTGTGACATCACCTCTTG	1155	
	GmGTADT12CR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCTATGGTTGTATCTATGGGATAGCAG		
GmADT12D	GmGTADT12DF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGGCTGCGTCGCGAATC	933	
	GmGTADT12DR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCTACCTTTGTAAGGTTAATCTGACGC		
GmADT13A	GmGTADT13AF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGCAGAGTCTTTCACCACC	1272	
	GmGTADT13AR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCGTCTCCCCGGGAGGAA		
GmADT13B	GmGTADT13BF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGCGTGTGGTTGATCATCCT	963	
	GmGTADT13BR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCCCATACCCGAAGAAATGTGGC		
GmADT17A	GmGTADT17AF	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGGCTCTTAAGGCTGTATC	1209	
	GmGTADT17AR	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCGTTAAGACACTGAACTTCTATAATACT		
GmADTU4	GmGTADTU4F	GGGGACAAGTTTGTAC AAA AAA GCA GGC Ttc ATGGCGGCATCGCGAATCGTG	1158	
	GmGTADTU4R	G GGG ACC ACT TTG TAC AAG AAA GCT GGG TCCGTCAAGCTAGTGTCCACAGGATA		

The pEG101-GmADT plasmids were then transformed into *A. tumefaciens* GV3101 via electroporation. The transformed *Agrobacterium* was plated on a LB agar plates containing rifampicin (10 μg/mL), gentamycin (50 μg/mL), and kanamycin (50 μg/mL) for selection. Transformation success was determined via colony PCR using the gene-specific gateway primers. For BiFC studies, the pDONRZeo-GmADT plasmids were recombined into pEG201-YN and pEG202-YC Gateway LR Clonase® II Enzyme mix (Invitrogen), to form expression clones in which the GmADTs were translationally fused to the N-terminal YFP and C-terminal YFP sequences, respectively. The LR reactions were then transformed into *E. coli* strain DH5α following the method described for cloning into pEarleyGate101, to obtain pEG201-GmADT and pEG202GmADT. The pEG201-GmADT and pEG202-GmADT plasmids were then transformed into *A. tumefaciens* GV3101 as described above.

2.8 Plant infiltration

Single *A. tumefaciens* colonies containing different pEG101-GmADT constructs were grown separately at 28°C in infiltration media - LB broth containing 10 mM 2-N-morpholino-ethanesulfonic acid (MES) pH 5.6, 100 μM acetosyringonel, kanamycin (50 μg/mL), rifampicin (10 μg/mL), and gentamycin (50 μg/mL) - until an OD_{600} of 0.9-1.2 was reached. The *A. tumefaciens* cultures were spun at 775 *g* for 30 minutes, and the pellets were resuspended in Gamborg's solution - 3.2 g/L Gamborg's B5 vitamin mix, 20 g/L sucrose, 10 mM MES pH 5.6, 200 μM acetosyringone and milliqu water – to a final OD_{600} of 1.0. To activate the *Agrobacterium* virulence genes the suspensions were incubated at room temperature with gentle agitation for one hour. *N. benthamiana* leaves were

infiltrated with the bacterial suspensions using a syringe to gently infiltrate the suspension into the leaves via the underside. The plants were returned to the growth chamber at the conditions described in section 2.1.

For BiFC, equal volumes of the *Agrobacterium* cultures containing corresponding interaction partners were thoroughly mixed before being co-infiltrated into *N. benthamiana* leaves. Infiltration was performed as described above.

2.9 Confocal microscopy

At 48 hours post-infiltration, *N. benthamiana* leaves were visualized using the OlympyusFV1000 confocal microscope using a 60X water immersion objective lens. For YFP visualization, the excitation wavelength was set to 514 nm and emission was collected at 520-550 nm. For chloroplast-visualization, the natural autofluorescence produced by the chlorophyll was harnessed by exciting the chlorophyll at 600 nm and emission was collected at 640-700 nm.

Chapter 3: Results

3.1 The soybean *ADT* gene family contains 9 members

Identification of *GmADTs* was done using a keyword search with the following search terms: "ADT, PDT, arogenate dehydratase". These searches retrieved a cumulative 183 unique genes; many of the genes in this list were actually not arogenate or prephenate dehydratases, making this method too non-specific. In order to narrow down this list to sequences that actually showed sequence similarity to the two ADTs pulled-down in a previous co-IP experiment, *Glyma.13G319000 and Glyma.12G181800* (Dastmalchi *et al.*, 2016), these two ADT sequences were used in BLAST searches of the soybean genome. The *Glyma.13G319000* BLAST search yielded two putative *ADT*s, and the *Glyma.12G181800* BLAST search yielded nine putative *ADT*s, two of which overlapped with the *ADT*s retrieved from *Glyma.13G319000* BLAST search. To exhaust the database, each of these nine *ADT* sequences were used in their own BLAST searches, which ultimately yielded two additional *ADT* sequences, bringing the total list of putative *ADT*s to 11 candidates. The peptide sequences encoded by the 11 putative *GmADT* genes were aligned with *Arabidopsis* ADTs to look for regions of high conservation. One candidate *GmADT*, *Glyma.19G053000*, was missing large portions of sequence (see Appendix A), and also had FPKM expression values of zero, indicating it is not expressed (Table 3.1). Thus it was removed from the list of candidate *GmADTs*. Another of these 11 candidate genes, *Glyma.09G004200*, showed very low expression values (ranging in value from 0-0.103 FPKM) in published RNAseq data on Phytozome (see Table 3.1).

Multiple attempts to clone this gene using cDNA prepared from RNA isolated from different soybean tissues were not successful. Therefore, *Glyma.09G004200* was also removed from the list of candidate genes in this study, leaving a final list of 9 *GmADTs.*

The list of *GmADT* family members, with their coding sequence length, number of splice variants, and protein molecular weight is shown in Table 3.2. The majority of ADTs are predicted to localize to the chloroplast, with the exception of GmADT12B, GmADT17A and GmADT13B, which don't have strong enough prediction values for any specific subcellular location, and GmADT12C, which is predicted to localize to the mitochondrion. The coding sequence percent identity comparisons range from 56% between *GmADT11A and GmADT13B* to 96% identity between *GmADT13B and GmADT12C* (Table 3.3)*.* The amino acid percent identity of the mature proteins range from 60% between GmADT12B and GmADT13B to 99% between GmADT12A and GmADT13A (Table 3.3).

Table 3.1| Published tissue-specific *GmADT* FPKM expression values from Phytozome

Gene Name	Gene Locus	Flower	Leaves	Nodules	Pod	Root	Root Hairs	Seed	SAM	Stem
GmADT11A	Glyma.11G189100	2.86	0.83	1.26	2.46	1.66	1.87	2.13	1.70	1.95
GmADT12A	Glyma.12G181800	15.25	33.84	10.14	12.16	23.08	12.50	5.73	7.58	20.85
GmADT12B	Glyma.12G085500	28.77	4.75	4.35	9.69	11.95	4.78	3.07	3.16	13.12
GmADT12C	Glyma.12G193000	20.23	53.03	9.55	16.69	12.36	5.14	5.44	10.56	14.50
GmADT12D	Glyma.12G072500	4.90	6.68	4.77	4.86	1.27	4.60	12.22	5.50	4.01
GmADT13A	Glyma.13G319000	17.78	22.88	14.272	14.04	17.46	14.49	5.27	17.63	24.73
GmADT13B	Glyma.13G309300	1.57	2.50	0.69	1.48	0.47	0.08	0.95	1.66	0.53
GmADT17A	Glyma.17G012600	0.70	0.53	4.44	1.77	5.50	1.88	2.37	1.94	1.51
GmADTU4	Glyma.U021400	11.35	17.98	12.86	23.87	9.02	17.30	29.69	14.79	19.22
--------	Glyma.09G004200	0	0.04	0	0.06	0.33	0.03	0	0.10	0
--------	Glyma.19G053000	0	0	0	0	0	0	0	0	0

Data collected from:
https://phytozome.jgi.doe.gov/jbrowse/index.html?data=genomes%2FGmax&loc=Chr09%3A330689..338585&tracks=Transcripts%2CAlt_Transcripts%2CPASA_assembly%2CBlastx_protein%2CBlatx_FabidaeGrape%2CTranscriptExpression&highlight=).
SAM: shoot apical meristem

Table 3.2|Characteristics of putative *arogenate dehydratase* gene family members in soybean.

Gene Name	Locus Name*	Locus range	Predicted Mol. Weight (kDa)**	Coding Sequence Length (bp)	Splice Variants	Predicted Subcellular Localization (TargetP)
GmADT11A	Glyma.11G189100	Chr11:26084890..26087312	32.94	1287	1	Chloroplast
GmADT12A	Glyma.12G181800	Chr12:34248924..34251406	32.75	1278	1	Chloroplast
GmADT12B	Glyma.12G085500	Chr12:6895173..6896459	32.85	1287	1	Any other location
GmADT12C	Glyma.12G193000	Chr12:35450459..35456613	31.80	1155	1	Mitochondrion
GmADT12D	Glyma.12G072500	Chr12:5344610..5350276	24.21	933	3	Chloroplast
GmADT13A	Glyma.13G319000	Chr13:41357376..41359383	32.72	1275	1	Chloroplast
GmADT13B	Glyma.13G309300	Chr13:40509929..40514843	35.13	963	1	Any other location
GmADT17A	Glyma.17G012600	Chr17:970156..977601	31.85	1209	3	Any other location
GmADTU4	Glyma.U021400	scaffold_21:3463451..3468589	31.70	1158	2	Chloroplast

*Locus name is respective to the Phytozome categorization.

**Molecular weight is of mature protein.

Table 3.3|Pairwise coding region and amino acid sequence comparisons of the putative soybean arogenate dehydratase family

	GmADT13A	GmADT13B	GmADT12A	GmADT12B	GmADT12C	GmADT12D	GmADT17A	GmADTU4	GmADT11A
									Amino Acid**
GmADT13A		61.65	99.32	92.23	65.60	70.45	62.98	69.01	91.89
GmADT13B	60.39		60.88	60.00	96.50	63.18	65.05	61.97	60.68
GmADT12A	95.06	60.05		91.89	64.89	70.00	62.28	68.66	91.55
GmADT12B	84.29	59.70	84.85		64.18	70.00	64.36	67.61	96.63
GmADT12C	60.39	96.96	60.05	59.70		63.34	63.82	61.86	60.47
GmADT12D	59.58	61.54	59.43	59.58	61.69		68.18	97.27	69.09
GmADT17A	57.59	66.32	57.24	57.82	66.20	65.76		62.32	63.32
GmADTU4	60.23	61.29	59.57	59.88	62.82	95.93	63.27		66.9
GmADT11A	84.51	56.19	84.40	94.52	58.65	59.73	57.59	60.47	
Nucleotide*									

*nucleotide percent identities were calculated using the coding sequence corresponding to the mature protein

** amino acid percent identities were calculated using mature protein sequences

3.2 *GmADTs* share conserved motifs with *Arabidopsis ADTs* and other organisms

To determine if the putative GmADTs contain the conserved motifs necessary for ADT enzymatic activity, a multiple sequence alignment of GmADTs with AtADTs was performed. As shown in Figure 3.1, all nine GmADTs contained the three domains present in the well-characterized *Arabidopsis* ADTs: The N-terminal transit peptide, the catalytic domain, and the regulatory C-terminal ACT domain. The transit peptides were highly variable, with GmADT13B lacking significant portions of this domain. There was a high degree of conservation across most motifs in the catalytic and ACT domains (Figure 3.1). In bacterial PDTs and P-proteins, there have been multiple site-directed mutagenesis studies conducted to determine which of the conserved residues are important for PDT activity. Hsu *et al.* (2004), determined the following residues, when mutated, significantly reduce PDT activity: Glu59(Glu64), Thr177(Thr183), Arg178(Arg184) and Phe179(Phe185) [residue numbers in reference to the mature *E. coli* P-protein shown in the alignment in Appendix A; in brackets are the residue locations according to the *C. glutamicum* PDT used in the study (Hsu *et al.*, 2004)]. Likewise, Zhang *et al.* (2000) determined the following residues to be important for catalytic activity: Asn60(Asn160), Ser107(Ser208), Gln114(Gln215) and Thr177(Thr278) [residue numbers in reference to the mature *E. coli* P-protein shown in the alignment in Appendix A; in brackets are the residue locations according to the *E. coli* P-protein used in the study (Zhang *et al.*, 2000)]. Additionally, Tan *et al.* (2008) determined the Thr177-Arg178-Phe179 motif is located at the active site of the PDT protein, making them critical for catalytic activity. Furthermore, Tan *et al.* (2008) identified a number of individual residues involved in hydrogen-bonding to the TRF active

site, including Glu59, Asn60, Gln114 and Asn175 (as marked by asterisks (*) in Appendix A). In order to look for these important residues in GmADTs, a second multiple sequence alignment was performed using additional characterized *Pinus pinaster* ADT sequences. The result revealed that Glu59 and Asn60 are conserved across all PDTs and ADTs looked at so far, along with Ser107 and Arg178. Interestingly, GmADT13B and GmADT12C both had Gln114Leu and Thr183Ser substitutions (shown in Appendix A). GmADT13B and GmADT12C also had substitutions at Gln114, Asn175 and Phe179, along with GmADT17A, which had a substitution at Asn60.

There have also been multiple studies that have determined critical residues in the regulatory ACT domain. Two main motifs, the ESRP and GALV motifs, were previously determined to be critical for regulatory function via site-directed mutagenesis in the *E.coli* PDT P-protein (Pohnert *et al.,* 1999) (Appendix A). When plant sequences were aligned with bacterial PDT sequences, the full ESRP domain was almost entirely conserved across every organism included, except for PpADT-H, that was missing the 'SR', and notably, GmADT12D appears to be missing this motif altogether (Appendix A). The GALV motif was less conserved, and none of the plant sequences (i.e. ADTs) had the alanine (only the *E. coli* P-protein had the alanine in this motif). The leucine residue from this motif was conserved across all sequences, both bacterial and plant, with many plant sequences having a "GVLF" motif (Appendix A).

Figure 3.1| Multiple sequence alignment of GmADTs and AtADT. The amino acid sequences of the GmADTs and Arabidopsis ADTs (AtADT) were aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/). The alignment was visualized using the boxshade server (https://embnet.vital-it.ch/software/BOX_form.html). Black shading indicates identical residues and grey shading indicates similar amino acid residues. Dashed lines indicate gaps. The boxes of different shades of blue indicate the corresponding ADT domains identified in AtADTs as shown in the diagram on the top: ▭ , transit peptide domain; ▭ , catalytic ADT domain; and ▭ , regulatory ACT domain (adapted from Cho, *et al.* 2007).

H₃N⁺ — **TP** — **ADT** — **ACT** — COO⁻

$$\text{H}_3\text{N}^+ - \boxed{\text{TP}} - \boxed{\text{ADT}} - \boxed{\text{ACT}} - \text{COO}^-$$

```
AtADT4    -----Y------QAATSCDLKFRS-TDPTSRNKCFSHAIPKRVAVTCGYRSESFS-PNG-SVSRSDWQSSCAILSSKVASVENTG--------GLADK-AAVNGHTNGS-   89
AtADT5    --MQT------SPAFSCDLKSVIQPNLTAKKARYSHVNGKRVSVRCSYRSESFS-PNG-GSSRADWQSSCAILASKVVSAENSS------------SVAVNGHSNGS-   89
GmADT13A  --MQS------SPPTSNALNLK-----HVF-RPRLGASSRISVKCAFGFEPVS--G-GSSRADWQSSCAILASKVVSQEQP-SA--DQNSGTTDH-AAVNGHKA---   87
GmADT12A  --MQT------SPPTSNALNLK-----HVL--RPRRVAPSRIFVKCAFASEPASM-G-GSSRADWQSSCAILASKVVSQEQT-SA--DQNGGTADH-AAVNGHKA---   88
GmADT11A  --MQT------NQPCPNPLNYL-----TRARHSAIRFAPTRFTVKCGYGFESAS-SQG-GASRADWQSSCAILASKVVSQREDSPA---HGGDNNNN-GAVNGHNA---   91
GmADT12B  --MQT------TPPYPNPLNYL-----TRARHSVIRVAPTRFTVKCGYGFESAS-SQG-GASRADWQSSCAILASKVVSQREDSPA---DGDDHDNN-GDVNGHNA---   91
AtADT3    --MRT------LPSHTPA-TVT-----TAAR--------RRHVIHC-AGK--RSDSFS-NSSSSDWQSSCAILSKVNSQEQSESLSSNSNGSSSYHVSAVNGHNNG-A   84
AtADT6    --MKA------SSS-SPI-L---------------------GAS-QPA--TATALIARSGRSEWQSSCAILSKVISQEESESLPVPPVSGG--VDHLNGHNSAAA     71
AtADT2    MAAHTVR----LSPATQLHGGISSN-------LSPPNRKPNNSI---------VRMGCGSSKRFRIVTVLA--------SLREN---------------DANGRD     62
GmADT12D  MAASR-V----AHPPPYLHR-----------QSPPSDAAPSLN--------L---TL-PKRYRNLGICA--------SLRG------------------------     47
GmADTU4   MAASR-V----AHPPPYPHR-----------QSPPSDAPPSLN--------L---TLNPKRYRNLGIRA--------SLRGH------------------EEKNKN   54
GmADT13B  ------------------------------------------------------------------------------------------------------------    0
GmADT12C  MA-------------------V-------VAG---------SQLRCHNFRRTRHMALILS--WEKQQLKRAQKCG--------------GDENPWKP         51
AtADT1    MAFRCFPIWVCPQTTH--------H-------RSPL------MGL---------AEPDADKRRRFCLWECS-----SSASQRAVTAI----------EGEIPFSR   60
GmADT17   MATKAVSIWGCYKPPQLGVGVSNS-------HSTL------IGN--------LRHDYDKCR---RWECCCL----GVLAQRATTAV-----------EDEGPSVP   67

AtADT4    --VNT---GVVAVES-TN------GKLAPAQPLTITDLSPAPLHGSSLRVAYQGVPGAYSEAAAGKAYPNCDAIPCDQFLVAFQAVELWIADRAVLPVENSLGGSIHRNY   187
AtADT5    --VDT---SVVESKSQHN------GKPGLIQPLTITDLSPAPSHGSTLRVAYQGVPGAYSEAAAGKAYPNSEAIPCDQFLVAFQAVELWIADRAVLPVENSLGGSIHRNY   188
GmADT13A  AVSDF---QVVEIGNLAEANN----KPLPPKPLTIDDLSPAPMHGSKLRVAYQGVPGAYSEAAAGKAYPNGEAIPCDQFEVAFQAVELWIADRAVLPVENSLGGSIHRNY   190
GmADT12A  AVSDT---QVVEIGNLAQANN----KPLPPKPLTISDLSPAPMHGSKLRVAYQGVPGAYSEAAAGKAYPNGEAIPCDQFEVAFQAVELWIADRAVMPVENSLGGSIHRNY   191
GmADT11A  AVTNI---NIVEVKA-DDENI----KPIPPKPLTISDLLPAPMHGSQLRVAYQGVPGAYSEAAAGKAYPNGEAIPCDQFEVAFQAVELWIADRAVLPVENSFGGSIHRNY   193
GmADT12B  AVTNI---NIVEVKA-DGENI----KPIPPKPLTISDLLPAPMHGSQLRVAYQGVPGAYSEAAAGKAYPNGEAIPCDQFEVAFQAVELWIADRAVLPVENSFGGSIHRNY   193
AtADT3    GVSDL---NIVEFNNN---------QSIQSKKPLSISDLSPAPMHGSNLRVAYQGVPGAYSEAAAGKAYPNCQAIPCDQFLVAFQAVELWVDRAVLPVENSLGGSIHRNY   183
AtADT6    RVPGV---NIVPIEKSDSNPLVPQHRHNPLKPLSMTDLSPAPMHGSNLRVAYQGVPGAYSEAAAGKAYPNCQAIPCDQFEVAFQAVELWIADRAVLPVENSLGGSIHRNY   178
AtADT2    N------SV-VRA--MEVKKIFEDSPLLPKPLSSNQLTESVSNGSRVRVAYQGVPGAYSEAAAKAYPNCEAVPCEEFDTAFEAVERWLVDRAVLPIENSLGGSIHRNY   161
GmADT12D  ---DKPHSVEL-RAT-TSDVVVSRDLLSLPRPLSSTQLSASVSGSCLRVAYQGVPGAYSEAAAKAYPNCEAVPCEQFLTAFDAVERWHVAYHKLQDAGAVASSAAAIYGLNILDQDIQDDSDNVTR  152
GmADTU4   DLZADKPHSVEL-RATSTPDDVVSRDLLSLPRPLSSTQLSASVSDSRLRVAYQGVPGAYSEAAAKAYPNCEAVPCEQFLTAFDAVERWHVAYHKLQDAGAVASSAAAIYGLNILDQDIQDDSDNVTR  163
GmADT13B  ----M-RVVDHEPRDGDVSYGLHKDLVSLPKPLSISDIVAASDDHAKVRISYKGIPGSYSBDAALKAYPNCETVSQNDFEPAFRAVEIWMADKNVILPIENTSGGSIHRNY   105
GmADT12C  VKIVM-RVVDHHPRDGDVSYGLHKDLVSLPKPLSISDIVAASDDHAKVRISYKGIPGSYSBDAALKAYPNCETVSQNDFEPAFRAVEIWMADKNVILPIENTSGGSIHRNY   160
AtADT1    EIKK--SSDELGLTQETQSLSFHRDLSMLPKPLTANSLYSSDGDDSKVRISEQGIPGAYSETAALKAFPNCETVPCEQFPAAFDAVELWIVDRAVLPIENSVGGSIHRNY   168
GmADT17   PLVD--SSGAADGVHQNESKGFHKDLNLLPKPLTAIDLTSSYERDGSKVRVAYQGLPGAYSBDAALKAYPRCETVPCDNFEPAFRAVELWLVNRTVLPIENSVGGSVHRNY   175

AtADT4    DLLLLRHRLHIVGEVQIPVHHCLLALPGVRTDCVSRVISHPQALAQTEHSLDVLTPHAAREAFHDTAAAAEYISANDLHDTAAVASARAAELYNLQILADGIQDDPGNVTR   297
AtADT5    DLLLLRHRLHIVGEVQIPVHHCLLALPGVRTECVSINKLTPKAALTEAFHDTAAAAEYIAANNLHDTAAVASARAAELYGLNVMADGIQDDGNVTR               298
GmADT13A  DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLTRVISHPQALAQCEHTLTKLGINVAREAVDDTAGAAEFVASNNLRDTAAIASARAAELYGLNVMADGIQDDPSNVTR   300
GmADT12A  DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLTRVISHPQALAQCEHTLTKLGINVAREAVDDTAGAAEFVAANNLRDTAAIASARAAELYGLNVMADGIQDDPSNVTR   301
GmADT11A  DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLERVISHPQALAQCEHSLTKLGLNTVAREAVDDTAGAAEFVARNDLRSTAAIASARAAELYGLQILADGIQDDPNNVTR   303
GmADT12B  DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLTRVISHPQALAQCEHSLTKLGLTVTREAVDDTAGAAEFVAANDLRDTAAIASARAAELYGLQILADGIQDDPNNVTR   303
AtADT3    DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLTRVISHPQGLAQCEHTLTKLGLNVAREAVDDTAGAAEFIAANNIRDTAAIASARAAEIYGLEILEDGIQDDASNVTR   293
AtADT6    DLLLLRHRLHIVGEVQLPVHHCLLALPGVRKEFLTRVISHPQGLAQCEHTLTKLGLNVAREAVDDTAGAAEFIASNNLRDTAAIASARAAEIYGLEILEDGIQDDVSNVTR   288
AtADT2    DLLLLRHNLHIVGEVKLAVRHCLLANHGVNIEDLRRVISHPQALAQCENTLTKLGI--VREAVDDTAGAAKQLAFENLNDAAAVASEKAARIYGLNIVAKDIQDDCDNVTR   269
GmADT12D  DLLLLRHSLHIVGEVNFAVCHCLMANHGVKREDLKRVLSHPQALAQCENTLTKFGI--VREAVDDTAGAAKHVAYHKLQDAGAVASSAAAIYGLNILDQDIQDDSDNVTR   260
GmADTU4   DLLLLRHSLHIVGEVNFAVRHCLMANHGVKREDIKRVLSHPQALAQCENTLTKFGI--VREAVDDTAGAAKHVAYHKLQDAGAVASSAAAIYGLNILDQDIQDDSDNVTR   271
GmADT13B  DLLLCHRLHIVGEVQLATNLALLALPGVRTEILKRVLSHSQAFELSDDFLTKLGV--ARENVDDTAGAAQILASNGLYDAGAIASIRAAEIYGLNVLADRIQDDSEIISR   213
GmADT12C  DLLLLRHRLHIVGEVQLATNLALLALPGVRTEILKRVLSHSQAFELSNDFLTKLGV--ARENVDDTAGAAQILASNGLYDAGAIASIRAAEIYGLNVLADRIQDDSEIISR   268
AtADT1    DLLLLRHRLHIVQEVHLPVNHCLLGVPGVKKEDIKCVISHPQALDQCVNSLNNLGI--QRISAKDTATAAQTVSSSGKIDVGAIASVRAANIYGLDILADPIQDDVNNVTR   276
GmADT17   DLLLLRHRLHIVGEVQLRVNHCLLGLPGVRKEIRAVVSHPQAFAQCETTLSDLGA--VRIAARDTAAAAQTVASNCARDTGAIASSRAAEVYGLDILADPRIQDDDENITR   283

AtADT4    FLMLAREPIIPRTDRPFKTSIVFAAQEHKGTSVLFKVLSAFAFRDISLTKIESRHHNRPLRVVGDGSFGTSKNFEYIFYVDFEASMAEFRAQNALAEVQEYTSFLRVLG   407
AtADT5    FLMLAREPIIPRTDRPFKTSIVFAAQEHKGTSVLFKVLSAFAFRNISLTKIESRHQNCPRVVLFFEYIGTSKFFEYLFYVDFEASMAEFRAQNALAEVQEYTSFLRVLG   408
GmADT13A  FVMLAREPIIPRTDRPFKTSIVFAH--DKGTSVLFKVLSAFAFRNISLTKIESRPHRNRPIRLVDDANYGTAKFFEYLFYVDFEASMAEVRAQNALAEVQEFTSFLRVLG   408
GmADT12A  FVMLAREPIIPRTDRPFKTSIVFAH--DKGTSVLFKVLSAFAFRNISLTKIESRPHRNRPIRLVDDANYGTAKFFEYLFYVDFEASMAEVRAQNALAEVQEFTSFLRVLG   409
GmADT11A  FVMLAREPIIPRTDRPFKTSIVFAH--DKGTSVLFKVLSAFAFRNISLTKIESRPHRSRPIRVVDDESEGTAKFFEYLFYIDFEASMAEVRAQNALAEVQEFTSFLRVLG   411
GmADT12B  FVMLAREPIIPRTDRPFKTSIVFAH--DKGTSVLFKVLSAFAFRNISLTKIESRPHRSRPIRVVDDESEGTAKFFEYLFYIDFEASMAEVRAQNALAEVQEFTSFLRILG   411
AtADT3    FVMLAREPIIPRTDRPFKTSIVFAH--EKGTSVLFKVLSAFAFRDISLTKIESRPNHNVPIRLVDEANVGTAKFFEYMFYIDFEASMAESRAQNALSEVQEFTSFLRVLG   401
AtADT6    FVMLAREPIIPRTDRPFKTSIVFAH--EKGTSVLFKVLSAFAFRDISLTKIESRPNHNIPIRLVDEANVGTAKFFEYMFYVDFEASMAEFRAQNALSEVQEFTSFLRVLG   396
AtADT2    FLMLAREPIIPCTNRPFKTSIVSL--BEGPGVLFKAIAVFALRQINLTKIESRPLRKHPLRAS-----GGLKMFDYLFYVDFEASMADEVAQNALRHLPEFAILRVLG   372
GmADT12D  FLMLAREPIIPGTDRRFKTSIVSL--BEGPGVLFKAIAVFALRQINLTKVL*----------------------------------------------------------   311
GmADTU4   FLMLAREPMIPGTDREFKTSIVSL--BEGPGILFKAIAVFALRQINLTKIESRELRNQPIRADD--SNNSKYFDYLFYIDFEASMAEQSAQNALRHIKEFAILRVLG   376
GmADT13B  YLVLARDPIIPKADKPFKTSIVBTL--BEGPGVLFKAIAVFALRDINLNKIESRPQRNRPLRVVDDSNIGTAKYFDYLFYIDFDASMTBRAQNALCHLQEFAILRVLG   321
GmADT12C  YLVLARDPIIPKADKPFKTSIVBTL--BEGPGVLFKAIAVFALRDINLNKIESRPQRNRPLRVVDDSNIGTAKYFDYLFYIDFDASMTBRAQNALCHLQEFAILRVLG   376
AtADT1    FLILAREPMIPRTDRPYKTSIVBSL--BEGPGVLFKAIAVFALRDINLSKIESRELRVRVVDGSNNGSAKYFDYLFYVDGSMADIRACHALGHLQEFASFIRILG    384
GmADT17   FLVLAREPIIPGTDREHKTSIVBSL--BEGPGVLFKALAVFAMRDINLSKIESRPLKQRSLRVVDHLNEGSATYFDYLFYIDIFEASMAEFRAQNALCQLCISSCSWLLSY   391

AtADT4    SYPMDMTPWSMTSTEEA--*-----------------   425
AtADT5    SYPMDMTPWSTLPSEDV--*-----------------   426
GmADT13A  SYPMDMTPWTPSSRGD*-----------------   425
GmADT12A  SYPMDMTPWTPSSRGD*-----------------   426
GmADT11A  SYPIDMSPWTPSCPRQN--*-----------------   429
GmADT12B  SYPIDITPWTPSSPRQN--*-----------------   429
AtADT3    SYPMDMTSWSPSSSSSS--S--STFSL*----------   425
AtADT6    SYPMDMTPWSPTSSTSS--*-----------------   414
AtADT2    SYPVDTPML*-----------------   382
GmADT12D  -----------------------------------------
GmADTU4   SYPVDTSLT*-----------------   386
GmADT13B  PAFLYKVGIIRKHCLSICCNEQVTISQNKIIICHPAD*   359
GmADT12C  CYPIDT-TI*-----------------   385
AtADT1    CYPMDTVR---*-----------------   393
GmADT17   GYSHIEVQCLN*-----------------   403
```

Smith (2014) determined that a single amino acid residue, Phe341, is critical for secondary PDT activity of ADTs. When the sequences of the GmADTs were analyzed, five of them (GmADT13B, GmADT12C, GmADT17, GmADT12D and GmADTU4) displayed Phe to Leu substitution, at Phe215 (residue location in relation to the mature AtADT5 protein used in Appendix A). These five GmADTs also contain a PDT conferring domain (PAC) (El-Azaz *et al.*, 2016).

3.3 *GmADT* gene structure and phylogenetic analysis

Based on gene structures, *GmADTs* can be divided into two groups: those with introns and those without. *GmADT11A, GmADT12A, GmADT12B and GmADT13A* having no introns, while *GmADT12C, GmADT12D, GmADT13B, GmADT17A* and *GmADTU4* have introns (Figure 3.2). The intron-containing genes all have at least 8 exons, that are very similar in size and distribution. *GmADT12C, GmADT13B* and *GmADT17A* contain large introns (approximately ≥1kb) immediately after the first exon, and another large intron (approximately ≥1kb) after the sixth exon. *GmADT12D* and *GmADTU4* each have large introns (approximately 1.5kb in size) after the third exons. All intronic *GmADTs* have more than 8 exons, except *GmADT12D*, with *GmADT12C* and *GmADT17A* having large introns of approximately 1 kb in size separating the last two exons.

To delineate the GmADTs based on sequence similarities a neighbour-joining tree was constructed using mature protein sequences of characterized plant ADTs and putative GmADTs. As seen in Figure 3.3, the GmADTs formed three distinct subgroups similarly to those formed by AtADTs. GmADT17A, GmADT13B and GmADT12C grouped in 'Subgroup I' with AtADT1. GmADT12D and GmADTU4 grouped into 'Subgroup II' with

AtADT2. GmADT13A, GmADT12A, GmADT11A and GmADT12B grouped in 'Subgroup III' with AtADT3, AtADT4, AtADT5 and AtADT6. The GmADTs with intronic sequences grouped with *Arabidopsis* ADTs that have introns, and likewise with the intronless ADTs. In each subgroup, GmADTs clustered together and each GmADT clade contained paralogs except for GmADT17. The P-protein sequence (without the chorismate mutase domain) from *E. coli* was used to root the tree.

A more extensive neighbor-joining tree with annotated ADT sequences from a variety of plant and microorganism species, shown in Appendix B, was used to explore the evolutionary history of GmADTs. Sequences from other legume species, *Vigna angularis, Trifolium pratense, Medicago truncatula* and *Phaseolus vulgaris* were used to look for any potential legume-specific pattern of ADT grouping, signifying potential importance to isoflavonoid synthesis, which is a feature unique to legumes. However, a legume-specific grouping pattern was not observed. Additionally, sequences from more primitive lycophytes, liverworts and mosses were used to determine if duplication of GmADTs occured before angiosperm speciation, that is very early-on in soybean evolution, which would signify their importance to plants. Based on the phylogeny, where GmADTs group amongst the lycophytes, liverworts and mosses, along with the other plant ADT sequences, it is clear that GmADTs did in fact duplicate before angiosperm speciation. Furthermore, algae and bacterial ADT and PDT sequences, respectively, were used to root the tree (Appendix B).
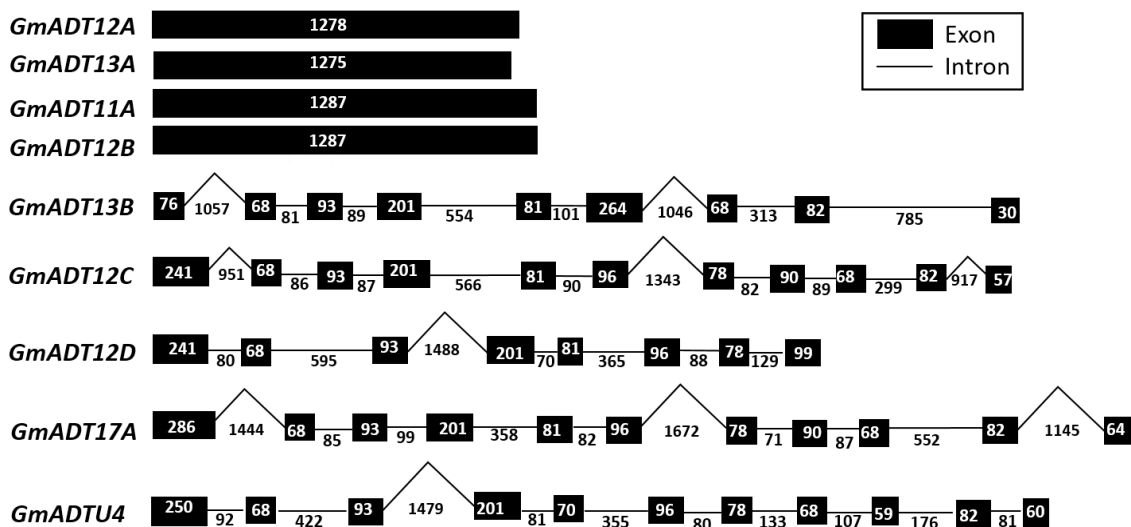
Figure 3.2| Gene structure of GmADTs. Using the Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html) annotations, the gene structures of all nine *GmADTs* were determined. Lengths of bars have been approximated and are not to scale in accordance with exon and intron base pair sizes, indicated by the numbers on the figure.
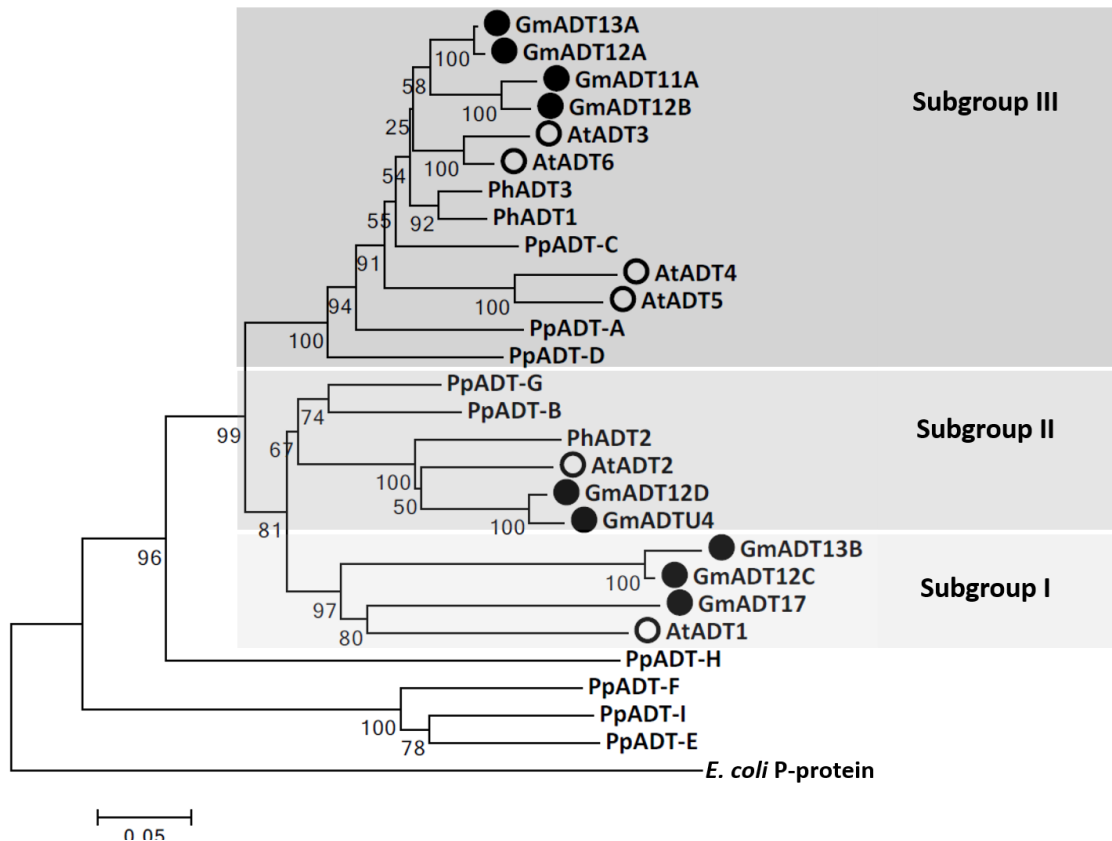
Figure 3.3| Phylogenetic relationship between GmADT genes and characterized plant ADTs. A neighbour-joining tree of the mature protein sequences of GmADTs, AtADTs, PpADTs and PhADTs was constructed using 1000 bootstrap replications, with the percentage of replicate trees in which the respective taxa clustered together shown. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site (indicated by scale). The *E.coli* P-protein sequence was used to root the tree. Evolutionary analyses were conducted in MEGA7 (Kumar, *et al.* 2016). Black circles represent the 9 GmADTs and white circles represent the characterized AtADTs. Labels indicate the subgroups identified in *Arabidopsis* that the GmADTs fall into. *E. coli* indicates the characterized *E. coli* P-protein. Abbreviations: At, *Arabidopsis thaliana;* Pp, *Pinus pinaster;* Ph, *Petunia hybrida.* Accession numbers: PhADT1, ACY79502.1; PhADT2, ACY79503.1; PhADT3, ACY79504.1; PpADTA, APA32582.1; PpADTB, APA32583.1; PpADTC, APA32584.1; PpADTD, APA32585.1; PpADTE, APA32586.1; PpADTF, APA32587.1; PpADTG, APA32588.1; PpADTH, APA32589.1; PpADTI, APA32590.1; AtADT1, OAP11955.1; AtADT2, OAP02204.1; AtADT3, ABD67752; AtADT4, ABD67753.1; AtADT5, ABD67754.1; AtADT6, OAP14989.1; *E. coli,* WP_115444483.1.

3.4 *GmADT* expression analysis

<u>3.4.1 Tissue-specific expression</u>

In order to make inferences into their potential roles in soybean, the tissue-specific expression of each *GmADT,* along with *GmIFS2,* was examined. Expression levels of each of the candidate *GmADTs* was obtained from the Phytozome database (Table 3.1), and plotted in a heatmap (Figure 3.4). The tissue-specific expression for *GmIFS2* was also included in the heat map, as GmIFS2 was used as the bait protein in the Co-IP assay and *GmIFS2* and *GmADT* were potentially interacting partners (Dastmalchi *et al.* 2016). The results revealed that the *GmADT* family members show unique tissue-specific expression profiles. *GmADT13B*, *GmADT13A*, *GmADT12A, GmADT12D* and *GmADT12C* show very high expression in leaf tissue. *GmADT13A* and *GmADT12A*, along with *GmADT12B* and GmADTU4, also show high expression in stem, and *GmADT13B* also shows high expression in the shoot apical meristem. *GmADT12A*, as well as *GmADT12B,* show high expression in root tissue as well. *GmADT17A* shows high expression in root and nodule tissues, while *GmADT11A* appears to show highest expression in tissues related to reproduction, namely the flowers, pods and seeds. *GmADT12D* and *GmADTU4* show a pronounced expression in seeds, and *GmADTU4* also shows a high expression in pod and stem tissues. *GmIFS2* showed the highest expression in root, leaf and root hair tissues.
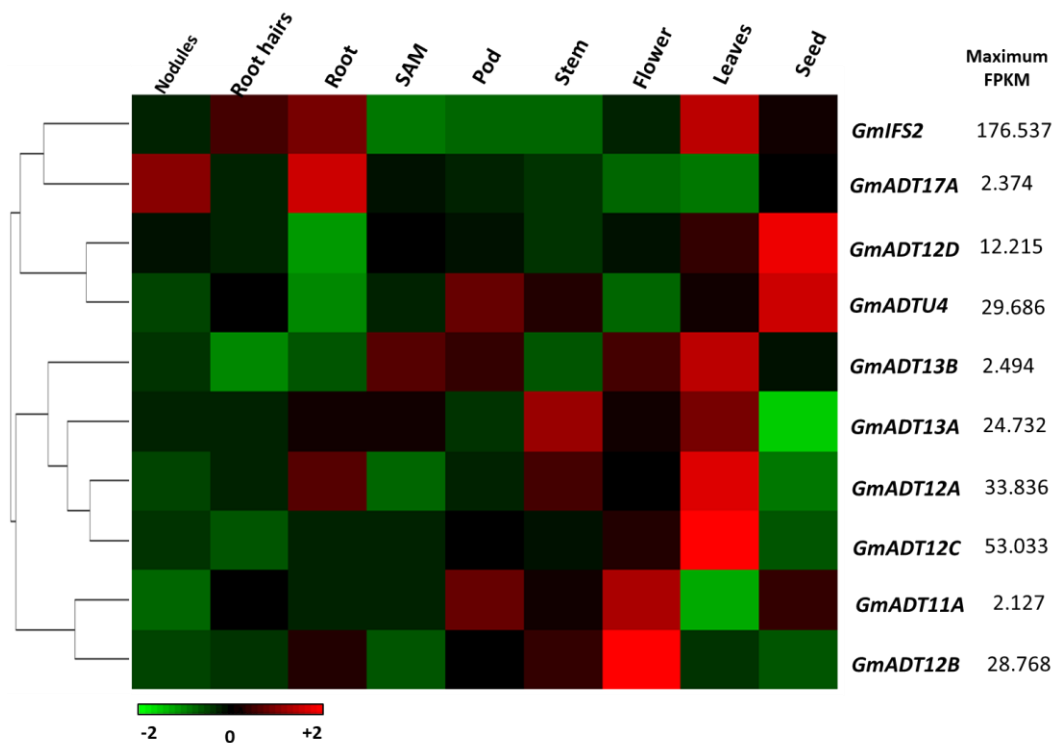
Figure 3.4| Tissue-specific expression profile of *GmADT* genes in soybean. Soybean RNA-seq data across different tissues (https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax) were normalized and a heatmap was constructed. Transcript abundance is indicated by the color scale below the heat map with a gradient from red (high) to green (low). Numbers to the right indicate the maximum value of fragments per kilobase of million mapped reads (FPKM) for each respective gene. SAM: shoot apical meristem.

3.4.2 Expression in response to stress

It has been well documented that the biosynthesis of flavonoids, isoflavonoids, and other metabolites of the phenylpropanoid pathway are induced in response to both biotic and abiotic stresses. Therefore, it follows that *GmADT* expression will also be induced, as more phe is required during these stressful times. Therefore, the expression of *GmADTs* in response to both biotic and abiotic stress was analyzed.

3.4.2.1 Expression in response to biotic stress

Silver nitrate (AgNO$_3$) has been documented as an effective means to mimic *P. sojae* infection as it elicits the plant's stress response by inducing the production of glyceollins, which are phytoalexin isoflavonoids (Sukumaran *et al.*, 2018). Therefore, RNA extracted from AgNO$_3$ -treated soybean hypocotyls at various time-points was used in RT-PCR reactions as a means of visualizing changes in gene expression for each *GmADT*. Out of 9 putative *GmADTs, GmADT13A*, *GmADT12A*, *GmADT17A, GmADT12B* and *GmADTU4* displayed induced expression upon AgNO$_3$ treatment, compared with water-treated control samples (Figure 3.5). *GmADT13A*, *GmADT12A*, *GmADT17A* showed increased transcript levels in treated samples compared to controls from 6 hours until 48 hours post-treatment, after which the expression of both the control and treated samples decreased. *GmADT12B* showed increased transcript levels in treated samples compared to controls for the entire duration of the experiment. *GmADTU4* expression was induced 12 h post treatment, and showed increased transcript levels until 48 hours post-treatment.
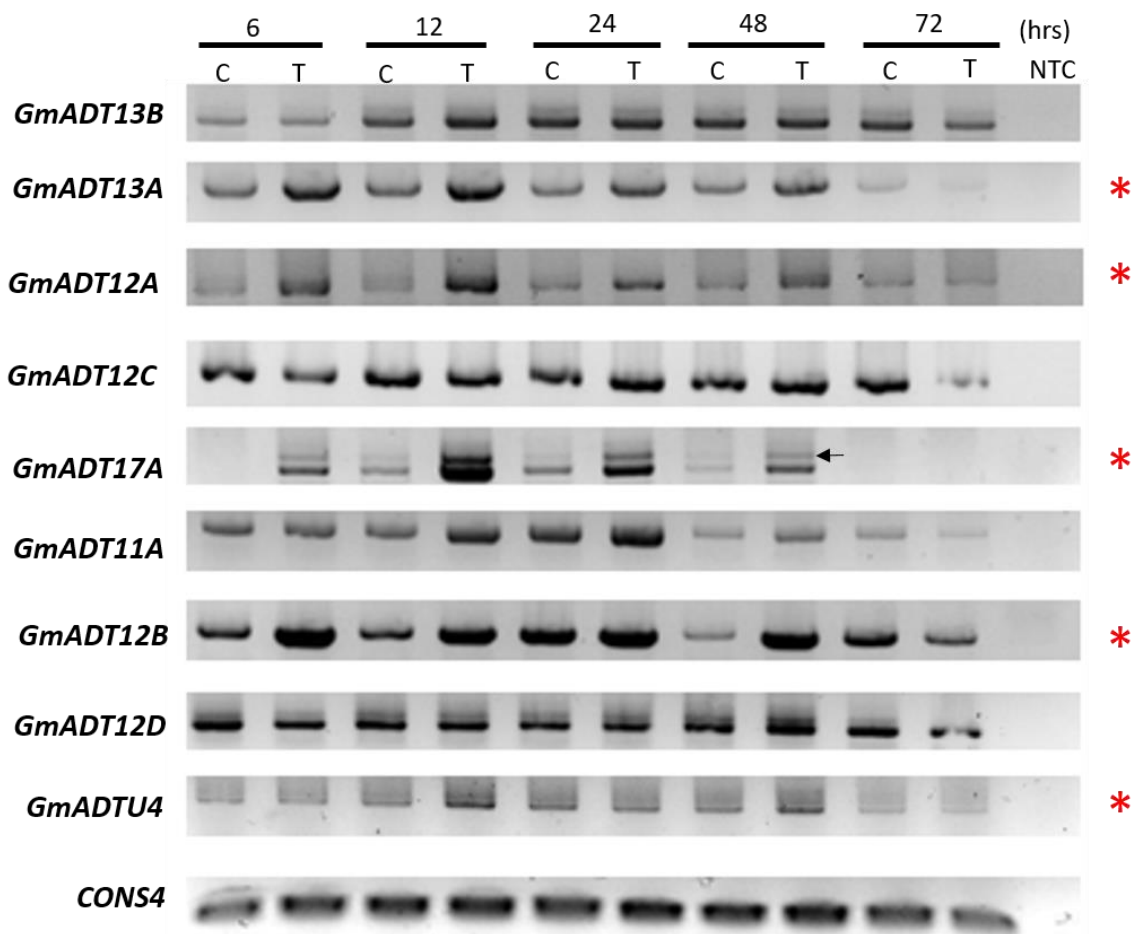
Figure 3.5| Expression of *GmADTs* in response to biotic stress. Total RNA (1 μg) was extracted from AgNO$_3$–treated (T) and control (C) soybean hypocotyls. Expression analysis was conducted by RT-PCR with GmADT gene-specific primers. NTC indicates no template control. Arrow indicates primary *GmADT17A* transcript. CONS4 was used as a loading control. A red asterisk indicates GmADTs that show stress-induced expression.

<u>*3.4.2.2 Expression in response to abiotic stress*</u>

In order to gain insight into whether *GmADTs* are induced in response to abiotic stressors, expression levels of *GmADTs* were studied in response to drought and flood stress. Publically available RNAseq data on both drought and flood stressed soybeans were extracted and analyzed (Chen *et al.*, 2016b). The results revealed that the *GmADTs* do in fact have different expression profiles in response to abiotic stresses (Figure 3.6). Even though expression levels of *GmADTs* in the three biological replicates used in the experiments were not consistent, the results show an overall pattern of differential gene expression relative to the control. *GmADT11A* shows a clear upregulation in all three replicates that were exposed to drought stress, whereas *GmADT12C and GmADT13B* were downregulated in response to both flood and drought stress. In one of the flood replicates, *GmADT17A* showed an increased expression, while being downregulated in all three drought replicates. *GmADT12B* was consistently upregulated in all three flooding replicates, while *GmADT12A* showed increased expression in two; both genes were downregulated in the flood treatments. No difference in the expression levels of *GmADTU4* was observed between the control and stress-exposed samples. *GmADT13A* and *GmADT12D* both appeared to show the highest expression in the control samples. These results indicate that there is stress-specific differential expression of *GmADT* gene family members.
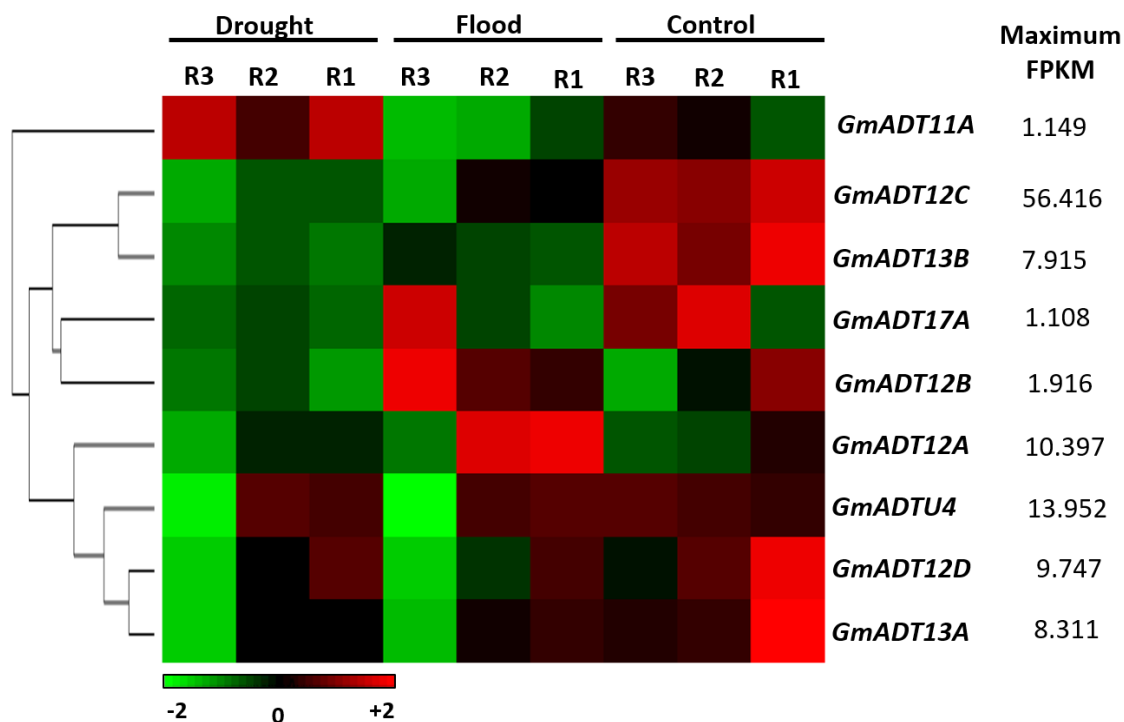
Figure 3.6| Expression of *GmADTs* in response to drought and flood stress. Publicly available RNA-seq data from a study in which soybean plants were exposed to stressed conditions of either flooding or drought, was used to create a heatmap to visualize Gm*ADT* expression. (Accession: PRJNA324522, https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA324522). Transcript abundance is indicated by the color scale below the heat map with a gradient from red (high) to green (low). Numbers to the right indicate the maximum value of fragments per kilobase of million mapped reads (FPKM) for each respective gene.

3.5 *GmADT* promoter analysis

The promoter landscape of all nine *GmADTs* was analyzed, in order to gain insight into the conditions that may influence *GmADT* expression. To characterize the promoters, the promoter element database, PLACE, was used (Higo *et al.*, 1999). This database classifies each promoter element by specific function determined through literature, so it provides a more thorough and specific understanding beyond the GO descriptors. The promoter elements in each promoter were manually sorted to determine which ones were stress-response related and which were developmental. The results of the search can be found in Appendix D, where the promoter element names, target sequences, predicted functions and literature source, and the number of each of these elements in each gene are shown. These numbers were used to determine the percentage of each of the promoter element categories present in each gene, which were plotted in pie charts (Figure 3.7). The promoter elements were grouped into the following categories: Root/nodulation (as isoflavonoids are involved in nodulation), phenylpropanoid pathway-related, abiotic stress (as phenylpropanoid metabolites are involved in abiotic stress response), hormone-induced (stress-response hormones such as jasmonate, salicylic acid, etc), pathogen-response (these are elements upregulated in response to pathogen-infection but not necessarily via hormone recognition) and light response (both stress and non-stressed). Developmentally-related elements were also looked at and grouped into the following categories: seed/embryonic development, pollen development, and gibberellin response (this hormone has many important roles in development).
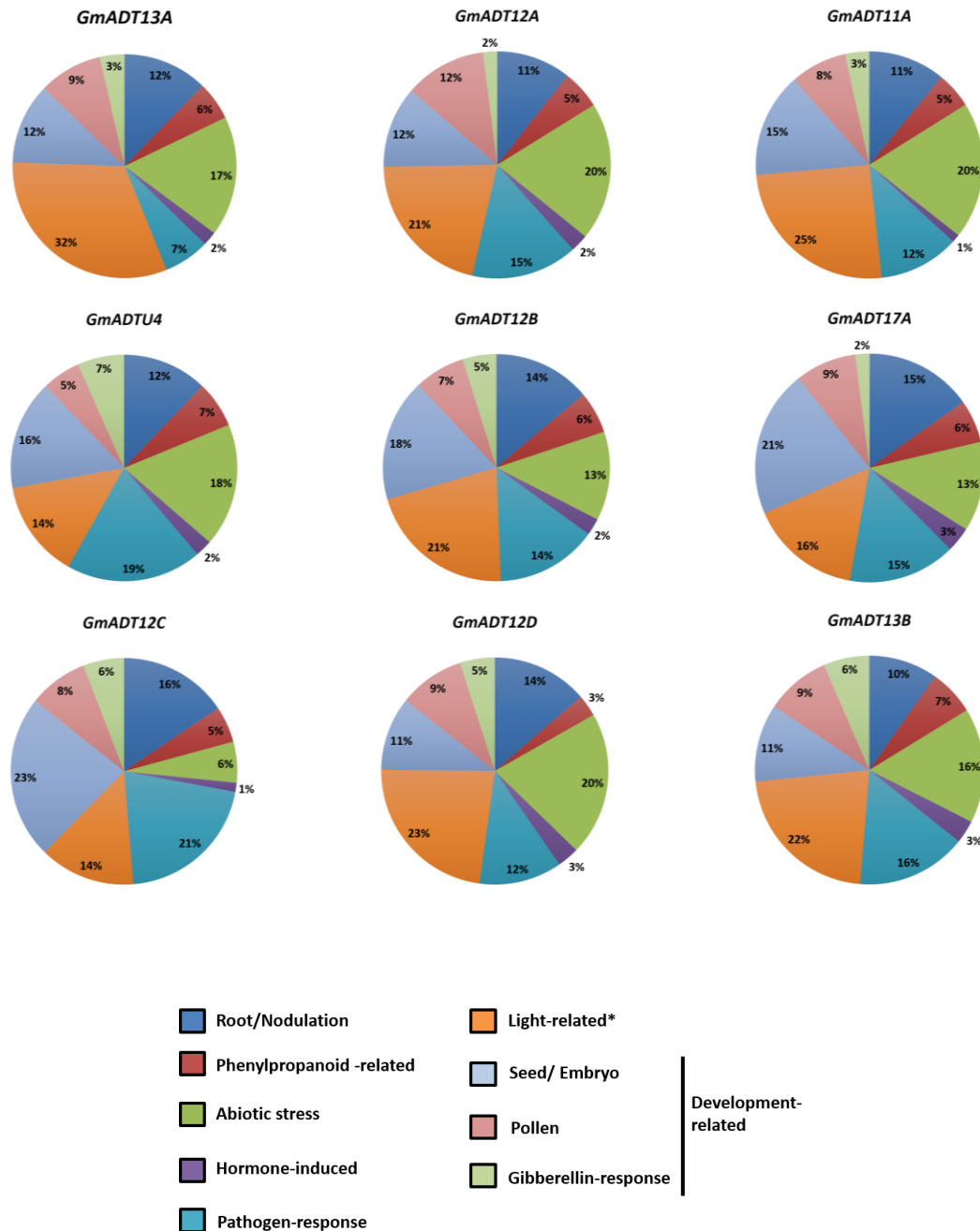
Figure 3.7| Promoter sequence analysis of *GmADTs*. The PLACE database was mined, using the 1000 bp upstream promoter regions for elements involved in stress or development. Percentage of the elements under each category are shown. *Light-related contains elements related to both stressed and non-stressed light-response.

The 9 *GmADTs* show overall similarity in promoter element composition, with categories like light-response and abiotic stress-related elements showing the highest percentages for the majority of *GmADTs.* However, there are differences in proportions of each category between the *GmADTs*.

3.6 GmADTs primarily localize to chloroplasts

Using the TargetP prediction software, the predicted subcellular localization of each of the *GmADTs* was determined (Table 3.2). To verify these predictions, and to aid in verifying the ADT-IFS interaction detected in the co-IP done by Dastmalchi *et al.* (2016), the GmADTs were translationally fused to YFP. These translational fusions were then transiently expressed in *N. benthamiana* leaves, and visualized by confocal microscopy. To visualize the chloroplast location (in order to confirm if GmADT*s* are chloroplastic), the red autofluorescence generated by chlorophyll was taken advantage of and used as a marker (control). Eight out of the nine GmADTs showed chloroplastic localization (Figure 3.8). Specifically, a globular expression pattern in the outer edges of the chloroplast was observed, indicative of stromule localization. In addition to the globular pattern, GmADTU4 displayed localization in additional thin, elongated structures (Figure 3.8 - see Appendix C for additional images). Among nine GmADTs, GmADT13B was not localized to the chloroplast; instead it showed a very clear nuclear and cytosolic localization. To determine if the presence of GmADT13B-YFP (approximately 67 kDa) was due to its passive diffusion into nucleus, the fusion protein size was increased to approximately 96 kDa by adding mCherry to the translational fusion. It was found that GmADT13B-mCherry-YFP also localized to the cytosol and nucleus (Figure 3.8).
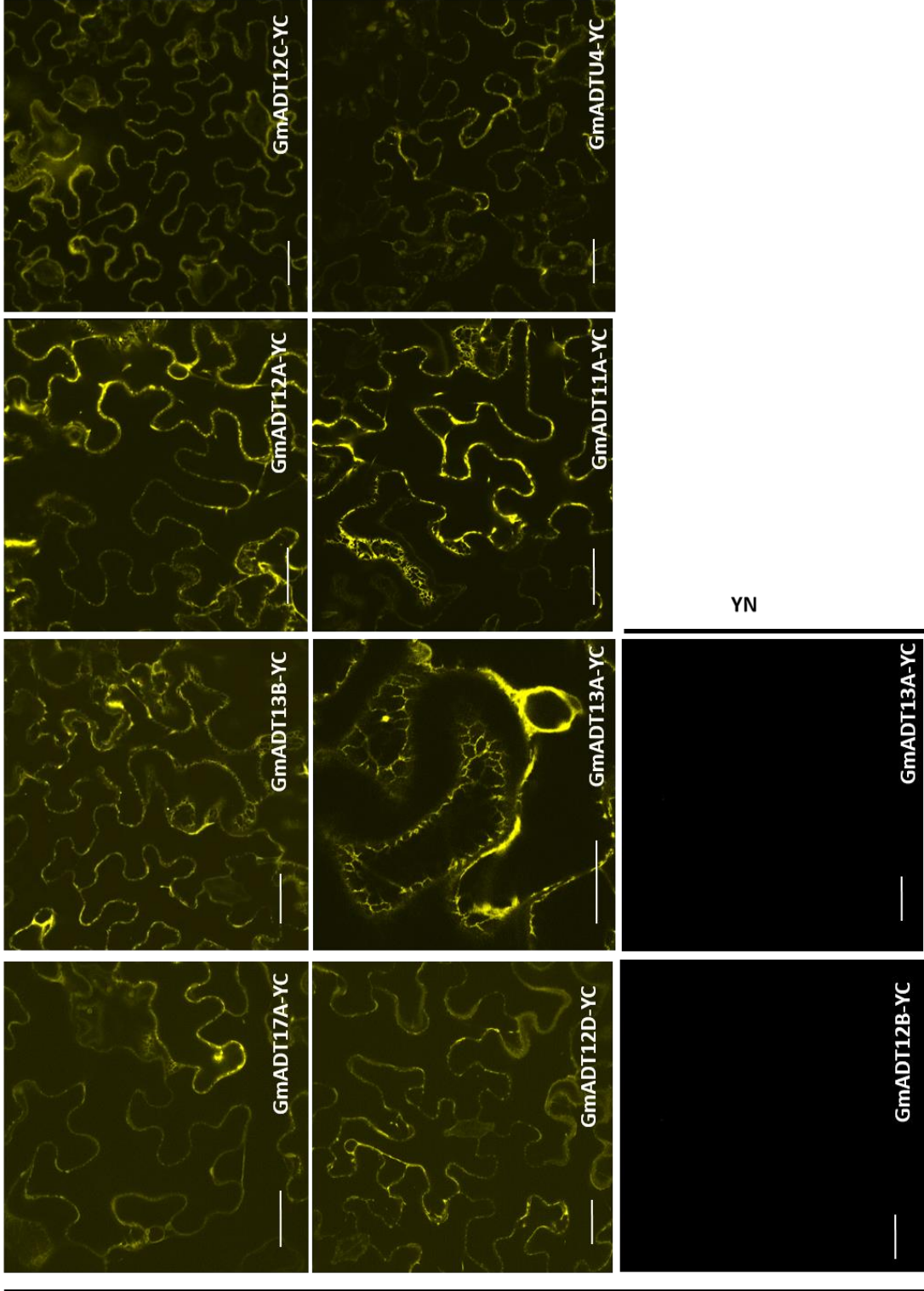
3.7 GmADT and GmIFS do interact *in-vivo*

To confirm if the GmADT and GmIFS interaction initially detected via Co-IP actually occurs *in-planta*, BiFC assays were conducted. GmADTs was translationally fused to the N-terminal of YFP and GmIFS2 was translationally fused to the C-terminal of YFP (and vice versa for the reciprocal combinations). In this assay, if the two proteins interact (either directly or indirectly in the form of a complex with other proteins and the two YFP halves are close enough to come together and complete the fluorophore, fluorescence will be emitted at the location of the interaction. As with the subcellular localization, potential interactions were viewed using confocal microscopy.

The interaction between GmADT and GmIFS was confirmed to occur *in-vivo*, for eight out of nine GmADTs, with GmADT12B being the only exception (Figure 3.9). Reciprocal combinations are shown in Appendix E. Fluorescence was observed in a reticulate pattern, indicating the interaction is occurring at the ER surface, where GmIFS is located. Similar results were obtained for the reciprocal combinations (Appendix E). Co-infiltration of each GmADT-YN with an empty YC (and the corresponding reciprocal combination) were used as negative controls, in which no fluorescence was observed.

Figure 3.8| Subcellular localization of the GmADTs. A translational fusion of GmADT-YFP was transiently expressed in *N. benthamiana* leaf and visualized with confocal microscopy. Confirmation of localization was performed through co-localization of the GmADT-YFP fusion with the chloroplast autofluorescence (in red). The white arrowhead is directed at the nucleus for GmADT13B. Scale bars represent 30 µM. Shown is the vector T-DNA cassette showing translational fusions of GmADTs with YFP or GmADT13B with mCherry and YFP.
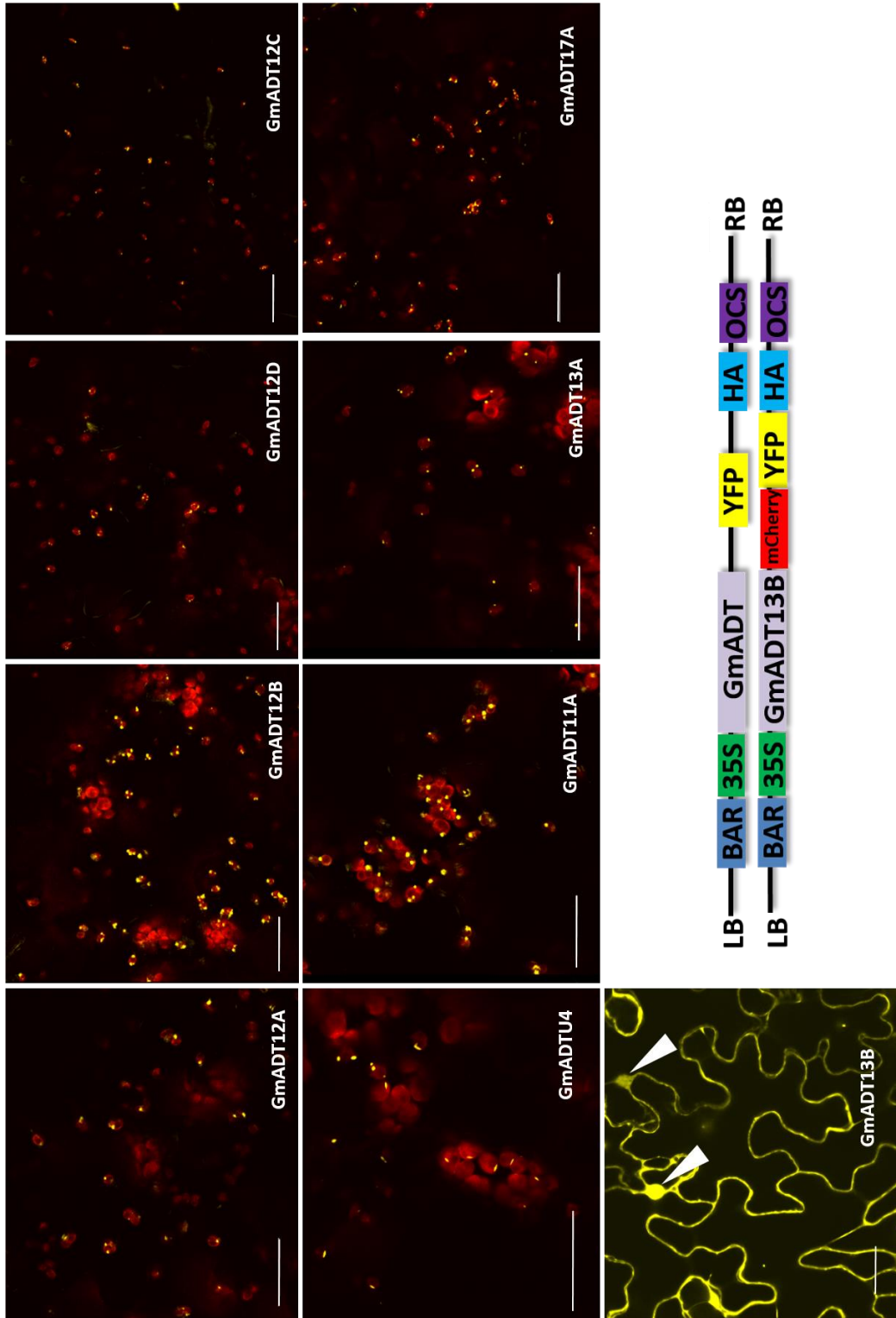
Figure 3.9| GmADT and GmIFS2 interact *in planta* at the ER. Bi-directional interaction between GmADTs and GmIFS2 was assayed by co-expression of translational fusions with N–terminal (YN) and C–terminal (YC) fragments of YFP in *N. benthamiana*. Shown are one set of co-expressions, between the GmADT-YC and GmIFS-YN. As a negative control, GmADT-YC vectors were co-expressed with empty YN vectors. Scale bars represent 30 μM.

GmADT12C-YC

GmADTU4-YC

GmADT12A-YC

GmADT11A-YC

GmADT13B-YC

GmADT13A-YC

YN

GmADT13A-YC

GmADT17A-YC

GmADT12D-YC

GmADT12B-YC

GmIFS2-YN

Chapter 4- Discussion

Isoflavonoid biosynthesis is a legume-specific novelty with invaluable roles in the plants. From their role in facilitating rhizobial communication during nodulation, as phytoalexins - released in response to pathogenic attack-, and the immense health benefits they provide in the human diet, the importance of isoflavonoids cannot be understated. Previous work has elucidated an isoflavonoid biosynthetic metabolon, in which key enzymes in the pathway aggregate spatially and temporally in order to maximize the efficiency of isoflavonoid synthesis (Dastmalchi *et al.* 2016). This metabolon was first discovered through co-IP, which identified previously unknown interactions of isoflavonoid biosynthetic enzymes with IFS, the first committed step in isoflavonoid synthesis. One of these interactors was GmADT, a chloroplastic enzyme far upstream of IFS in the phenylpropanoid pathway. ADT in its own right is a critical enzyme as it catalyzes the last step in the biosynthesis of phe, the precursor to the phenylpropanoid pathway.

In this study, 9 putative ADTs family in soybean were identified. The GmADTs formed these distinct phylogenetic subgroups in the phylogenetic tree, suggesting their potentially unique evolutionary histories. The gene family members were determined to have unique tissue-specific expression patterns, as well as expression in response to abiotic and biotic stress, with some genes being differentially expressed while others are not. Most importantly, the GmADT-GmIFS2 interaction was confirmed to be occurring at the ER surface, introducing a new direction of investigation, as to how the enzymes, predicted to be localized in separate membrane-bound compartments, interact.

4.1 Soybean ADTs form distinct subgroups that suggest their importance to the isoflavonoid metabolon

The *in-silico* mining approach identified nine putative *ADTs* present in soybean. Soybean is a paleopolyploid, meaning many of its genes are present in large families. The phylogenetic analysis showed that putative GmADTs form 3 distinct subgroups (Figure 3.3), analogous to the subgroups formed by the characterized *AtADTs* (Cho *et al.* 2007). Despite the use of peptide sequences for the phylogenetic analysis, the subgroups appear to be formed from the presence or absence of introns; GmADT13A, GmADT12A, GmADT11A and GmAT12B have no introns and group together, whereas GmADTU4 and GmADT12D and GmADT13B, GmADT12C and GmADT17A, have introns and also group together in respective separate subgroups. (Figures 3.2 and 3.3). The GmADTs, along with ADTs from other plants, duplicated early on, as is evident in the phylogenetic tree where some GmADT homologs can be seen diverging from the same common ancestors as primitive plant species (Appendix B). This phenomenon has been seen in other studies characterizing ADTs (Dornfeld *et al.*, 2014, El-Azaz *et al.*, 2016), suggesting that the ADT enzyme evolved early on in vascular plant development, implying its importance to plant survival.

Large gene families are often seen in genes involved in pathogen defense, stress signaling and developmental processes (Cannon *et al.*, 2004). When multiple copies of a gene exist in an organism, if one gene copy is lost, there is often minimal effect as the other gene copies pick up the slack, in a process known as functional compensation or functional redundancy (Kafri *et al.*, 2009). Duplicate copies of genes are often lost over

time, unless there are certain selection pressures that promote retention in the genome, such as involvement in complexes and/or tightly regulated pathways, where a loss of one member can affect the entire complex (Edger and Pires 2009). ADT is positioned in a critical pathway location, producing phe that is used for both protein and phenylpropanoid synthesis; reduced phe levels could drastically affect both of these vital processes. Therefore, the retention of so many *GmADT* isoforms in the genome could be as a result of this critical pathway location, where in the event of a null mutation in one of the family members, other isoforms can maintain phe levels through functional compensation

4.2 GmADTs contain many conserved motifs and residues that suggest their activity

As in Arabidopsis, both intronless and intronic *ADTs* exist in soybean (Figure 3.2). Interestingly, the intron-containing genes - *GmADT13B, GmADT12C, GmADT12D, GmADT17A* and *GmADTU4* - are the same genes whose gene product contain the residues and/or motifs discussed in section 3.2 that suggest PDT activity. They form distinct subgroups of their own, with *GmADT17A, GmADT13B and GmADT12C* having a similar intron-exon pattern and grouping into subgroup I. These three genes also are the same three genes that are missing the critical residues determined to be necessary for catalytic and regulatory function. GmADT12D and GmADTU4, also have a similar pattern and they group into subgroup II. Notably, the *GmADTs* also have similarly sized introns located relatively at the same distances, and forming at the same intron-exon splice sites as seen in *AtADTs* in the respective subgroups, suggesting a commonality of evolution amongst

all ADTs. There appears to be no significant effect of intron-exon pattern on transcript or protein size between *GmADTs* of similar total exon length, seen in Table 3.2, where most *GmADTs* are similar in size. However, *GmADT12D* lacks the additional exons that the other *GmADTs* have, naturally resulting in truncated transcript and protein sequences. The significance of these intron-exon junctions has yet to be explored in-depth.

GmADTs showed a high level of conservation in their catalytic and ACT domains, amongst each other as well as with well-characterized *Arabidopsis* ADT family members (Figure 3.1, Appendix A). GmADT17A, GmADT13B and GmADT12C contain multiple substitutions at residues experimentally determined to be critical to PDT function and regulation. Most notable is The Thr177Ser substitution in GmADT13B and GmADT12C at the highly conserved TRF motif (Thr177-Arg178-Phe179) (see Appendix A). Interestingly, Zhang *et al.* (2000) tested a Thr177Ser mutation in their kinetic analysis, which resulted in a 40-fold decrease in PDT activity. This same Thr177 residue, along with Arg178 and Phe179, were also shown to be critical for PDT activity by Hsu *et al.* (2004), and were also shown to be present at the active site of PDT, implicating them in PDT function (Tan *et al.*, 2008). Furthermore, Tan *et al.* (2008) identified a number of individual residues involved in hydrogen-bonding to the TRF active site, at which GmADT13B, GmADT12C and GmADT17A had substitutions (marked with asterisks in Appendix A). However, as mentioned further on in the current section (4.1), GmADT17A, GmADT13B and GmADT12C all contain conserved residues that suggest they likely possess secondary PDT-activity. Given this, it is unclear why they would contain substitutions that have been shown to *decrease* PDT activity. As it stands, there is no current ADT crystal structure.

Therefore, it is unclear what roles these residues and motifs play in arogenate substrate binding. Given the structural similarity between prephenate and arogenate, perhaps these residues and motifs are necessary for arogenate catalysis, and the substitutions at these locations in GmADT13B, GmADT17A and GmADT12C allow them to have additional affinity to prephenate.

Allosteric phe binding is reserved to the C-terminal ACT domain of the protein, specifically at the ESRP and GALV motifs (Zhang *et al.*, 1998). Mutants in the ESRP and GALV significantly reduce feedback regulation and result in increased levels of phe and tyrosine (Pohnert *et al.* 1999, Yamada *et al.*, 2008). In addition, Tan *et al.* (2008) determined that the ESRP and GALV domains are critical for PDT subunit binding, as PDTs require tetramerization to be active. In fact, the serine residue of the ESRP domain, and the leucine residue of the GALV domain, make direct contact with phe (marked with asterisks (*) in Appendix A) (Tan *et al.*, 2008). In the current study, the leucine residue in the GALV motif was conserved across all sequences (Appendix A). Interestingly, GmADT12D is missing the ESRP domain entirely. In a study where *Arabidopsis* mutants resistant to *m*-tyrosine were studied, an ADT2 mutant, *adt2-1D,* with a serine-to-alanine substitution at the 'S' residue of the ESRP motif resulted in extremely increased phe levels, but also abnormal leaf development, resistance to cabbage looper growth and increased salt tolerance (Huang *et al.*, 2010). Since GmADT12D is missing this motif entirely, and is also an ADT2-like protein, as is evident from the phylogeny tree, it suggests that perhaps GmADT12D is involved in keeping phe levels high in the tissues it is expressed in, ensuring these tissues have increased phenylpropanoid metabolites for stress tolerance.

Five of the GmADTs (GmADT13B, GmADT12C, GmADT17, GmADT12D and GmADTU4) were determined contain a Phe to Leu substitution (Phe215Leu, in reference to the *AtADT5* sequence in Appendix A). The leucine residue in this position elicits PDT activity (Smith, 2014). Interestingly, these five GmADTs also group with AtADT1 or AtADT2 that have demonstrated secondary PDT activity (Cho *et al.*, 2007). In addition to the Phe215Leu substitution, a second region known as the PAC or 'PDT conferring domain' was discovered, present across all plant and algae lineages, that determines if a plant ADT also has a secondary PDT functionality (El-Azaz *et al.*, 2016). The same five GmADTs that have the Phe215Leu substitution contained the PAC domain as well (Appendix A). These were also the same five GmADTs that contain introns. These sequence conservations indicate that, like *Arabidopsis* and *Pinus pinaster*, some soybean ADTs may have also retained the secondary PDT pathway.

4.3 GmADTs predominantly localize to the chloroplast stromules

Eight out of the nine GmADTs were found to localize to the chloroplasts, specifically the stromules, as is evident by the point, globular fluorescence pattern, rather than an all-over chloroplastic signal (Figure 3.8). This is consistent with ADTs studied from *Arabidopsis, Petunia hybrida* and *Pinus pinaster*, that also show the same localization pattern (Maeda *et al.*, 2010, El-Azaz *et al.*, 2016, Bross *et al.*, 2017). GmADTU4, in addition to stromule localization, displayed thin, elongated YFP signals, at what appears to be the chloroplast equatorial plate (Figure 3.8, Appendix C). These signals are reminiscent of the localization signals of *Arabidopsis* ADT2, which was determined to have a role in chloroplast division (Bross *et al.*, 2017, Abolhassani Rad *et al.*, 2018), as its localization

followed similar patterns as known proteins involved in this division. GmADTU4 grouped into the same subgroup as AtADT2, suggesting a potential role in chloroplast division for GmADTU4. It is important to note that this study did not use native promoters for the transient protein expression, but rather the 35S constitutive expression promoter present in the pEG expression vectors used for gene cloning. However, the chloroplastic localization of GmADTs seen in this study is supported by the transit peptides the GmADTs possess, as well as evidence from previous studies that have also shown similar chloroplastic localization (Maeda *et al.*, 2010, El-Azaz *et al.*, 2016, Bross *et al.*, 2017).

The localization of GmADTs to stromules further suggests ADTs may play a role in stress response, thus explaining GmADT interaction in the isoflavonoid metabolon. There have been multiple studies thus far that have shown a direct link between stromule dynamics and stress-response. For instance, NRIP1 is an NB-LRR receptor interacting protein required for full resistance against the *Tobacco mosaic virus* as part of the effector triggered immunity (Caplan *et al.*, 2008). NRIP1 is a chloroplastic-localized protein, but upon binding the viral effector molecule p50, NRIP1 is actually triggered to localize in the nucleus and cytoplasm, channeled through the stromules into the nucleus, through stromule-nuclei connections. The authors also observed that artificial application of hydrogen peroxide and salicylic acid – signs of cell stress- could alone induce stromule formation (Caplan *et al.*, 2015). Further to this, microscopy experiments determined that stromule formation increases during the day, via light-sensitive redox signals, or when ROS were upregulated within the chloroplasts due to stress, suggesting a redox signal-dependent pathway to initiate stromule formation (Brunkard *et al.*, 2015). In addition to

rhizobacteria interactions, stromules also likely play a role in fungal arbuscular mycorrhiza interactions (Fester *et al.*, 2007). Upon colonization of root cells by the mycorrhiza, chloroplast numbers increase, and become connected through tubular projections of the chloroplasts that form a network surrounding the fungus penetrations (Fester *et al.*, 2007). Furthermore, upon this fungal colonization, plastid division is significantly increased (Fester *et al.*, 2007), suggesting a potential role for GmADTU4 in plant-symbiont interactions given its potential role in plastid division. Altogether, the localization of GmADTs to the stromules, that have a well-documented role in stress-responses, coincides with GmADTs being involved in the isoflavonoid metabolon.

GmADT13B was the only ADT isoform that did not show chloroplastic localization, and instead was in the cytosol and nucleus. The lack of chloroplast expression may be explained by large portions of the transit peptide not being present in GmADT13B (Figure 3.1). Bionda *et al.* (2010) determined that a transit peptide of at least 60 amino acids is required for translocation into the chloroplast. In its transit peptide region, GmADT13B only has 44 amino acid residues, whereas the other GmADTs have transit peptide regions upwards of approximately 100 amino acids. Therefore, this small transit peptide region may hinder the ability GmADT13B to be translocated into the chloroplast, explaining the presence in the cytosol and not the chloroplast.

The nuclear localization of GmADT13B was surprising, as no nuclear localization signal (NLS) was predicted in the GmADT13B sequence. A similar result was found in *Arabidopsis,* where ADT5 was found to be localizing to the chloroplasts and nucleus , with no sequence being predicted by online prediction software tools (Abolhassani Rad, 2017).

Using deletion constructs, it was determined that the region responsible for nuclear localization was in the ACT domain, at a specific motif, which allows for an interaction with a PDAT1 enzyme, that AtADT5 'piggybacks' into the nucleus (Abolhassani Rad, 2017). GmADT13B does not contain this motif, but it contains an additional 35 amino acids in its ACT domain that may allow an interaction with and subsequent 'piggybacking' on another nuclear-localized protein, similarly to ADT5 piggybacking PDAT1. Interestingly, GmADT13B has the highest molecular weight out of all nine GmADTs (Table 3.2) and yet is the only GmADT to localize in the nucleus, further illustrating that this localization is not due to passive diffusion, but rather a deliberate localization.

4.4 *GmADTs* have unique promoter landscapes that point towards upregulation in response to various stresses

When genes are present in large gene families, these different members often development new functions and roles. This often occurs via rearrangements and duplications in the promoter regions, rather than the coding sequences (Langham *et al.*, 2004). Therefore, in order to further characterize potential roles *GmADTs* may play, as well as aid in explaining why GmADTs would be interacting with GmIFS, a promoter analysis was done. All GmADT promoters contain similar distributions of the different categories used, but each with slight variations (Figure 3.7). These variations in promoter landscapes may indicate the different conditions in which the *GmADTs* are upregulated. For instance, *GmADT12C* has the smallest proportion of hormone-induced promoter elements while showing the greatest proportion of root/nodulation related elements, indicating it may play a role in nodule formation. Whereas *GmADT13A* shows the greatest

proportion of light-induced elements, indicative of its high leaf tissue-specific expression. Or, for example, *GmADTU4*, *GmADT12A*, *GmADT11A* and *GmADT12D* show the greatest proportions of abiotic stress related elements, indicating they could be more involved in abiotic responses like drought and/or flooding, which can be seen in Figure 3.6 where *GmADT11A* and *GmADTU4* showed increased expression in response to drought stress, and *GmADT12A, GmADTU4* and *GmADT12D* were increased in response to flooding.

Despite the altered proportions in the *cis*-element categories, the *GmADTs* all contain significant proportions of elements associated with isoflavonoid expression and/or function. For instance, isoflavonoids are important metabolites for promoting nodulation formation in roots, as they facilitate communication with rhizobial symbionts in the soil. The *GmADTs* all had relatively high proportions of root/nodulation-related promoter elements, suggesting *GmADTs* are induced to promote isoflavonoid synthesis for nodulation. Additionally, isoflavonoid biosynthesis is induced during plant stress-response to both biotic stresses like pathogens as well as abiotic stress conditions like drought or temperature stress. *GmADTs* all had promoter elements associated with abiotic stress-response, pathogen-response and light-response, corresponding to the various conditions isoflavonoid synthesis is upregulated. Likewise, there were also phenylpropanoid-related promoter elements present in the *GmADTs*, indicative of phe being required for the phenylpropanoid pathway. There were also significant proportions of promoter elements associated with seed and pollen development, as well as elements involved in gibberellin signaling, a key hormone in a variety of developmental processes. This suggests *GmADTs* are also highly involved in not only plant defense, but

developmental processes as well. Environmental stresses like temperature extremes or drought and flooding cause alterations in the photosynthetic pathways that result in accumulations of reactive oxygen species, that are then transmitted to the nucleus through retrograde signaling (Chan *et al.*, 2016). Since *GmADTs* show distinct promoter landscapes associated with stress-response, they may act as signaling partners in the retrograde signaling pathway to facilitate communication between the chloroplasts and nucleus during stressed-conditions, to promote the synthesis of phenylpropanoid metabolites.

4.5 Differential tissue-specific *GmADT* expression suggests their isoform-specific roles in metabolic pathways

Large gene families allow for individual family members to undergo their own unique selection pressures and evolution, resulting in new expression patterns and roles for the different family genes (Airoldi and Davies, 2012). By this reason, given the large *GmADT* family size, the different *GmADT* members were likely to display unique expression patterns. Even though *GmADT*s are expressed in all soybean tissues, the gene family members differed in their temporal and spatial expression patterns (Figure 3.4 and Figure 3.5). In one study that looked at differential contributions of each *AtADT* to anthocyanin production, an important pigment for UV protection, it was determined that the *AtADTs* differentially contributed to anthocyanin production (Chen *et al*. 2016). Given their exposure to the sun, leaves would likely need the highest concentration of anthocyanins for UV-protection, suggesting the *GmADTs* most highly expressed in leaves – *GmADT13A, GmADT13B, GmADT12A and GmADT12C*- could be contributing more to

anthocyanin synthesis. This can be seen in the promoter analysis (Figure 3.7), where these genes show significant portions of their promoter make-up in the light-related, abiotic stress categories. Similarly, *AtADT4* and *AtADT5* have the biggest contribution to lignin synthesis, a crucial molecule for cell-wall formation and structural support in stem tissue (Corea *et al.*, 2012b). The expression analysis of the *GmADTs* showed that *GmADT13A, GmADT12A,* and *GmADT12B* were increased in the stem, and grouped in the same clade as *AtADT4* and *AtADT5,* suggesting these *GmADTs* are possibly important for shunting phenypropanoid metabolism towards lignin synthesis. Lastly, increased expression of *GmADTU4* and *GmADT12D* in seed tissue (Figure 3.4), and their grouping with *AtADT2* (Figure 3.3), which has recently been shown to be necessary for seed development (El-Azaz *et al.*, 2018), implicating these two *GmADTs* in seed development roles.

The expression pattern of *GmIFS2* was also analyzed, in order to better understand the results of previous co-IP experiments (Dastmalchi *et al.* 2016). *GmIFS2* expression was consistent with the expression of *GmADT12A* – one of the enzymes pulled down in the co-IP assay (Dastmalchi *et al.* 2016) – both of which show high root tissue expression (Figure 3.4). *GmADT13A,* which was also co-precipitated with GmIFS2, shows a reduced root expression, but a high leaf tissue expression that correlates with a high leaf expression for *GmIFS2*. Furthermore, the FPKM values used in the heatmap are transcript-specific, and may not reflect the translation levels.

Many of the *GmADTs* showed a reduction in expression under either drought or flood condition, or in some cases both (Figure 3.6), which may be a result of reduced protein synthesis during stressful times. However *GmADT11A, GmADTU4,* and

GmADT12D showed induced expression under drought conditions in soybean. This result is consistent with the tissue-specific expression mentioned above, as these were also highly expressed in seed tissue, and dehydration is required during seed formation. *GmADTs* with increased expression under flood condition were also expressed at higher levels in stem (mentioned above). Given their expression in stem tissues, which experience hypoxia and need to increase oxygen transport during flooding (Pucciariello *et al.*, 2014), perhaps this condition elicited their upregulation, as the production of lignin for structural support in water-logged roots may be important. Accordingly, these *GmADTs* showed significant proportions of their promoter make-up in the abiotic stress category, indicating these *GmADTs* may be more responsive to flooding and hypoxia stress conditions.

Lastly, the apparent increase in expression of *GmADT* genes upon biotic stress clearly implicates *GmADTs* as being actively involved in these stress response pathways (Figure 3.5), as it has been well-documented for phenylpropanoid enzymes, including those specific to isoflavonoids (Zabala *et al.*, 2006). The *GmADTs* that showed increased expression in response to AgNO$_3$ treatment also contain significant proportions of pathogen-response promoter elements, suggesting that their expression is induced in response to the various signaling pathways that occur in response to biotic stress. To further validate the interaction between GmADT and GmIFS2 and what it means for the isoflavonoid metabolon, ADTs in other studies have been shown to be upregulated in response to stresses. An ADT6-like soybean ADT (gene identifier not specified by authors) was found to be upregulated in response to Asian soybean rust infection, in a coordinated

fashion with ispG - an important enzyme in the methylerythritol 4-phosphate (MEP) pathway that produces isoprenoids, which are important defense molecules (Hossain *et al.*, 2018), clearly demonstrating that *GmADTs* are upregulated in response to biotic stress.

4.6 GmADTs interact with GmIFS2 *in-vivo*, at the ER, expanding the current knowledge of the isoflavonoid metabolon

Two GmADTs, GmADT12A and GmADT13A, were pulled-down in the co-IP where GmIFS2 was used as a bait protein. This does not necessarily mean only GmADT12A and GmADT13A interact with GmIFS2 in live plant cells; there are many factors that result in potential false-positives, as well as the potential of existing interactions to go undetected. For instance, many protein-protein interactions are transient, existing only in certain tissues at certain times. Therefore, if the co-IP assay uses a specific tissue that was collected at a certain time or condition where the protein-protein interaction of interest was not occurring, that interaction would remain undetected. Furthermore, co-IP assays are performed using conditions that minimize potential protein denaturation; however, these conditions may not accurately reflect those occurring in the cell at the time of the interaction *in-planta.* Therefore, this may induce protein-protein interactions that are in-fact false positives, but it may also result in actual interactions going undetected. Therefore, results gained from co-IP assays must be confirmed through other methods. Here, using BiFC assay, the interaction between GmADTs and GmIFS2 was validated.

Interaction between GmADTs and GmIFS2 was found to occur, specifically at the ER surface, where GmIFS2 is anchored within the ER membrane (Figure 3.9) (Dastmalchi

et al., 2016). This interaction occurs for all GmADTs except GmADT12B, not just GmADT13B in the cytosol. Validation of this interaction helps to confirm the notion that GmADT is directly involved in isoflavonoid biosynthesis, more so than just passively producing phe for the phenylpropanoid pathway. Isoflavonoid biosynthesis shows increased expression in response to biotic stress, in their role as phytoalexins, and also play an important role in nodule formation in roots, by promoting interaction with symbiotic bacteria and fungi in the soil. *GmADTs* appear to be upregulated in response to the same stimuli as *GmIFS2*, and therefore could be involved in the metabolon to direct metabolic flux away from the other phenylpropanoid pathway metabolites and towards the isoflavonoinds. This ability of an enzyme to shunt competing pathways towards one particular set of products was demonstrated in GmIFS (Li et al., 2011)*.* GmIFS competes with another enzyme at the flavonoid- isoflavonoid pathway junction for the narinigen substrate, and is able to block flux towards the flavonoids (Li *et al.*, 2011). In fact, expression of GmIFS in a non-legume plant, resulted in induced increased expression of other upstream enzymes in the isoflavonoid pathway (Li *et al.*, 2011), suggesting that it may also trigger the increased expression of *GmADTs.*

While the interaction between GmADTs and GmIFS2 inherently makes sense, how it takes place has yet to be explained. One potential explanation for this interaction is that some of the GmADT molecules that get translated simply remain in the cytosol, while others are brought to the chloroplast via their transit peptides (Figure 4A). It has been previously determined that cleavage of the transit peptide is not in fact required for enzymatic function (Cho *et al.*, 2007, Bross *et al.*, 2017). Therefore, it is also possible for

GmADTs to remain in the cytosol and still be active members of the isoflavonoid metabolon. Furthermore PPY-AT, the enzyme that synthesizes phe through the PDT pathway, is actually localized in the cytosol, and phenylpyruvate is transported out of the chloroplast for PPY-AT to convert phenylpyruvate to phe (Yoo *et al.*, 2013). Therefore, it is possible that arogenate may also be transported out of the chloroplast to cytosolic ADTs, where it can be converted to phe (Figure 4A). GmADTs that potentially possess secondary PDT activity may also function in the cytosol, where they may synthesize phe using prephenate transported to the cytosol (Figure 4A). Evidence of both plastidic and cytosolic shikimate pathway enzymes has been proposed, indicating that co-existing cytosolic and plastid branches of phe synthesis are possible as well (Hrazdina and Jensen, 1992). Previous studies have shown that plastidic chorismate mutase is feedback regulated by aromatic amino acids, whereas cytosolic chorismate mutase is feedback regulated by phenylpropanoid and other metabolites derived from aromatic amino acids (Hrazdina and Jensen, 1992). This suggests that plastid-localized isozymes are responsible for amino acid production to be used for protein synthesis, whereas some GmADT translated molecules may be retained in the cytosol if increased phenylpropanoid metabolites are present (which would be the case in a stress-response situation). This may explain the result from the abiotic-stress response, in which many of the *GmADTs* showed a decreased expression in response to flooding and drought. In the leaf tissues used in that study for RNA extraction, perhaps *GmADT* expression decreased because upon translation, the cytosolic ADTs were induced to produce phenylpropanoid metabolites, reducing the need for overall protein synthesis. Dual-targeting of enzymes

has been shown in other biosynthetic networks as well, such as the isoprenoid biosynthesis network. This dual-targeting occurs via alternate transcription of one single gene, with one translated product transported to the chloroplasts and mitochondria and the other version being cytosolic. Each compartmentalized isoform can then regulate flux in their respective localizations (Heinig *et al.*, 2013).

Another possible explanation is what is known as the trans-organelle continuity (Figure 4B). In this phenomenon, the chloroplast and ER membranes are bound together with very strong forces, where the ER essentially forms a "mesh" surrounding the chloroplasts and the stromules (Barton *et al.*, 2018), and stromule growth and retraction directly correlates to the cortical ER rearrangement (Schattat *et al.*, 2011). This ER-chloroplast continuity potentially facilitates the movement of metabolites such as lipids between the organelles via the plastid stromules (Wang and Benning, 2012). The continuity may also allow the movement of larger molecules, like proteins, which may help to explain how GmADTs can be trafficked through the continuity to reach GmIFS2 at the ER surface (Figure 4B). For example, using photobleaching experiments, it was determined that proteins could flow through stromules (Kohler, 1997). Furthermore, it has also been shown that under stressed conditions stromal proteins in the chloroplasts can be released as vesicles via the stromules, without damaging the chloroplast membrane structure (Ishida *et al.*, 2008). Perhaps GmADTs can also be released from the stromules through vacuoles which can then fuse with the ER membrane, essentially delivering GmADTs to GmIFS2. There are also multiple accounts of proteins being channeled through the stromules into other organelles, like NRIP1 being channeled to the

nucleus through the stromules, or the stromule-facilitated interaction of the chloroplast chaperone cpHSC70-1 and the nuclear-localized AbMV movement protein, both as mentioned previously (Caplan *et al.* 2005, Krenz *et al.*, 2012).

Using a photoconvertible red-to-green fluorescent protein (mEosFP) in differential colouring of stromules from separate chloroplasts, it was shown that there was no fusion of membranes, and fluorescent proteins were not seen moving between the two distinct plastids, suggesting that in-fact stromules do not allow protein trafficking (Schattat *et al.*, 2012). However, this is in stark contrast to a largely growing body of work suggesting the very opposite. Perhaps the trafficking does not happen through membrane fusion between different stromules, or between stromules and other organelles, but rather via vacuole fusion between two sets of membranes, whereby proteins are sent in packages between the two membrane sets, as discovered by (Ishida *et al.*, 2008).

Figure 4| Proposed models for the GmADT-GmIFS2 interaction. A. While most GmADTs get translocated to the chloroplast, some may be retained as functionally active in the cytosol, requiring arogenate transport into the cytosol (represented by the question mark as this transport is currently not known). GmADTs in the cytosol may also possess the secondary PDT activity, where they can also synthesize phenylalanine using prephenate transported to the cytosol (also represented by a question mark). Remaining in the cytosol would enable the GmADT-GmIFS2 interaction (⟷). **B**. The trans-organelle continuity may facilitate the GmADT-GmIFS2 interaction, where the membranes of the chloroplasts and ER are able to form strong junction sites, allowing for the movement of metabolites and other macromolecules between the two organelles. The trans-organelle continuity is shown at the chloroplast and ER membrane interface, within the grey circle.

A

SHIKIMATE PATHWAY
E4P+PEP → Chorismate
CM
PPA–AT
Prephenate → Arogenate → L-Phenylalanine
PDT
Phenylpyruvate
Cytosol
Thylakoids
Stroma
Inner Membrane
Outer Membrane
PPY–AT
L-Phenylalanine
Stromule
ADT
Transit peptide directs transport into chloroplast
ADT
Chloroplast
IFS
ADT/PDT
Functionally active remaining in cytosol
Endoplasmic Reticulum

B

SHIKIMATE PATHWAY
E4P+PEP → Chorismate
Cytosol
CM
PPA–AT
Prephenate → Arogenate
PDT
Phenylpyruvate
Chloroplast
L-Phenylalanine
Thylakoids
Stroma
Inner Membrane
Outer Membrane
PPY–AT
L-Phenylalanine
Stromule
Trans-organelle continuity
Endoplasmic Reticulum
ADT
IFS

Chapter 5: Concluding Remarks

This study demonstrated *in-planta* GmADT and GmIFS2 interaction, specifically at the ER surface, where GmIFS2 is localized. In order to explain how this otherwise unlikely interaction could be happening, it was necessary to determine *why* it was happening-specifically, why an enzyme far upstream of isoflavonoid synthesis, in a totally separate organelle, would be involved in this metabolon.

Based on the results reported in this study, it is clear that GmADTs have roles beyond passive phe synthesis. The evolutionary clades formed between GmADTs, but also with AtADTs and other characterized ADTs studied give potential clues as to what function they serve. Furthermore, the conserved residues in GmADTs may also suggest the enzymatic activity of GmADTs, in whether or not they have retained the secondary PDT functionality. Interestingly, there was distinct and separate clade formation of the *Arabidopsis* ADTs that have been confirmed to have PDT activity along with certain GmADTs, making these likely ADT/PDT candidates. The *GmADTs* show unique tissue-specific expression profiles, which may also indicate their roles. Promoter analysis revealed several promoter element categories present in the *GmADTs*, further implicating certain *GmADTs* in specific roles involved with biotic or abiotic stress, nodulation, or developmental roles. In fact, expression analysis showed that some *GmADT* members are induced upon biotic stress, and that most of the GmADTs localize to the chloroplast stromules, which are known to be involved in cellular stress-responses. Put together, these results strongly suggest *GmADTs* have active roles in stress response and nodulation, beyond passive synthesis of phe, validating the GmIFS2-GmADT interaction.

Stromules form interconnected networks that can facilitate protein transport. Therefore, one explanation for the mechanism that allows GmADT-GmIFS2 interaction is the theory of the trans-organelle continuity. The GmADT-GmIFS2 interaction may also be simply explained by the translated GmADT proteins interacting with GmIFS2 before they are imported into the chloroplasts. Regardless of how the interaction occurs, the results of this study validate GmADTs in the isoflavonoid metabolon. Perhaps GmADTs, by participating in the metabolon, bypass the other antagonist phenylpropanoid pathways to optimize isoflavonoid production in concert with GmIFS2, to drive the flux of shared metabolites into the isoflavonoid branch of the network. This profound interaction extends our knowledge on metabolons producing specialized metabolites beyond our original grasp.

References

Abolhassani Rad, S. (2017) The mystery of nuclear localization of arogenate dehydratases from *Arabidopsis thaliana,* PhD thesis, The University of Western Ontario, London, ON. .

Abolhassani Rad, S., Clayton, E.J., Cornelius, E.J., Howes, T.R. and Kohalmi, S.E. (2018) Moonlighting proteins: putting the spotlight on enzymes. *Plant Signal Behav*, **13**, e1517075.

Airoldi, C.A. and Davies, B. (2012) Gene duplication and the evolution of plant MADS-box transcription factors. *J Genet Genomics*, **39**, 157-165.

Andersson, M.X., Goksor, M. and Sandelius, A.S. (2007) Optical manipulation reveals strong attracting forces at membrane contact sites between endoplasmic reticulum and chloroplasts. *J Biol Chem*, **282**, 1170-1174.

Ayers, A.R., Ebel, J., Finelli, F., Berger, N. and Albersheim, P. (1976) Host-Pathogen Interactions: IX. Quantitative Assays of Elicitor Activity and Characterization of the Elicitor Present in the Extracellular Medium of Cultures of *Phytophthora* megasperma var. *sojae*. *Plant Physiol*, **57**, 751-759.

Barton, K.A., Wozny, M.R., Mathur, N., Jaipargas, E.A. and Mathur, J. (2018) Chloroplast behaviour and interactions with other organelles in *Arabidopsis thaliana* pavement cells. *J Cell Sci*, **131**.

Boydston, R., Paxton, J. and Koeppe, D. (1982) Glyceollin: A Site-Specific Inhibitor of Electron Transport in Isolated Soybean Mitochondria. *Plant Physiol*, **72**, 151-155.

Bross, C.D., Corea, O.R., Kaldis, A., Menassa, R., Bernards, M.A. and Kohalmi, S.E. (2011) Complementation of the pha2 yeast mutant suggests functional differences for arogenate dehydratases from *Arabidopsis thaliana*. *Plant Physiol Biochem*, **49**, 882-890.

Bross, C.D., Howes, T.R., Abolhassani Rad, S., Kljakic, O. and Kohalmi, S.E. (2017) Subcellular localization of Arabidopsis arogenate dehydratases suggests novel and non-enzymatic roles. *J Exp Bot*, **68**, 1425-1440.

Brunkard, J.O., Runkel, A.M. and Zambryski, P.C. (2015) Chloroplasts extend stromules independently and in response to internal redox signals. *Proc Natl Acad Sci U S A*, **112**, 10044-10049.

Cannon, S.B., Mitra, A., Baumgarten, A., Young, N.D. and May, G. (2004) The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol*, **4**:10.

Caplan, J.L., Kumar, A.S., Park, E., Padmanabhan, M.S., Hoban, K., Modla, S., Czymmek, K. and Dinesh-Kumar, S.P. (2015) Chloroplast Stromules Function during Innate Immunity. *Dev Cell*, **34**, 45-57.

Caplan, J.L., Mamillapalli, P., Burch-Smith, T.M., Czymmek, K. and Dinesh-Kumar, S.P. (2008) Chloroplastic protein NRIP1 mediates innate immune receptor recognition of a viral effector. *Cell*, **132**, 449-462.

Chan, K.X., Phua, S.Y., Crisp, P., McQuinn, R. and Pogson, B.J. (2016) Learning the Languages of the Chloroplast: Retrograde Signaling and Beyond. *Annu Rev Plant Biol*, **67**, 25-53.

Chen, Q., Man, C., Li, D., Tan, H., Xie, Y. and Huang, J. (2016a) Arogenate Dehydratase Isoforms Differentially Regulate Anthocyanin Biosynthesis in *Arabidopsis thaliana*. *Mol Plant*, **9**, 1609-1619.

Chen, W., Yao, Q., Patil, G.B., Agarwal, G., Deshmukh, R.K., Lin, L., Wang, B., Wang, Y., Prince, S.J., Song, L., Xu, D., An, Y.C., Valliyodan, B., Varshney, R.K. and Nguyen, H.T. (2016b) Identification and Comparative Analysis of Differential Gene Expression in Soybean Leaf Tissue under Drought and Flooding Stress Revealed by RNA-Seq. *Front Plant Sci*, **7**, 1044.

Cho, M.H., Corea, O.R., Yang, H., Bedgar, D.L., Laskar, D.D., Anterola, A.M., Moog-Anterola, F.A., Hood, R.L., Kohalmi, S.E., Bernards, M.A., Kang, C., Davin, L.B. and Lewis, N.G. (2007) Phenylalanine biosynthesis in *Arabidopsis thaliana*. Identification and characterization of arogenate dehydratases. *J Biol Chem*, **282**, 30827-30835.

Corea, O.R., Bedgar, D.L., Davin, L.B. and Lewis, N.G. (2012a) The arogenate dehydratase gene family: towards understanding differential regulation of carbon flux through phenylalanine into primary versus secondary metabolic pathways. *Phytochemistry*, **82**, 22-37.

Corea, O.R., Ki, C., Cardenas, C.L., Kim, S.J., Brewer, S.E., Patten, A.M., Davin, L.B. and Lewis, N.G. (2012b) Arogenate dehydratase isoenzymes profoundly and differentially modulate carbon flux into lignins. *J Biol Chem*, **287**, 11446-11459.

Dastmalchi, M., Bernards, M.A. and Dhaubhadel, S. (2016) Twin anchors of the soybean isoflavonoid metabolon: evidence for tethering of the complex to the endoplasmic reticulum by IFS and C4H. *Plant J*, **85**, 689-706.

Dastmalchi, M. and Dhaubhadel, S. (2014) Soybean Seed Isoflavonoids: Biosynthesis and Regulation. In *Phytochemicals- Biosynthesis, Function and Application*. Vancouver, British Columbia: Springer, pp. 1-21.

Dastmalchi, M. and Dhaubhadel, S. (2015) Soybean chalcone isomerase: evolution of the fold, and the differential expression and localization of the gene family. *Planta*, **241**, 507-523.

Dopheide, T.A., Crewther, P. and Davidson, B.E. (1972) Chorismate mutase-prephenate dehydratase from *Escherichia coli* K-12. II. Kinetic properties. *J Biol Chem*, **247**, 4447-4452.

Dornfeld, C., Weisberg, A.J., K, C.R., Dudareva, N., Jelesko, J.G. and Maeda, H.A. (2014) Phylobiochemical characterization of class-Ib aspartate/prephenate aminotransferases reveals evolution of the plant arogenate phenylalanine pathway. *Plant Cell*, **26**, 3101-3114.

Edger, P.P. and Pires, J.C. (2009) Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res*, **17**, 699-717.

El-Azaz, J., Canovas, F.M., Avila, C. and de la Torre, F. (2018) The arogenate dehydratase ADT2 is essential for seed development in Arabidopsis. *Plant Cell Physiol*, **doi:10.1093/pcp/pcy200**.

El-Azaz, J., de la Torre, F., Avila, C. and Canovas, F.M. (2016) Identification of a small protein domain present in all plant lineages that confers high prephenate dehydratase activity. *Plant J*, **87**, 215-229.

Emanuelsson, O., Brunak, S., von Heijne, G. and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, **2**, 953-971.

Fester, T., Lohse, S. and Halfmann, K. (2007) "Chromoplast" development in arbuscular mycorrhizal roots. *Phytochemistry*, **68**, 92-100.

Hanada, K., Kuromori, T., Myouga, F., Toyoda, T., Li, W.H. and Shinozaki, K. (2009) Evolutionary persistence of functional compensation by duplicate genes in Arabidopsis. *Genome Biol Evol*, **1**, 409-414.

Hanson, M.R. and Sattarzadeh, A. (2008) Dynamic morphology of plastids and stromules in angiosperm plants. *Plant Cell Environ*, **31**, 646-657.

Heinig, U., Gutensohn, M., Dudareva, N. and Aharoni, A. (2013) The challenges of cellular compartmentalization in plant metabolic engineering. *Curr Opin Biotechnol*, **24**, 239-246.

Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*, **27**, 297-300.

Hossain, M.Z., Ishiga, Y., Yamanaka, N., Ogiso-Tanaka, E. and Yamaoka, Y. (2018) Soybean leaves transcriptomic data dissects the phenylpropanoid pathway genes as a defence response against *Phakopsora pachyrhizi*. *Plant Physiol Biochem*, **132**, 424-433.

Hrazdina, G. and Jensen, R. (1992) Spatial organization of enzymes in plant metabolic pathways. *Annu Rev Plant Biol*, **43**, 241-267.

Hsu, S.K., Lin, L.L., Lo, H.H. and Hsu, W.H. (2004) Mutational analysis of feedback inhibition and catalytic sites of prephenate dehydratase from *Corynebacterium glutamicum*. *Arch Microbiol*, **181**, 237-244.

Huang, T., Tohge, T., Lytovchenko, A., Fernie, A.R. and Jander, G. (2010) Pleiotropic physiological consequences of feedback-insensitive phenylalanine biosynthesis in *Arabidopsis thaliana*. *Plant J*, **63**, 823-835.

Ishida, H., Yoshimoto, K., Izumi, M., Reisen, D., Yano, Y., Makino, A., Ohsumi, Y., Hanson, M.R. and Mae, T. (2008) Mobilization of rubisco and stroma-localized fluorescent proteins of chloroplasts to the vacuole by an ATG gene-dependent autophagic process. *Plant Physiol*, **148**, 142-155.

John L. Giannini, Donald P. Briskin, Jana S. Holt and Paxton, J.D. (1988) Inhibition of Plasma Membrane and Tonoplast HW-Transporting ATPases by Glyceollin. *J. Physiol. Biochem.* **78**, 1000-1003.

Jung, E., Zamir, L.O. and Jensen, R.A. (1986) Chloroplasts of higher plants synthesize L-phenylalanine via L-arogenate. *Proc Natl Acad Sci U S A*, **83**, 7231-7235.

Kafri, R., Springer, M. and Pilpel, Y. (2009) Genetic redundancy: new tricks for old genes. *Cell*, **136**, 389-392.

Kaplan, D.T., Keen, N.T. and Tuomason, I.J. (1980) Studies on the mode of action of glyceollin in soybean Incompatibility to the root knot nematode, Meloidogyne incognita. *Physiol. Plant Path.* **16**, 319-325.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*, **14**, R36.

Kohler, R.H. (1997) Exchange of Protein Molecules Through Connections Between Higher Plant Plastids. *Science*, **276**, 2039-2042.

Krenz, B., Jeske, H. and Kleinow, T. (2012) The induction of stromule formation by a plant DNA-virus in epidermal leaf tissues suggests a novel intra- and intercellular macromolecular trafficking route. *Front Plant Sci*, **3**, 1-21.

Kumar, S., Stecher, G. and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*, **33**, 1870-1874.

Langham, R., Wlash, J., Dunn, M., Ko, C., Goff, S. and Freeling, M. (2004) Genomic Duplication, Fractionation and the Origin of Regulatory Novelty. *Genetics*, **166**, 935-945

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Li, X., Qin, J.C., Wang, Q.Y., Wu, X., Lang, C.Y., Pan, H.Y., Gruber, M.Y. and Gao, M.J. (2011) Metabolic engineering of isoflavone genistein in *Brassica napus* with soybean isoflavone synthase. *Plant Cell Rep*, **30**, 1435-1442.

Liberles, J.S., Thorolfsson, M. and Martinez, A. (2005) Allosteric mechanisms in ACT domain containing enzymes involved in amino acid metabolism. *Amino Acids*, **28**, 1-12.

Maeda, H. and Dudareva, N. (2012) The shikimate pathway and aromatic amino Acid biosynthesis in plants. *Annu Rev Plant Biol*, **63**, 73-105.

Maeda, H., Shasany, A.K., Schnepp, J., Orlova, I., Taguchi, G., Cooper, B.R., Rhodes, D., Pichersky, E. and Dudareva, N. (2010) RNAi suppression of Arogenate Dehydratase1 reveals that phenylalanine is synthesized predominantly via the arogenate pathway in petunia petals. *Plant Cell*, **22**, 832-849.

Messina, M. (2010) Insights gained from 20 years of soy research. *J Nutr*, **140**, 2289S-2295S.

Neve, E.P. and Ingelman-Sundberg, M. (2008) Intracellular transport and localization of microsomal cytochrome P450. *Anal Bioanal Chem*, **392**, 1075-1084.

Pohnert, G., Zhang, S., Husain, A., Wilson, D.B. and Ganem, B. (1999) Regulation of phenylalanine biosynthesis. Studies on the mechanism of phenylalanine binding and feedback inhibition in the *Escherichia coli* P-protein. *Biochemistry*, **38**, 12212-12217.

Pucciariello, C., Voesenek, L.A., Perata, P. and Sasidharan, R. (2014) Plant responses to flooding. *Front Plant Sci*, **5**, 1-2.

Ralston, L., Subramanian, S., Matsuno, M. and Yu, O. (2005) Partial reconstruction of flavonoid and isoflavonoid biosynthesis in yeast using soybean type I and type II chalcone isomerases. *Plant Physiol*, **137**, 1375-1388.

Rippert, P., Puyaubert, J., Grisollet, D., Derrier, L. and Matringe, M. (2009) Tyrosine and phenylalanine are synthesized within the plastids in Arabidopsis. *Plant Physiol*, **149**, 1251-1260.

Schattat, M., Barton, K., Baudisch, B., Klosgen, R.B. and Mathur, J. (2011) Plastid stromule branching coincides with contiguous endoplasmic reticulum dynamics. *Plant Physiol*, **155**, 1667-1677.

Schattat, M.H., Griffiths, S., Mathur, N., Barton, K., Wozny, M.R., Dunn, N., Greenwood, J.S. and Mathur, J. (2012) Differential coloring reveals that plastids do not form networks for exchanging macromolecules. *Plant Cell*, **24**, 1465-1477.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178-183.

Sepiol, C.J., Yu, J. and Dhaubhadel, S. (2017) Genome-Wide Identification of Chalcone Reductase Gene Family in Soybean: Insight into Root-Specific GmCHRs and *Phytophthora sojae* Resistance. *Front Plant Sci*, **8**, 1-15.

Siehl, D.L. and Conn, E.E. (1988) Kinetic and regulatory properties of arogenate dehydratase in seedlings of *Sorghum bicolor* (L.) Moench. *Arch Biochem Biophys*, **260**, 822-829.

Smith, M. (2014) Changing the substrate specificity of arogenate dehydratases (ADTs) from Arabidopsis thaliana. M.Sc. thesis, The University of Western Ontario, London, ON.

Stenmark, S., Pierson, D. and Jensen, R. (1974) Blue-green bacteria synthesize L-tyrosine by the pretyrosine pathway. *Nature*, **247**, 290-292.

Styranko, D. (2011) Characterizing Arogenate Dehydratase dimers. M.Sc. thesis, The University of Western Ontario, London, ON.

Subramanian, S., Graham, M.Y., Yu, O. and Graham, T.L. (2005) RNA interference of soybean isoflavone synthase genes leads to silencing in tissues distal to the transformation site and to enhanced susceptibility to *Phytophthora sojae. Plant Physiol*, **137**, 1345-1353.

Subramanian, S., Stacey, G. and Yu, O. (2006) Endogenous isoflavones are essential for the establishment of symbiosis between soybean and *Bradyrhizobium japonicum*. *Plant J*, **48**, 261-273.

Sukumaran, A., McDowell, T., Chen, L., Renaud, J. and Dhaubhadel, S. (2018) Isoflavonoid-specific prenyltransferase gene family in soybean: GmPT01, a

pterocarpan 2-dimethylallyltransferase involved in glyceollin biosynthesis. *Plant J.* **doi: 10.1111/tpj.14083**

Tan, K., Li, H., Zhang, R., Gu, M., Clancy, S.T. and Joachimiak, A. (2008) Structures of open (R) and close (T) states of prephenate dehydratase (PDT)--implication of allosteric regulation by L-phenylalanine. *J Struct Biol*, **162**, 94-107.

Todd, J.J. and Vodkin, L.O. (1996) Duplications That Suppress and Deletions That Restore Expression from a Chalcone Synthase Multigene Family. *The Plant Cell*, **8**, 687-699.

Tyler, B.M. (2007) Phytophthora sojae: root rot pathogen of soybean and model oomycete. *Mol Plant Pathol*, **8**, 1-8.

Vogt, T. (2010) Phenylpropanoid Biosynthesis. *Mol Plant,* **3**, 2-20.

Wang, Z. and Benning, C. (2012) Chloroplast lipid synthesis and lipid trafficking through ER-plastid membrane contact sites. *Biochem Soc Trans*, **40**, 457-463.

Whatley, J.M., McLean, B. and Juniper, B.E. (1990) Continuity of chloroplast and endoplasmic membranes in *Phaseolus vulgaris*. *New Phytology*, **117**, 209-217.

Xia, T.H., Ahmad, S., Zhao, G.S. and Jensen, R.A. (1991) A single cyclohexadienyl dehydratase specifies the prephenate dehydratase and arogenate dehydratase components of one of two independent pathways to L-phenylalanine in *Erwinia herbicola*. *Arch Biochem Biophys*, **286**, 461-465.

Yamada, T., Matsuda, F., Kasai, K., Fukuoka, S., Kitamura, K., Tozawa, Y., Miyagawa, H. and Wakasa, K. (2008) Mutation of a rice gene encoding a phenylalanine biosynthetic enzyme results in accumulation of phenylalanine and tryptophan. *Plant Cell*, **20**, 1316-1329.

Yanagihara, K., Ito, A., Toge, T. and Numoto, M. (1993) Antiproliferative effects of isoflavones on human cancer cell lines established from the gastrointestinal tract. *Cancer Res*, **53**, 5815-5821.

Yoo, H., Widhalm, J.R., Qian, Y., Maeda, H., Cooper, B.R., Jannasch, A.S., Gonda, I., Lewinsohn, E., Rhodes, D. and Dudareva, N. (2013) An alternative pathway contributes to phenylalanine biosynthesis in plants via a cytosolic tyrosine:phenylpyruvate aminotransferase. *Nat Commun*, **4**:2833.

Zabala, G., Zou, J., Tuteja, J., Gonzalez, D.O., Clough, S.J. and Vodkin, L.O. (2006) Transcriptome changes in the phenylpropanoid pathway of *Glycine max* in response to Pseudomonas syringae infection. *BMC Plant Biol*, **6**:26.

Zernova, O.V., Lygin, A.V., Pawlowski, M.L., Hill, C.B., Hartman, G.L., Widholm, J.M. and Lozovaya, V.V. (2014) Regulation of plant immunity through modulation of phytoalexin synthesis. *Molecules*, **19**, 7480-7496.

Zhang, S., Pohnert, G., Kongsaeree, P., Wilson, D.B., Clardy, J. and Ganem, B. (1998) Chorismate mutase-prephenate dehydratase from *Escherichia coli.* Study of catalytic and regulatory domains using genetically engineered proteins. *J Biol Chem*, **273**, 6248-6253.

Zhang, S., Wilson, D.B. and Ganem, B. (2000) Probing the catalytic mechanism of prephenate dehydratase by site-directed mutagenesis of the *Escherichia coli* P-protein dehydratase domain. *Biochemistry*, **39**, 4722-4728.

Zhao, G., Xia, T., Fischer, R.S. and Jensen, R.A. (1991) Cyclohexadienyl Dehydratase from *Pseudomonas aeruginosa*. *J. Biol. Chem.*, **267**, 2487-2493.

Appendices

Continued…

Appendix A| Multiple sequence alignment of mature proteins of GmADTs with characterized plant ADTs. Critical residues and/or motifs identified by the respective studies to be critical for enzyme function or regulation are shown by colored boxes: Hsu *et al* 2004, ▯, important for catalytic activity in bacterial PDTs; Zhang *et al.* 2000, ▯, important for catalytic function; El-Azaz *et al.,* 2016, ▯, important for conferring PDT activity in plant lineages. Specifically, the PDT-conferring residue is at a single Alanine residue, Ala205 (▮) (in reference to PpADT-G); Smith 2014, ▯, a Phe215Leu substitution (in reference to AtADT5) that results in PDT activity to non-PDT AtADTs. Pohnart *et al.* 1999, ▯, important in phe binding and allosteric regulation, including the ESRP and GALV domains. ▮ indicates single residue substitutions within individual GmADT sequences. ▮ indicates the additional residues in the ACT domain of GmADT13B that may be involved in binding to another protein that is nuclear localized.

Appendix B| Phylogenetic relationship between GmADTs and plant and bacterial ADT/PDT sequences. A neighbour-joining tree was constructed of the mature protein sequences of GmADTs, AtADTs, PpADTs and PhADTs, along with mature protein sequences of annotated ADT sequences of other plant and algae species and PDT sequences of bacterial species. 1000 bootstrap replications were used, with the percentage of replicate trees in which the respective taxa clustered together shown. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. Black circles represent the 9 GmADTs and white circles represent the characterized AtADTs. Labels indicate the subgroups identified in *Arabidopsis* that the GmADTs fall into. *Phaseolus vulgaris* sequences are identified by their respective Phytozome genome annotation units. *Gm, Glycine max; At, Arabidopsis thaliana;* Pp, *Pinus pinaster*; Ph=, *Petunia hybrida.* Accession numbers for those not listed in tree: PhADT1, ACY79502.1; PhADT2, ACY79503.1; PhADT3, ACY79504.1; PpADTA, APA32582.1; PpADTB, APA32583.1; PpADTC, APA32584.1; PpADTD, APA32585.1; PpADTE, APA32586.1; PpADTF, APA32587.1; PpADTG, APA32588.1; PpADTH, APA32589.1; PpADTI, APA32590.1; AtADT1, OAP11955.1; AtADT2, OAP02204.1; AtADT3, ABD67752; AtADT4, ABD67753.1; AtADT5, ABD67754.1; AtADT6, OAP14989.1; *E. coli,* WP_115444483.1; *Salmonella enterica,* WP_094894131.1; *Staphylococcus aureus,* WP_050961432.1.
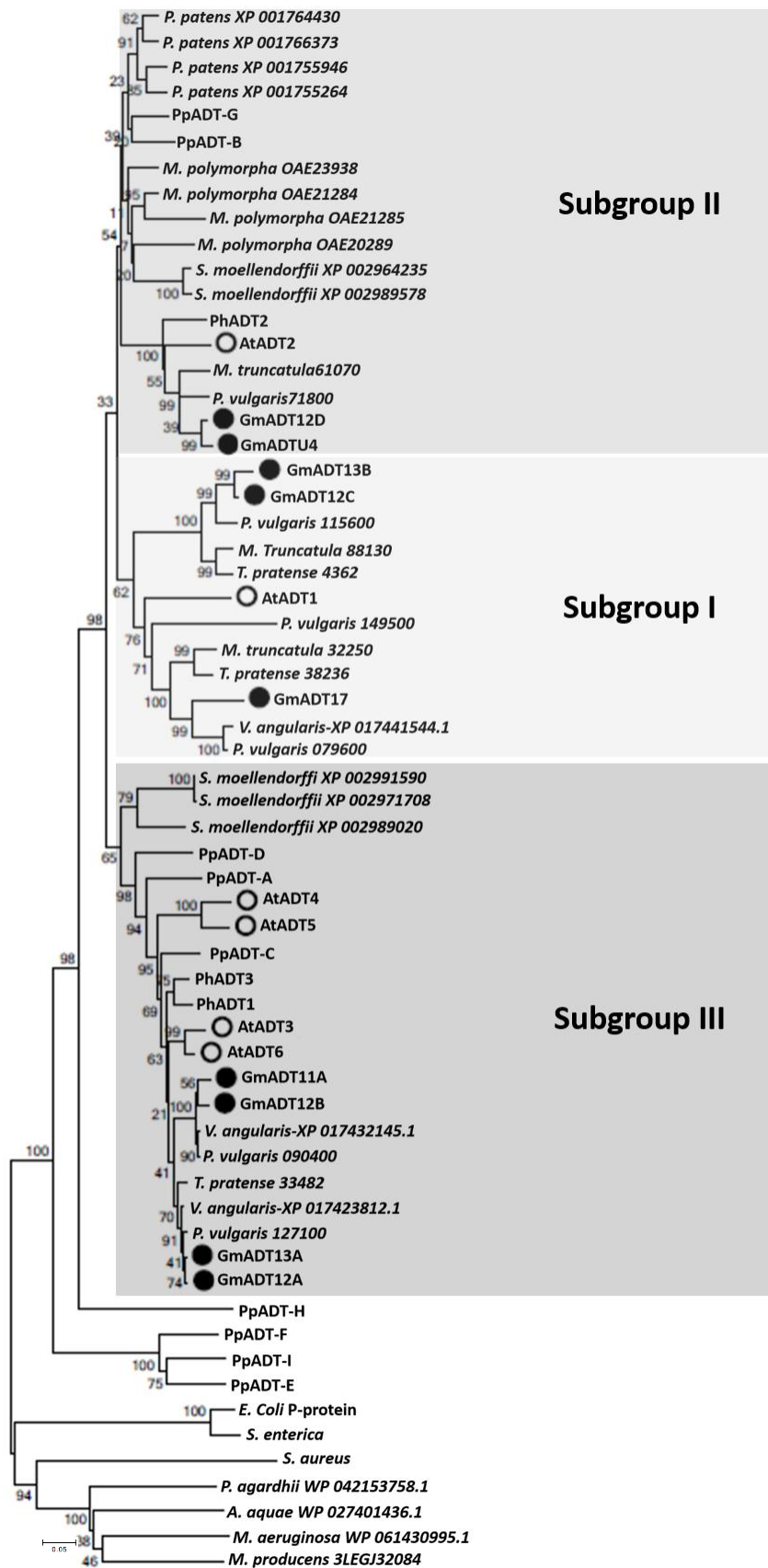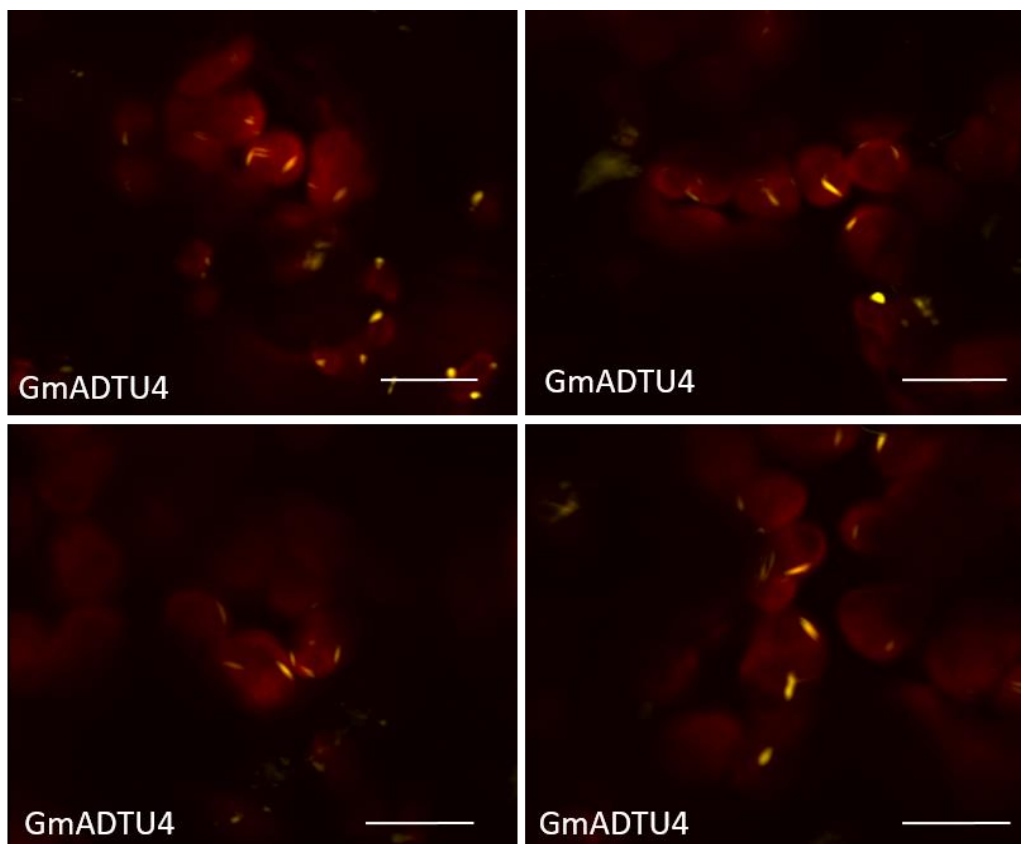
Appendix C| Close-up images of the subcellular localization of GmADTU4. Additional images of GmADTU4 fluorescence are shown. Scale bars represent 30 µM. A schematic diagram showing the stages of chloroplast division is shown beneath the images. Chloroplasts are shown in red, and the ring structures of proteins involved in division are shown in yellow (representing YFP where GmADTU4 potentially aids in division).

Appendix D| Specific element promoter count and description

Category	Motif Name	Description	13A	12A	11A	U4	12B	17A	12C	12D	13B
Root/Nodulation	WUSATAg	expressed in cells of the root apical meristem	0	0	0	1	1	0	1	0	0
	ROOTMOTIFTAPOX1	root hair growth	16	9	10	12	12	24	13	13	8
	OSE1ROOTNODULE	consensus sequence of promoters activated in infected root cells	3	4	1	2	2	0	4	2	2
	NODCON1GM	nodulation	3	4	1	2	2	0	4	2	2
	NODCON2GM	nodulation	1	1	4	1	4	3	1	4	2
	RAV1AAT	root-specific promoter	1	2	3	2	2	3	1	0	1
	RHERPATEXPA7	root-hair specific	0	1	0	0	0	0	0	1	0
Phenylpropanoid Pathway-Related	BOXLCOREDCPAL	transcriptional activator of PAL enzyme	0	2	0	0	3	1	0	0	0
	EBOXBNNAPA:	tissue-specific, light-activated element of phenylpropanoid-pathway genes	6	6	6	6	0	4	2	0	4
	MYBPLANT	consensus sequence of promoters in phenylpropanoid-pathway genes	0	0	0	0	3	1	0	0	0
	PALBOXLPC	consensus sequence of promoter element found in PAL enzyme	0	1	0	0	0	0	0	0	0
	MYBPZM	core promoter element gene that activates flavonoid genes	0	1	0	1	4	1	1	0	0
	MYBCORE	binding site for two myb proteins involved in regulation of flavonoid genes	3	1	1	4	0	5	3	1	4
Abiotic Stress	PREATPRODH	cis-element for proline and hypoosmolarity induced expression	1	0	0	0	0	1	0	1	0
	MYCCONSENSUSAT	MYC recognition site found in the promoters of the dehydration-responsive gene	6	6	6	6	0	4	2	0	4
	ACGTATERD1	required for early response to dehydration gene	4	6	2	4	2	2	0	6	2
	GT1GMSCAM4	Plays a role in pathogen- and salt-induced SCaM gene expression	5	7	6	2	5	2	0	3	1
	AMMORESIIUDCRNIA1	Involved in ammonium-response	0	0	0	1	0	0	0	0	0
	LTRECOREATCOR15	Core of low temperature responsive element (LTRE)	1	0	0	1	0	0	0	1	1
	SURECOREATSULTR11	Core of sulfur-responsive element (SURE)	1	2	3	2	0	2	0	0	1
	ACGTATERD1	required for early response to dehydration gene	4	6	2	4	2	2	0	6	2
	EECCRCAH1	responsive element to low carbon dioxide	1	0	1	1	1	4	1	0	0
	TAAAGSTKST1	Target site for Dof1 protein controlling guard cell specific gene expression	1	5	6	2	3	1	3	3	0
	ANAERO1CONSENSUS	promoter element of anearobic genes involved in fermentive pathway	3	3	2	4	0	1	0	2	0
	ANAERO2CONSENSUS	promoter element of anearobic genes involved in fermentive pathway	1	0	0	0	0	0	0	0	2
	CURECORECR	copper responsive element	4	2	2	2	4	4	0	8	6
	MYB1AT	MYB recognition site found in dehydration responsive elements	0	1	2	1	1	2	2	1	5
	SREATMSD	sugar-responsive element	1	0	2	0	0	0	1	2	0
	QARBNEXTA	response to wounding and tensile stress	0	0	0	1	0	1	0	0	0
	HBOXCONSENSUSPVCHS	response to wounding and tensile stress	0	0	0	0	1	0	1	0	0
Hormone-Induced	LECPLEACS2	elicits ethylene biosynthesis	2	1	0	1	0	3	1	2	1
	ERELEE4	ethylene-responsive element	0	1	0	0	1	0	0	1	0
	ARFAT	ARF (auxin response factor) binding site	1	2	1	1	0	2	0	0	1
	ASF1MOTIFCAMV	transcriptional activation of several genes by auxin and/or salicylic acid	1	1	0	0	2	0	0	0	1
	T/GBOXATPIN2	involved in jasmonate-induction of genes	0	0	1	1	0	1	0	1	0

Continued…

Appendix D continued| Specific element promoter count and description

Category	Motif Name	Description	13A	12A	11A	U4	12B	17A	12C	12D	13B
	ASF1MOTIFCAMV	transcriptional activation of several genes by auxin and/or salicylic acid	1	1	0	0	2	0	0	0	1
Pathogen Response	NTBBF1ARROLB	auxin induction of an oncogene	0	2	3	1	0	0	2	1	0
	WBOXNTERF3	Activation of ERF gene in response to wounding	0	1	3	5	1	4	5	2	3
	ARR1AT	Response to bacterial infection	11	16	8	10	13	11	14	9	10
	BIHD1OS	binding site of transcription involved in disease resistance response	0	4	1	2	2	3	1	2	2
	WBOXATNPR1	"W-box" found in promoter of Arabidopsis thaliana (A.t.) NPR1 gene	1	1	1	4	2	4	3	0	2
	WRKY71OS	W box elements within the pathogenesis-related class10 genes	1	6	4	8	5	7	7	4	6
	ELRECOREPCRP1	involved in WRKY activation of pathogenesis-related genes	0	0	0	2	1	1	0	1	1
Light Related (stress and non-stress)	PRECONSCRHSP70A	involved in light-induction of plastid-response genes	0	0	1	1	0	1	0	0	0
	GATABOX	light-induced expression of a chlorophyll-binding protein	13	9	8	6	10	5	9	8	9
	IBOX	conserved promoter sequence in light-respone genes	10	5	9	1	7	1	4	4	5
	GT1CONSENSUS	consensus sequence in light-response genes	22	18	13	9	9	7	3	10	12
	-10PEHVPSBD	involved in chloroplast-gene expression	1	3	2	1	1	4	1	2	0
	SORLIP5AT	consensus sequence in light-induced promoter	0	0	0	0	0	0	0	0	1
	INRNTPSADB	light-responsive transcription	4	2	5	0	2	6	1	7	3
	MARTBOX	light-responsive element	10	0	4	5	4	4	2	5	3
	CCA1ATLHCB1	myb transcription factor involved I phytochrome regulation	2	4	0	0	1	2	0	0	0
	CPBCSPOR	light-response in a chlorophyll-related genes	0	1	2	0	1	1	1	1	1
	HDZIP2ATATHB2	shade avoidance responses	0	0	0	0	2	1	2	1	0

Continued…

Appendix D continued| Specific element promoter count and description

Category		Motif Name	Description	13A	12A	11A	U4	12B	17A	12C	12D	13B
Development/Reproduction	**Seed**	CAATBOX1	legumin gene promoter consensus sequence	12	16	18	11	18	25	20	11	7
		SEF4MOTIFGM7S	element in a seed storage protein	7	1	5	7	6	6	3	2	2
		SEF1MOTIF	element in a seed storage protein	0	3	0	1	2	2	4	1	1
		AACACOREOSGLUB1	endosperm-specific expression of glutein genes	0	0	0	2	0	0	0	1	0
		NAPINMOTIFBN	upstream sequence in napin gene in seeds	1	1	0	0	0	1	0	0	0
		RYREPEATLEGUMINBOX	promoter sequence of legumin seed-storage proteins	0	0	0	0	0	1	1	0	1
		RYREPEATGMGY2	upstream sequence of glycinin genes	0	0	0	0	0	1	1	0	1
		RYREPEATBNNAPA	upstream sequence in the napA gene	0	0	0	0	0	1	1	0	1
		CEREGLUBOX2PSLEGA	cereal glutenin box in legumin gene	0	0	0	0	1	0	0	0	1
		SEF3MOTIFGM	binding sequence of the soybean embryo factor	1	0	0	1	1	1	1	0	0
		WBOXHVISO1	Endosperm-specific expression	1	1	3	3	0	3	3	1	2
	Pollen	POLLEN1LELAT52	pollen-specific activation	9	13	9	3	11	10	9	10	5
		QELEMENTZMZM13	pollen-specific promoter	0	0	0	2	0	0	1	0	1
		GTGANTG10	pollen-specific activation	9	10	5	4	1	7	3	5	8
	Gibberellin Response	TATCCAOSAMY	gibberlelin-mediated regulation of alpha-amylase gene	1	0	1	0	0	0	1	1	2
		GAREAT	gibberelin-responsive element	1	1	1	2	2	0	2	1	1
		CAREOSREP1	gibberellin-upregulated proteinase expression in seeds	1	0	0	0	0	0	0	0	1
		CARGCW8GAT	binding site for AGL15 (AGAMOUS-like 15)	2	0	2	4	2	2	2	0	2
		PYRIMIDINEBOXOSRAMY1A	gibberlelin-mediated regulation of alpha-amylase gene	0	2	1	3	1	0	0	3	1
		PYRIMIDINEBOXHVEPB1	gibberellin-upregulated proteinase expression in seeds	1	0	0	1	0	0	0	0	1
		AMYBOX2	upstream region of alpha-amylase gene	0	0	1	0	0	0	0	1	1
		AMYBOX1	upstream region of alpha-amylase gene	1	1	0	1	1	1	2	1	1

Appendix E| Reciprocal BiFC combinations (GmADT-YN + GmIFS2-YC) and the corresponding negative control (GmADT-YN + empty YC)
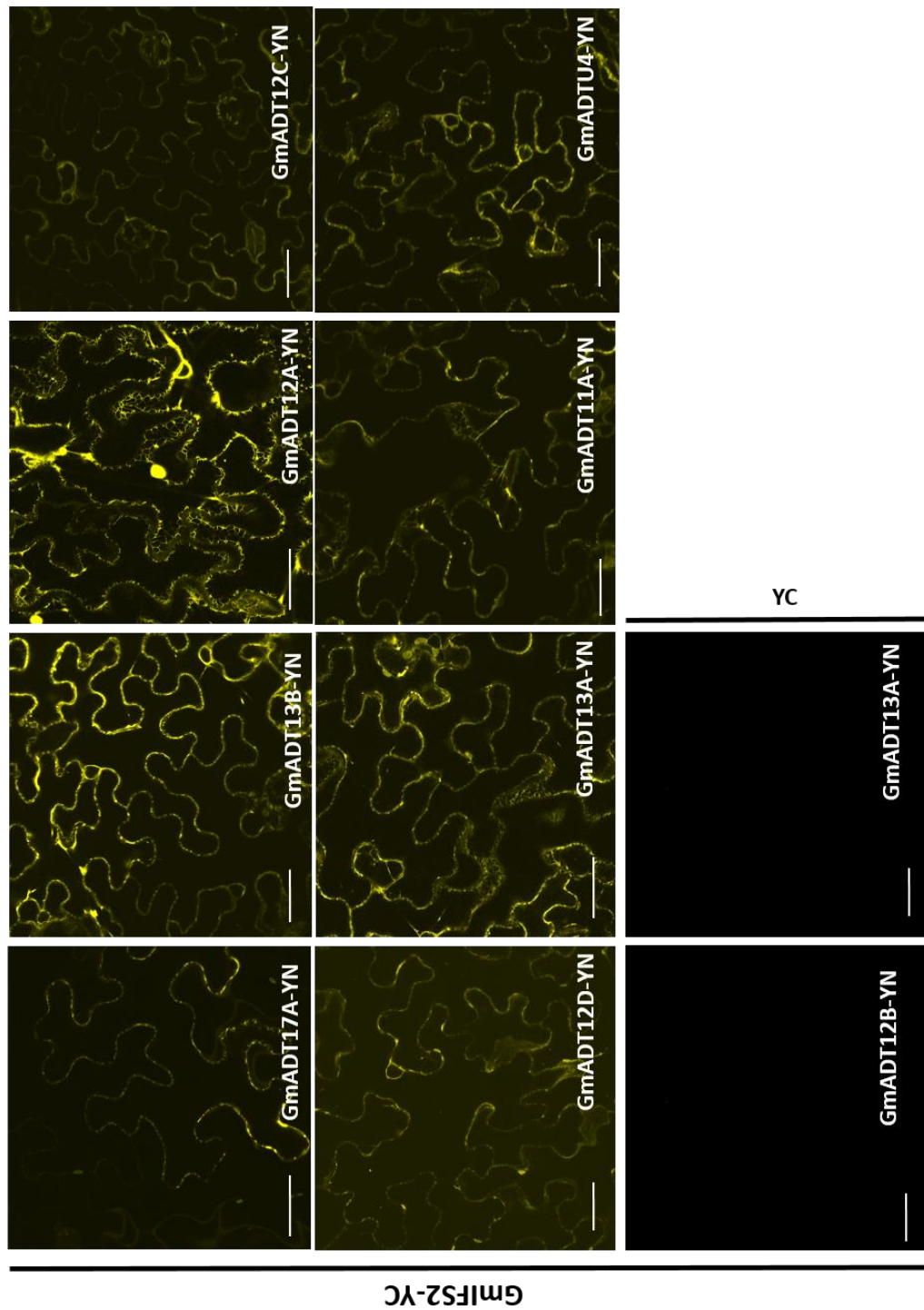
Curriculum Vitae

KELSEY PANNUNZIO, BSc, MSc Candidate

EDUCATION

MSc Candidate: Department of Biology (Cell and Molecular Biology)
The University of Western Ontario, London, ON; 2016-2018
Bsc: Double Major in Biochemistry and Biology
The University of Western Ontario, London, ON; 2011-2016

WORK EXPERIENCE

Graduate Research Assistant/ Research Affiliate
Agriculture and Agri-Foods Canada and Western University; 2016-2018
Teaching Assistant
Biology 2601A: Organismal Physiology, Fall Term; 2018
Biology 2485: Environmental Biology, Winter Terms; 2017, 2018
Biology 2483: Ecology, Fall Term; 2017
Biology 1001: Biology for Science, Fall Term; 2016
Lab Assistant Technician
Pollution Control Centre- Municipality of Leamington, Leamington, ON; May-August
2013-2016

VOLUNTEER EXPERIENCE

Community Outreach Volunteer
Agriculture and Agri-Foods Canada, London ON; 2017
Lab Assistant
Biochemistry Department, Western University, London, ON; 2013-2015
Chronic Pain Clinic
St. Joseph Hospital, London, ON; 2015-2016

AWARDS

-Western Graduate Research Scholarship (Western University, London, ON); 2016-2018
-Biology Graduate Student Travel Award (Western University, London, ON); 2018

RESEARCH PRESENTATIONS

Pannunzio, K. Uncovering a mystery of the isoflavonoid metabolon: identification and characterization of the *AROGENATE DEHYDRATASE* gene family in soybean. Graduate Research Forum, October 18-19, 2018, London, ON. *(Oral Presentation)*
Pannunzio, K., Dhaubhadel, S. identification and characterization of the *AROGENATE DEHYDRATASE* gene family in soybean. Plant Biology, July 14-18, 2018, Montreal, QC. *(Poster Presentation)*