A functional genomics approach in identifying the underlying gene for the E8 maturity locus in soybean (*Glycine max*)

by

Michael Sadowski

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Biology

in

Biology

Carleton University
Ottawa, Ontario

Abstract

Soybean is an economically important crop that has rapidly expanded throughout Western Canada and Northern regions. To continue this expansion, understanding the time of flowering and maturity pathway is an important factor for soybean adaptation. So far, eleven maturity loci have been identified for this pathway, however the underlying gene for one third of them remain unknown. The E8 maturity locus was previously identified in our lab on chromosome 4 using classical breeding practices and genome wide SSR marker analysis. A bioinformatics approach utilizing PIPE (Protein-protein Interaction Prediction Engine) along with a plethora of functional genomics resources and prediction tools has short listed this region down to 3 promising candidates; *Glyma.04G124600, Glyma.04G140000,* and *Glyma.04G101500*, all involved in light perception. Further analysis of these candidates will reveal the underlying gene for E8 and shed light on the flowering mechanism in the important food crop soybean.

Acknowledgements

First and foremost, I would like to thank my direct supervisor Dr. Bahram Samanfar for his mentorship, patience and support. I would like to thank all the Samanfar lab members; Doris for her amazing desserts, Martin for always bringing me along on his lunch breaks, and collectively, for their guidance in every step of this project. I would also like to thank Dr. Elroy Cober for his soybean breeding expertise and providing the cultivars necessary to perform this project.

In addition, a thank you my co-supervisor, Dr. Ashkan Golshani, along with committee members Dr. Shelley Hepworth and Dr. Leonid Savitch, for their insightful comments and guidance throughout this project.

Table of Contents

List of Tables

List of Illustrations

List of Appendices

List of Abbreviations

Abbreviation	Definition
AFLP	Amplified Fragment Length Polymorphism
AM	Association Mapping
AP2	APATELA2
CCA1	Circadian Clock Associated 1
CDF1	Cycling Dof Factor 1
CDS	Coding Sequence
CO	Constans
CRY1	Cryptochrome 1
DArT	Diversity Array Technology
DEL	Deletion
ELF4	Early-Flowering 4
EST	Expressed Sequence Tags
FAR1	Far-Red Impared Responsive 1
FHL	Far Elongated Hypocotyls Like
FHY3	Far-Red Elongated Hypocotyls 3
FKF1	Kelch Repeat, F-Box1
FT	Flowering Locus T
GAI	Gibberellic Acid Insensitive
GBS	Genotype By Sequencing
GI	Gigantea
GO	Gene Ontology
GWAS	Genome Wide Association Studies
HCMV	Human Cytomegalovirus
Hd1	Heading Date 1
Hd3a	Heading Date 3A
InDel	Insertions/Deletions
INS	Insertion
ISSR	Inter Simple Sequence Repeat
LD	Long-Day
LHY	Late Elongated Hypocotyl
LOF	Loss-Of-Function
LRR-RLK	Leucine-Rich Receptor Like Kinase
MAS	Marker Assisted Selection
MG	Maturity Group
MULE	Mutator-Like Element
NTC	No Template Control
PHYA	Phytochrome A
PHYB	Phytochrome B

PIPE	Protein-Protein Interaction Prediction Engine
POI	Protein Of Interest
PPI	Protein-Protein Interaction
PROVEAN	Protein Variation Effect Analyzer
PRR	Pseudo Response Regulator
QTL	Quantitative Trait Locus
RFLP	Restriction Fragment Length Polymorphism
RGA	Repressor Of Ga1-3
RLK	Receptor-Like Kinase
SAM	Shoot Apical Meristem
SCL3	Scarecrow-Like 3
SCR	Scarecrow
SD	Short-Day
SNP	Single Nucleotide Polymorphism
SSR	Simple Sequence Repeat
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TAP	Tandem Affinity Purification
TEM	Tempranillo
UTR	Untranslated Region
Y2H	Yeast Two-Hybrid

Chapter 1: Introduction

Soybean [*Glycine max* (L.) Merr.] is a short-day (SD) flowering plant that has a wide array of uses across many industries making it a highly valuable crop, both locally and globally. Soybean is grown for both human and animal consumption due to its high protein and oil content, about 20% and 40% respectively, while also providing many of the essential amino acids required for human diets [1]. For this reason, soybean is an excellent source to substitute meat and dairy and can provide a fat-free meal for animal feed. With increasing demand for more environmentally friendly products, soy oil has potential to replace petroleum in engines and lubricants and has already been utilized in the production of Canadian biofuels. Additionally, soybean promotes sustainable agriculture management practices as a key crop rotation partner due to its ability to naturally fix atmospheric nitrogen back into the soil [2]. With the multitude of uses for soybean there is major potential for future growth and expansion within Canada.

1.1 Soybean in Canada

Canada is the seventh largest global producer of soybean at approximately 6-7 million metric tons annually [3]. Within Canada, wheat is the leading principle field crop by area with 24.5 million acres, canola is the second crop with 20.9 million acres, soybean is the third at 5.7 million acres and fourth is corn for grain at 3.7 million acres, based on 2019 statistics from Statistics Canada [4]. Soybean originated (domesticated) from *Glycine soja*, a wild relative, in China 3,000 – 5,000 years ago and has only recently (~200 years ago) made its way into North America. Initially soybean production in Canada was limited

to southern Ontario, however in the mid-1970's soybean breeding programs developed early maturing lines that were successful throughout more Northern regions of the province [2], [5]. As the understanding of the genetics underlying time of flowering and maturity grew, so did the development of ultra-early maturing soybean lines and has thus resulted in the rapid expansion of soybean, mainly throughout Western Canada.

In 2000, Ontario contributed to approximately 85% of seeded soybean area, 2.25 million acres, with the rest of the land shared between Quebec and the Maritimes. The acreage in Ontario has relatively stayed the same, with fluctuations year to year, however seeded area throughout Canada has gone up to 5.7 million acres. Outlined in Figure 1, the principle crops of Canada can be visualized by province, with Ontario holding about 55 % of seeded soybean area in Canada, ~17% in Quebec, 23% in Manitoba, 2.7% in Saskatchewan, and 0.4% in Alberta, leaving the rest for the Maritimes [6]. Although the numbers for soybean seeded area in Western Canada are not relatively high, they were essentially nonexistent 20 years ago. Manitoba numbers were not recorded until 2001, Saskatchewan in 2013 and Alberta most recently in 2018 [4]. SoyCanada targets soybean seeded area in Western Canada to reach 6 million acres by 2027, about a 70% increase from current standings [7]. The opportunity to rapidly increase soybean production in Canada relies largely on the west due to their large availability of fertile land, which would also meet their need for sustainable crop managements and rotation practices. To meet the constraints of expanding soybean into more northern regions of Western Canada, research must be strengthened to further investigate the genetics underlying economically important pathways.

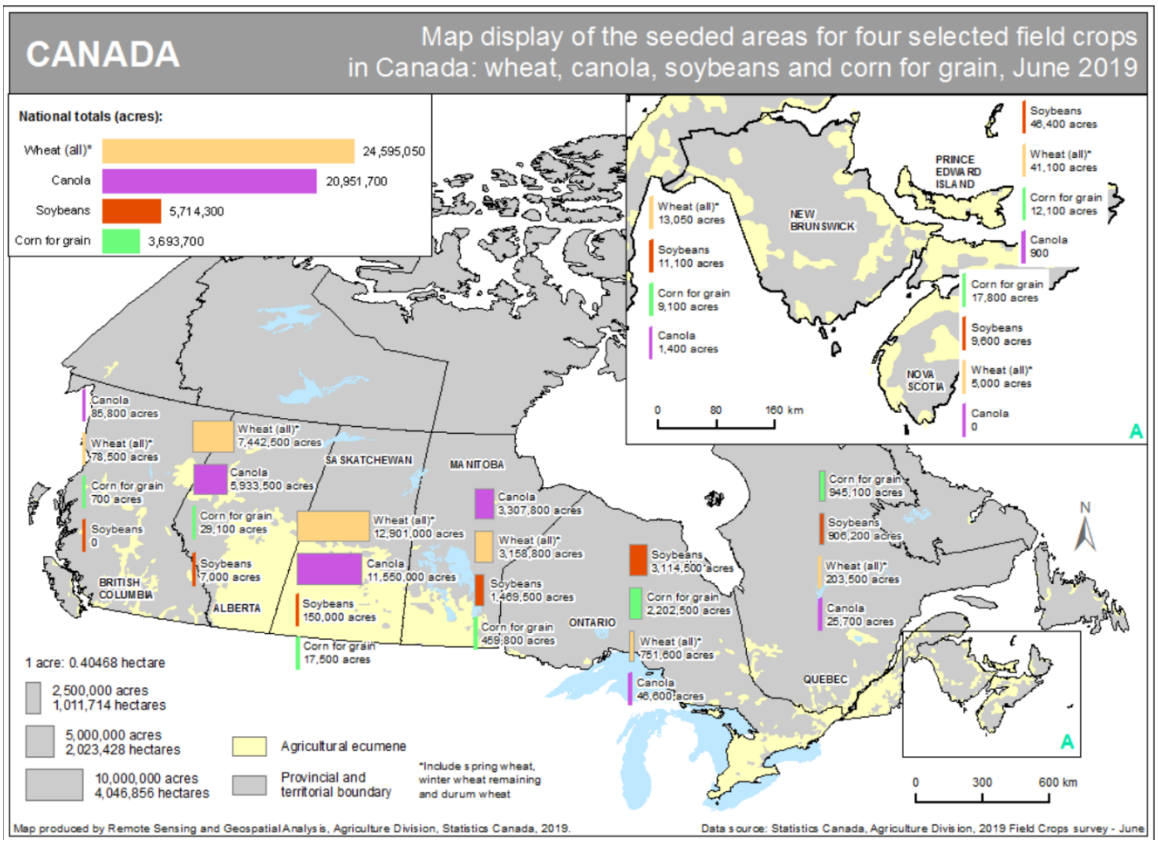


Figure 1: Map of Canada showing seeded area (in acres) by province for wheat, canola, soybean, and corn for grain. Adopted from Statistics Canada, a 2019 report [5].

1.2 Challenges for Western Canada

Although soybean has been well adapted to diverse ecological conditions for optimal growth worldwide, the desire to grow further north in Canada has many complicating factors such as longer days, shorter growing seasons, water stress, pests, diseases, and nutrient deficiencies such as iron [5], [8]. Among these challenges, overcoming the ability to grow soybean in long-day environments with short growing seasons is arguably the most important. Considering soybean is a short-day (SD) crop, flowering and ultimately time of maturity, become delayed when exposed to longer days and therefore cannot reach maturity before first frost. The genetics underlying time of

flowering and maturity hold the potential to increase this production, however there are many gaps of knowledge in understanding the key players that govern the control of this pathway [9]. Therefore, pushing soybean into more Northern regions of Canada is highly dependent on research towards better understanding the mechanisms underlying time of flowering and maturity, and integrating this knowledge into breeding programs where further elite varieties (early maturing and/or ultra-early maturing) can be developed.

1.3 Growth and development

Adapting a plant to regions that are widely different from their initial origin requires an extensive consideration of the underlying genetic pathways. Just as humans experience jet lag from crossing continents, due to circadian rhythm disorders, plants begin to lose or repress many key functions required for reproductive success [10]. Thus, understanding the growth and development of a species and the associated environmental and genetic factors carries important implications to the overall success of a plant.

The life cycle of soybean can be separated into three stages, vegetative stage, reproductive stage, and senescence (maturity). During the vegetative stage the cotyledons, visualized in Figure 2*,* germinate and emerge from the soil. This is known as the VE (vegetative emergence) stage of the growth cycle. Once cotyledons become exposed the hypocotyl begins to elongate and eventually produces unifoliate leaves and a root system [11], [12]. The plant is considered to be in VC (vegetative cotyledons) stage when unifoliate leaves are fully expanded. As the main stem continues to grow, the first trifoliate leaf begins to form, once fully expanded this stage is denoted as the V1 stage. All vegetative stages following VC are designated as Vn, where n can be determined by counting the number of nodes (where leaf stem attaches to the main stem), at the fully expanded trifoliate

leaves (Figure 2) [12]. The plant continues growth of its main stem and development of trifoliate leaves up until about the V5 stage where the transition to reproductive development begins via first flowers. The number of vegetative nodes along the main stem and the time to produce them is influenced by many factors such as, temperature, photoperiod and maturity genes. These factors also influence the morphology of the plant during vegetative growth [11].

Time of flowering and maturity, based on these various factors throughout vegetative development along with allelic variation at the soybean maturity loci, classifies soybean into maturity group [9]. These maturity groups indicate optimized soybean cultivars designed for their geographic region that can output the highest yield for that particular region. North American varieties are compiled into 13 maturity groups ranging from Southern Canada (MG 000, highest latitude) to Mexico and the Caribbean Islands (MG X, lowest latitude). Cultivars within each of these maturity groups have a narrow range of latitudes (150-250 km), signifying the importance of photoperiod on soybean growth and development [9].

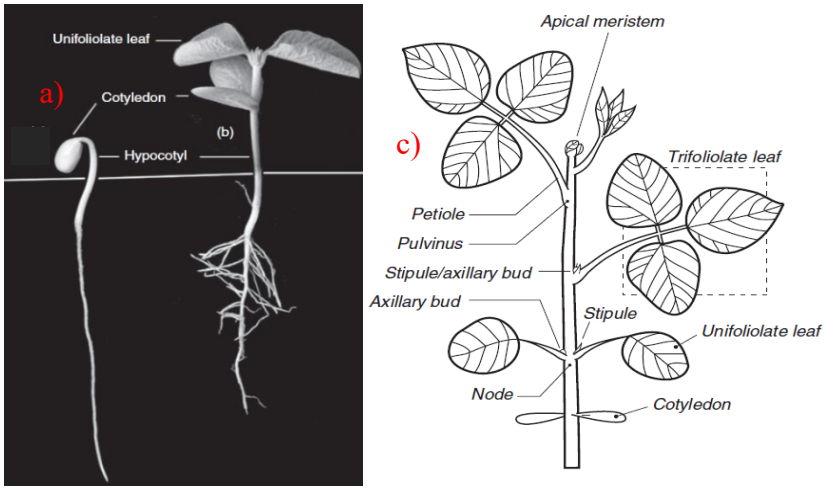


Figure 2: Diagram of soybean through vegetative stages, a) VE, emerging from the soil, b) VC, unifoliate leaves begin to unfold and root system is developed, c) about the V2 stage, showing all the components of the young plant [11]**. Copyright permission has been obtained.**

Following the vegetative stage is the reproductive stage (denoted as R1) and begins once the first flower has opened. R1 results from the transition of the vegetative meristem to the reproductive floral meristem [11], [12]. The reproductive period signals the plant to begin initiation of pods, seeds and maturation stages. The R1 and R2 stages focus on flower development, R3 and R4 on pod development, R5 and R6 on seed development and R7 and R8 on plant maturity [12]. Environmental and genetic factors continue to regulate reproductive development after flowering and are critical for adequate quality and yield. The R8 stage is reached when 95% of pods along the main stem have developed their mature pod colour and the plant begins to enter senescence stage [11].

Ultimately, the flowering stage in the life cycle of soybean indicates that the plant has detected optimal conditions for reproductive success and is thus one of the most important factors in crop production and yield. The process of flowering and maturity is highly complex and is an important component for adapting a crop to new geographic regions that relies heavily on understanding the underlying genetic mechanisms in response to environmental signals.

1.4 Time of flowering and maturity

Photoperiod is among the most important factors affecting soybean time of flowering and maturity, playing a key role in determining its geographic adaptability. Genetic variation within this pathway has allowed for widespread growth and adaptation of soybean worldwide [13]. Crops have certain requirements for timely flowering which, if not met, may delay or prevent flowering from occuring. Plants are classified as long-day (LD), when the day length (photoperiod) exceeds a certain threshold, or short-day (SD),

like soybean, where plants flower when the daylength is under a certain threshold (hours). However, soybean has the genetic diversity that harbors beneficial alleles in a large number of genes (due to polyploidization) allowing for adaptability to longer photoperiods, such as Northern Canada [13].

Many genes involved in these pathways are highly conserved across well-studied model plant species *Arabidopsis thaliana* and rice (*Oryza sativa*). These species can therefore be used as an aid to better understanding the underlying mechanisms in soybean [14]. Additionally, due to the complexity of the soybean genome with many redundant genes, low transformation efficiency, and long transgenic regeneration process, using a model plant to confirm gene function acts as an intermediate tool for faster, more efficient results, and has had success in the past with confirming the involvement of candidate genes at the other maturity loci on time of flowering and maturity [15].

1.4.1 **Molecular control of flowering in *Arabidopsis thaliana* & *Oryza sativa* (rice)**

Great advances have been made in model plants *Arabidopsis* and rice (*Oryza* sativa) to better understand the photoperiod pathways controlling flowering [14]. In *Arabidopsis,* flowering is controlled by a complex regulatory network based on various endogenous and environmental factors. Environmental signals such as photoperiod and temperature play an important role in determining floral transition. However, it is the response of internal pathways (such as the photoperiod pathway) to these environmental cues that holds the keys to geographic adaptation.

The photoperiod flowering pathway detects seasonal changes based on daylength (in the case of *Arabidopsis* – LDs) in combination with internal clocks to help determine

the optimal time to flower, ensuring reproductive success. At the helm of this pathway is *CONSTANS (CO),* a zinc finger transcription factor that integrates internal and external signals to regulate *FLOWERING LOCUS T (FT)*, which encodes florigen, a flowering hormone that moves from leaves through the phloem and into the shoot apical meristem (SAM) to initiate flowering (Figure 3) [16]. In addition to *CO,* several transcriptional repressors also regulate *FT* expression such as; *APATELA2 (AP2)*, *TEMPRANILLO 1 (TEM1),* and *TEM2*, however *CO* is the main transcriptional regulator of *FT*.

The circadian clock regulates the majority of *CO* expression by many circadian clock proteins, such as *KELCH REPEAT, F-BOX1 (FKF1), CIRCADIAN CLOCK ASSOCIATED 1 (CCA1), LATE ELONGATED HYPOCOTYL (LHY), PSEUDO RESPONSE REGULATOR (PRR), GIGANTEA (GI)*, and *CYCLING DOF FACTOR 1(CDF1)* [16], [17], [10]. *CDF1* expression is positively regulated by *CCA1* and *LHY* proteins in the mornings, which are then repressed by *TOC1* and *PRR* family proteins in the afternoon. These clock proteins keep *CO* expression low in the mornings ensuring that flowering does not occur too early in SD conditions. During LD, *CDF1* represses *CO* transcription in the mornings by binding to the promoter of *CO.* In the evenings, when peak day length is detected, *FKF1* is activated by detecting blue light thus forming a complex with *GI*, alleviating the repression of *CO* by *CDF1* via a ubiquitin-dependent degradation [18]. The repression of *CO* in the mornings ensures its expression in the afternoon under LD allowing activation of *FT,* subsequently resulting in initiation of photoperiodic flowering. However, under SD conditions a complex does not form between *FKF1* and *GI* and therefore levels of *FT* expression are too low for flower initiation. Stability of *CO* is important for *FT* floral activation, since peak *CO* accumulation occurs throughout the

night, various mechanisms ensure its activation only in the afternoon under LD. Additionally, *PHYTOCHROME A (PHYA)* and *PHYTOCHROME B (PHYB)* mediate far-red light and red light, respectively, also modulating the stability of *CO* [19]. In summary, *CO* is regulated by circadian clock proteins and light signals which, based on appropriate day length, cue *CO* to activate *FT* expression and thus initiate flowering (Figure 3).

Rice is a SD plant and therefore flowers under different regulatory mechanisms, however the *GI-CO-FT* module in *Arabidopsis* is conserved, with minor differences in the function of *CO* (Figure 3). *OsGIGANTEA (OsGI)*, *Heading Date 1 (Hd1)*, and *Heading date 3a (Hd3a)* are homologs of *Arabidopsis GI*, *CO*, and *FT*, respectively [18]. Rather than degradation of *Hd1* during the night, like *CO* in *Arabidopsis*, its function changes based on daylength. In SD *Hd1* activates the expression of *Hd3a*, however during LD *Hd1* activity flips and represses the expression of *Hd3a* [18]. Additionally, rice contains two important genes for the regulation of flowering among the photoperiodic pathway, *Earlyheading date1 (Ehd1)* and *Grain number, plant height, and heading date 7 (Ghd7)*, which are specific to grasses and function in the activation of *Hd3a* [18].

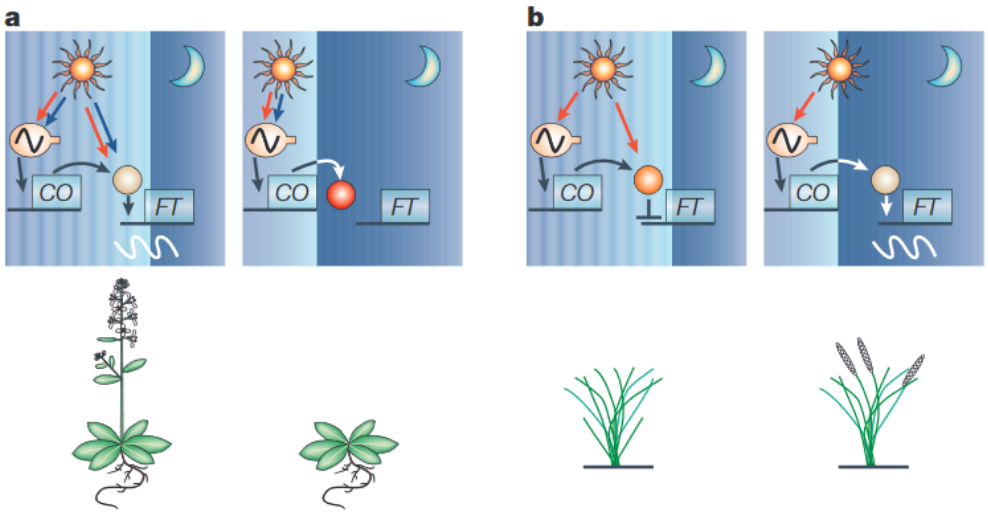


Figure 3: Florigen signaling pathway differences in *Arabidopsis (a)* and rice (b) showing how florigen is activated both in LD and SD in rice, however not in *Arabidopsis*[18]. Copyright permission has been obtained.

A phylogenetic analysis, Figure 4, with photoperiod flowering pathway genes conserved across *Arabidopsis,* rice and soybean was previously conducted in our lab to identify which two species share closer sequence homology. Genes were chosen based on well-known association with the photoperiod flowering pathway that were available among all three species, including the highly conserved *GI-CO-FT* module. Although rice, like soybean is a SD plant, our phylogenetic analysis, Figure 4, has confirmed that commonly known photoperiod flowering genes among the three species are more closely related between *Arabidopsis* and soybean. This may stem from differences in physiology, as soybean and *Arabidopsis* are both dicots whereas rice is a monocot. *Arabidopsis* contains a large number of mutants, a small genome, short life cycle and has previously been used to elucidate the function of many soybean genes [20]–[22], resulting in an excellent model plant for this project [23].

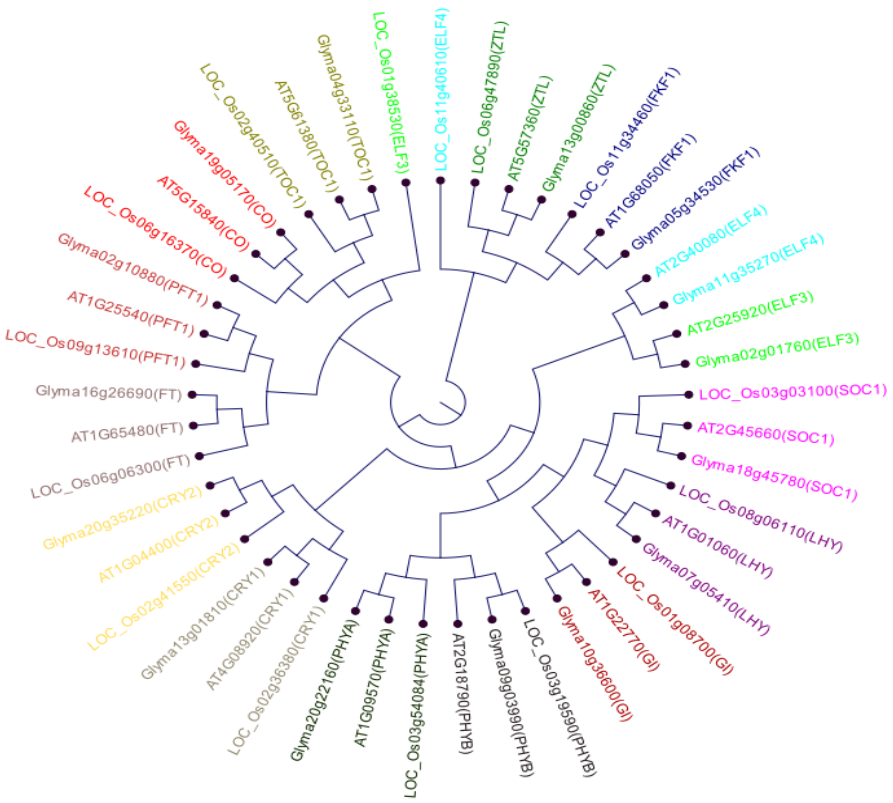


Figure 4: Phylogenetic analysis of photoperiod flowering pathway genes in <u>Arabidopsis,</u> rice and soybean using amino acid sequence. The phylogenetic tree was constructed using MEGA-X software with 1000 bootsraps. The colours above separate the different photoperiod flowering genes.

1.5 Molecular basis of flowering in soybean

With 11 known/confirmed genes involved in this pathway and over 100 QTLs (Quantitative Trait Loci) in some way associated with time of flowering and maturity, the understanding of this pathway is still largely unknown. To date, there are 11 major loci controlling time of flowering and maturity identified in soybean, known as the *E* loci, E1-E11 and J [24]–[31], with the underlying genes of E6, E7, E8 and E11 remaining unknown and E5, once thought to be its own loci, confirmed as an allele of E2 [32], [33]. A reduced sensitivity to LD photoperiods is mainly attributed to dysfunctional alleles at the E1-E4 maturity loci while the recessive alleles at the E6, E9 and J loci result in delayed time of flowering and maturity [34].

The major floral integrator in *Arabidopsis*, *FT*, is conserved throughout soybean, where 11 *FT-like* genes have been identified [*GmFT1a, GmFT1b, GmFT2a (E9), GmFT2b, GmFT2c, GmFT3a, GmFT3b, GmFT4 (E10), GmFT5a, GmFT5b,* and *GmFT6*]. Six of these *FT-like* genes were found to promote flowering in *Arabidopsis* ecotype Columbia (Col-o), *GmFT2a/b, GmFT3a/b,* and *GmFT5a/b* [35]. *GmFT4,* was identified to act as a flowering repressor while *GmFT6* had no effect on flowering. Of these *FT* homologs, *GmFT2a* was discovered as the E9 maturity gene, and *GmFT4* discovered as the E10 maturity gene in soybean. *GmFT2a* and *GmFT5a* have been found to play a major role in photoperiod flowering response, both of which promote flowering when overexpressed in *Arabidopsis* and soybean [36]. However, the role of *GmFT5a* may extend beyond photoperiod response as it was even found to be expressed under LD. The *Arabidopsis* homologs *PHYA* (*E3* & *E4*) and *GI* (*E2*) have also been found to regulate photoperiod

flowering in soybean. The recessive alleles at these three loci have been found to increase the expression of *GmFT2a* and thus promote flowering under LD [36].

Soybean contains 26 *CONSTANS-like (COL)* genes, with *GmCOL1a, GmCOL1b*, *GmCOL2a* and *GmCOL2b* showing closest sequence similarity to *Arabidopsis CO.* However, no evidence has suggested any of the *COL* genes play a role in flowering time regulation in soybean. Overall, the function of the *GI-CO-FT* mode is still largely unknown in soybean, indicating that there are still many undiscovered components in *FT* activation [36].

1.5.1 **E1 (*Glyma.06G207800*)**

The E1 locus was discovered by R. L. Bernard in 1971 and was found to have a large effect on time of flowering and maturity [37]. This loci has been mapped to the central region of chromosome 6 and corresponds to *Glyma.06G207800* [14]. E1 is identified as a legume specific transcription factor that functions as a flowering repressor and encodes a putative nuclear localization signal (NLS) along with a domain distantly related to the B3 domain [14]. Allelic variation at the E1 locus includes the E1 allele, which is fully functional and delays flowering while the recessive alleles promote flowering; e1-as is partially functional (amino acid substitution affecting the NLS), e1-fs (frameshift mutation), e1-nl (null allele), and e1-re (retrotransposon insertion) are non-functional alleles [38]. No homologs for E1 exist in *Arabidopsis* or rice, however homologs in other legumes are present with different roles, common bean (*Phaseolus vulgaris*) PvE1L inhibits flowering while, in *Medicago truncatula, MtE1L* does not have an effect on flowering [35]. The E1 locus and its homologs, *E1La* and *E1Lb* repress flowering in LD

conditions by downregulating *GmFT2a* and *GmFT5a* expression, which are two orthologues of *Arabidopsis FT,* and thus inhibit flowering, whereas under SD conditions the function of E1 is suppressed [35].

1.5.2 **E2 (*Glyma.10G221500*)**

The E2 maturity locus was discovered by R. I. Buzzell in 1971[39], and molecularly characterized by Watanabe et al., 2011[40], as an orthologue of the *GI* gene in *Arabidopsis*, *GmGIa (Glyma.10G221500*) located on chromosome 10. Two other *GI* homologs exist in soybean however the effects are unknown [35]. Three known alleles exist for E2, the functional dominant allele *E2,* recessive *e2* allele that encodes a truncated protein, and recessive e2-ns that results in a premature codon [41]. The recessive e2 alleles promotes flowering by upregulating *GmFT2a* expression under LD conditions while the dominant E2 allele suppresses flowering [35]. The expression of E2 is down regulated by E1 and up regulated by E3 and E4, although further studies are required to understand if *GI* regulates flowering in soybean the same as *Arabidopsis*.

1.5.3 **E3 (*Glyma.19G224200*)**

The E3 locus was discovered by R. I. Buzzell 1971 [39], and molecularly characterized by Watanabe et al., 2009 [28] as a phytochrome A (*PHYA*) gene, *GmPHYA3* (*Glyma.19G224200),* located on chromosome 19. The 3 dominant E3 alleles are fully functional and repress flowering, *E3*, with both E3-Mi and E3-Ha having large ~2.5Kb insertions. The recessive alleles, e3- ns (nonsense mutation), e3-fs (frame shift), e3-Mo (allele from cultivar 'Moshidou Gong 503' containing SNP in exon), and e3-tr (13.3 kb

deletion) are non-functional and promote flowering [41]. E3 controls flowering in LD by responding to high red/far-red (R:FR) light ratios, regulating the expression of *GmFT2a* and *GmFT5a* by controlling the amplitude of expression of the *E1* and *E1L* genes [35].

1.5.4 **E4 (*Glyma.20G090000*)**

The E4 locus, discovered by Buzzell and Voldeng 1980 [42], is similar in function to E3 in that it is also a homolog of *Arabidopsis PHYA*, *GmPHYA2* (*Glyma.20G090000).* E4 has been molecularly characterized by Watanabe et al., 2011[40] found to reside on chromosome 20. Allelic variation at the E4 loci contains one functional allele, *E4* that represses flowering, and 5 non-functional alleles, The e4-SORE (*Ty1/copia-like retrotransposon*) has a large ~6kb insertion, e4-oto, e4-tsu, e4-kam, and e4-kes contain single base deletions that promote flowering [27], [38]. In contrast to E3, E4 controls flowering in LD conditions by responding to low R:FR ratios and similarly, as E3, regulates expression of *GmFT2a* and *GmFT5a* [35].

1.5.5 **E6 (*unidentified*) & J (*Glyma.04G050200*)**

E6 and J are associated with the long juvenile (LJ) trait, which extends the time of flowering and maturity of soybean grown in southern latitudes so there is an increase in yield [43]. Bonato and Vello, 1999 [25], discovered the E6 loci that delayed time of flowering and maturity. Although the underlying gene for E6 is still unknown, QTL mapping has revealed its location in chromosome 4, in close proximity to the J gene [44].

The J gene was discovered by Ray et al., 1995 [45], and has recently been molecularly characterized as an orthologue of *Arabidopsis EARLY FLOWERING 3*

(ELF3), Glyma.04G050200 , located on chromosome 4 [43]. The J gene contains 8 recessive alleles, *j-1to j-8*, which result in enhanced grain yield with delayed flowering. However, this discovery is mainly relevant to soybean cultivation in tropical regions although can also play important roles in the understanding of flowering pathways within soybeans potentially improving cultivation in northern regions [43].

1.5.6 **E7 (*unidentified*)**

Cober and Voldeng, 2001 [26], discovered the E7 loci and found that it affects photoperiod responsiveness with the recessive genotype resulting in early flowering and maturity. E7 was found to be tightly linked to maturity loci E1 and T (Tawny pubescence), all which reside on chromosome 6. SSR marker analysis performed by Molnar et al., 2003, [46], identified that the location of E7, linked to Satt319, came after T and E1, and that this region is homologous to a region on chromosome 19, linked to a pod maturity QTL. Additionally, Kong et al., 2018 [47], mapped a major flowering and maturity QTL to this same region om chromosome 6, however the search still continues in identifying the underlying gene for E7 .

1.5.7 **E8 (*unidentified*)**

Cober et al., 2010 [48] discovered E8 using backcross-derived near-isogenic lines, identifying a new recessive gene that flowers ~5-8 days earlier in LD conditions, extending a cultivar to a new MG, MG 000. Genome wide SSR (Simple Sequence Repeat) markers identified three candidate regions, one on chromosome 4 between SSR markers Sat_404 and Satt136, and the other two on chromosome 19, one near Satt313 and the other near

Satt166 [48]. There has been no other maturity gene identified on chromosome 4 whereas the other two candidate regions on chromosome 19 may have been inadvertently selected due to the presence of E3. Additionally, two recent papers published by Wang et al., 2017 [49], and Kong et al., 2018 [47], identified major QTLs on chromosome 4 associated with a flowering time trait that could be controlled by E8. Other than *E1La* and *E1Lb* as potential candidates for E8 due to their location on the same QTL, no other reports have suggested a likely candidate [35].

1.5.8 **E9 (*Glyma.16G150700*)**

E9 was discovered by Kong et al., 2014 [50], and mapped to a 245 kb region on chromosome 16, where two potential flowering genes (*GmFT2a* and GmFT5a) were located. Zhao et al., 2016 [31], later confirmed the underlying gene to be *GmFT2a, Glyma.16G150700*, an ortholog of *Arabidopsis FT* gene required for flower initiation. The recessive allele for *GmFT2a* contains a *Ty1/copia-like* retrotransposon that delays flowering due to a reduction in transcript abundance [35].

1.5.9 **E10 (*Glyma.08G363100*)**

E10 was discovered and molecularly characterized by Samanfar et al., 2017 [30], using a combined approach of classical breeding, bioinformatics and molecular biology related practices. E10 was found to correspond to *FT4*, *Glyma.08G363100,* another *FT* orthologue of *Arabidopsis* that acts downstream of E1 to repress flowering under LD conditions. The recessive allele at this loci was found to contain an amino acid substitution that resulted 5-10 days earlier maturity under SD conditions [30].

1.5.10 **E11 (*unidentified*)**

The E11 locus was discovered by Wang et al., 2019, [32] using next generation sequencing (NGS) approaches. A new QTL was identified on chromosome 7 that controls time of flowering and maturity by a single dominant gene, conferring early flowering. Based on amino acid sequence analysis of the 11 genes within this QTL, three candidates were found with sequence variation among contrasting genotypes, *Glyma.07G049000, Glyma.07G048500,* and *Glyma.07G049200*, annotated as *photosystem 1P subunit, Homeodomain-like superfamily protein,* and *metal tolerance protein A2*, respectively [32].

1.6 Molecular markers

Over the last few decades there has been a rapid increase in the understanding of plant genomes, mainly by better understanding the effects of genetic variants throughout plant populations. This is largely due to the advancements in marker technology at the DNA level, (DNA-based markers). Molecular markers were the first steppingstone in understanding the structure of genomes and have rapidly progressed in identifying the location and function of genes affecting economically important traits.

Generally speaking, there are three types of markers; morphological (observable traits), biochemical (isozymes, allelic variants of enzymes detected by electrophoresis or staining), and genomics-based markers (i.e. sites of variation in DNA); DNA-based markers are the most abundant and are not influenced by their environment or growth stage and therefore the most widely used today [51]. The use of DNA-based markers relies on variation in DNA sequences, these mostly include single nucleotide polymorphisms

(SNPs), insertions or deletions (InDels), variable number tandem repeats (mini- and microsatellites, or SSR), transposable elements, and so on [52]. DNA-based markers are generally short DNA sequences that correspond to a certain location on a chromosome and can identify differences (polymorphisms) between individual organisms and species [53]. They play vital roles such as species and allele identification, marker-assisted selection (MAS), molecular breeding, genetic mapping, quantitative trail locus (QTL) analysis, and genome-wide association studies (GWAS) [54].

Over the years there has been a wide variety of marker options developed for plant geneticists and breeders, of those, microsatellite (simple sequence repeats or SSRs), and single nucleotide polymorphisms (SNPs) markers have been extensively used [55]. SSRs are short tandem repeats of di- or tri-nucleotides (nucleotide repeats up to 8 exist but get rarer with larger size). SSR markers are highly reproducible, require small amounts of DNA, are highly polymorphic, codominant, abundant and frequently distributed in the genome [56]. SNP markers are single base-pair changes in DNA sequence, they also are highly reproducible, require small amounts of DNA, and are not as polymorphic as SSRs, however their abundance and frequency throughout the genome compensates [55]. Overall, the ideal molecular marker should be co-dominant, be highly reproducible, able to detect higher levels of polymorphism, and have an accuracy very close to 100% [56]. Molecular markers form the basis of many techniques used in this project.

1.7 Computational and systems biology

With the continual increase of available biological data, computational analysis becomes increasingly more important. Biology and computers are merging closer together

as the flood of data, mainly in the form of DNA, RNA and protein sequences grows, putting a larger demand on computers and computational scientists. Before computers became essential to biology, understanding gene function only involved performing extensive experiments mostly through low-throughout approaches. With the availability of continuously increasing computational power, computers can be used to predict data-based hypothesis and leads by filtering and sorting loads of important information. Of course, this is not without the large amount of experimental data that has been collected to date, since all computational predictions are made based on this information. This also leads into why computational methods and tools have so many limitations. Computer programs only analyze and predict based on how a cell or an organism is believed to function; they will never be able to replace living cells due to the complexity of living systems and the lack of their functional understanding. With that being said, the rapid advancements and cost reductions in whole genome sequencing, RNA sequencing, and overall molecular tools and methods, brings us one step closer to fully understanding how the individual components work both independently and together with other systems. This is a field called systems biology, where the goal is to understand cellular function in its entirety.

Systems biology aims to help understand biological systems by using computational methods and bioinformatic tools in combination with wet-lab, experimentally obtained data to help analyze, interpret, and even predict the function(s) occurring throughout an entire system. Much like physics and chemistry, biology is headed into a more precise science. There are currently three approaches that exist for generating an understanding of biological systems. The ‘top down’ approach utilizes functional genomics information together to build an understanding of key players involved in a system. The ‘bottom up’

approach first looks at the underlying properties, starts small and goes bigger. Finally, the 'middle out' approach starts between the top and bottom, working out towards either extremes. In this case for this project a 'top down' approach has been selected to help identify genes involved in time of flowering and maturity, and thus revealing the possible underlying gene for E8.

1.7.1 **Functional genomics**

Computational biology has flooded the field of functional genomics, which contributes to the overall understanding of biological systems by attempting to understand the function and interaction of genes and proteins. For this, functional genomics leverages experimental evidence of the -omics; genomics, transcriptomics, proteomics, and metabolomics. Genomics is the study of an organism's entire genome at the DNA level and holds the basis of understanding the role of genes in the development of an organism, its adaptation to different geographic regions, its response to environmental stimuli and so on. Transcriptomics analyzes the entire transcriptome (RNA) providing insight on the regulation of gene expression including coding (mRNA) and non-coding (rRNA, miRNA etc.). With the reduced cost of high-throughput RNA sequencing technologies, the transcriptome has provided a powerful tool to study the function and expression of genes. Proteomics analyzes the expression profile of proteins and their interactions with one another throughout the entirety of the organism (the proteome), by analyzing structure, function, post-translational modification and protein-protein interactions (PPI's). Metabolomics is the study of molecule metabolites produced by an organism under a certain set of conditions. Metabolites contain more diversity within an organism compared to genes and proteins and can therefore be used as fingerprints to identify certain cellular

processes within an organism [57]. Together, these fields can be used to overcome the major challenge in understanding the function and regulation of genes [58].

1.7.2 **Protein-protein interaction**

Proteins carry out the majority of biological function within an organism, mostly through their physical interaction with one another [59]. Therefore, it can be stated that protein-protein interactions (PPI) define an organism's functionality, development and responses to various stimuli [59].

Although there are many traditional experimental methods to test for PPIs, such as yeast two-hybrid (Y2H), and tandem affinity purification (TAP), they are labour and time intensive, costly, and generate poor accuracy with a high rate of false-positives and false-negatives [60]. Recently, computational methods have been used to construct PPI databases based on known experimental evidence, and even predict novel PPIs based on sequence structure and known interactions [59], [61], [62]. PPIs can and have been used [30], to identify or short-list potential candidate genes involved in complex pathways. This has been done by a concept called "guilt by association" where if a protein of interest (POI) is interacting with a group of proteins involved in a particular pathway (for example time of flowering and maturity) then that POI may also be involved in time of flowering and maturity [58]. PIPE (Protein-protein Interaction Prediction Engine) is an example of a bioinformatics tool that predicts proteome-wide PPI and will be discussed in more detail in the methods section [30], [60].

1.8 Genetic mapping

Traditionally maps are used to navigate to areas of geographic interest with symbols and elements that define characteristics within a region. Similarly genetic maps do the same by indicating the position of markers, genes and QTLs along a chromosome [53]. Following the laws of segregation, markers found in close vicinity to genes stay together through generations of offspring. The law of independent assortment states that chromosomes segregate independently of one another during meiosis, in other words the allele of a loci separates independently of the allele at another loci. This assortment occurring at random is considered to be in linkage equilibrium. However, when the assortment of alleles deviates significantly from their random behaviour, they are said to be in linkage disequilibrium [56]. Statistically analyzing the degree of linkage disequilibrium can be used to generate genetic maps and is a key component in association mapping.

Linkage maps are constructed by analyzing marker segregation. Considering genetic linkage occurs when alleles for a gene are inherited together, analyzing this pattern of segregation, or the frequency at which two genes become separated during meiosis, can reveal the genetic distance between two genes/loci and thus generate a linkage map [63]. Traits that are more complex and conditioned by multiple genes or loci are considered quantitative or polygenic traits. Mapping these traits involves QTL analysis, a statistical method that links a measurable phenotype with genotypic data and allows the mapping of complex traits to certain regions on a chromosome [64]. Arguably the two most important components to generating a successful QTL map are the mapping population and marker choice. The parents of a mapping population should differ (preferably have the trait of

interest at the highest or lowest extreme) for the traits of interest and the resulting population should segregate for these traits [65]. Typically the number of individuals for a mapping population ranges anywhere from 50-250, and even more (<500) for generating high resolution maps or for identifying QTLs with a small phenotypic effect (minor QTLs) [64]. There are several marker types that are commonly used for QTL mapping in plants, each with their own advantages and disadvantages, such as; restriction fragment length polymorphism (RFLP), amplified fragment length polymorphisms (AFLP), inter simple sequence repeat (ISSR), simple sequence repeat (SSR), expressed sequence tags (ESTs), diversity arrays technology (DArT), and single nucleotide polymorphism (SNP) [56]. The number of markers used in mapping studies depends on the marker choice, their availability and species under study. QTL mapping is very useful for identifying QTLs across the genome responsible for economically important traits. However there are disadvantages such as low allelic diversity, lower number of recombination events, time and labour intensive, lower specificity, and so on [56].

In contrast to QTL mapping, association mapping (AM) utilizes a diverse germplasm collection with contrasting geographic origins and therefore provides a significant association of molecular markers with a phenotypic trait. [66]. AM can use historical recombination events between markers and QTLs therefore providing higher resolution maps, a greater number of alleles, and saves more time in comparison to linkage mapping [65]. Recent advancements in computational power and high-throughput genotyping (next generation sequencing approaches such as GBS) technologies has also allowed for genome-wide association study (GWAS). GWAS is used to study genetic

variations in hundreds to thousands of varieties and associate them with complex traits by investigating hundreds of thousands of SNPs across the genome [67].

The abundance of markers along with the generation of high-resolution linkage maps has made marker-assisted selection (MAS) more feasible. MAS utilizes molecular tools and techniques to identify allelic variation and select for a very particular trait(s) of interest [51]. On the other hand, there is also conventional breeding, which is mostly based on phenotypic selection rather than relying on markers to identify the traits of interest. MAS can be advantageous in many ways such as when selecting for traits governed by minor QTLs. However, MAS is not meant to replace breeding practices. Regardless of technique and plant species, the main goal in agricultural science is increase yield and quality of crops, where both MAS and conventional breeding can be used hand in hand [51].

1.9 Purpose and objective

The objective of this research is to identify the underlying gene for the E8 maturity locus in soybean. Due to its high value to the Canadian economy there has been a substantial expansion of soybean throughout the western provinces of Canada. This is largely due to the efforts of soybean breeding programs in developing short season early maturing cultivars. Efforts to continue expanding into the northern areas of these regions are limited by the photoperiod capabilities of soybean, where flowering becomes too far delayed by the long days so much that it can no longer mature in time before frost hits. SoyCanada (https://soycanada.ca/) expects over 35 million acres of new land become available for soybean production in Western Canada by 2027 [7]. In order to meet these

expectations emphasis must be placed on identifying more players involved in time of flowering and maturity. Although the time of flowering and maturity pathway has been extensively investigated over the past few decades, there is a significant gap of knowledge in understanding the comprehensive mechanisms that govern the control of this important and key pathway as well as the cross-communication between other cellular processes. Continued efforts in the identification and characterization of genes involved in this pathway will lead to a better understanding of how the individual components work to regulate photoperiod flowering and timely maturity. Although the E8 locus is just one component within a highly complex process, identifying and elucidating the function of its underlying gene will further aid in its understanding while simultaneously assisting breeders in developing ultra-early maturing cultivars attuned for more northern regions.

The E8 locus has been previously characterized to a large region on chromosome 4 by three independent studies [47]–[49]. Considering the magnitude of genes present within this region, ~1000 genes, discovering the underlying candidate can be similar to finding a needle in a haystack. Beginning with the broad hypothesis that a lack of knowledge exists in understanding time of flowering and maturity in soybean, and that the underlying gene is present among the E8 region on chromosome 4, a bioinformatics approach was first employed to predict the genes (proteins) that may have an association with genes known to be involved in time of flowering and maturity. These predicted proteins (genes) were then further funneled down by various computational tools and resources such as observing LOF mutants, presence of SNPs, and RNA sequence data, providing a manageable list of candidates. Finally, applying traditional molecular biology related practices such as

sequencing and expression analysis via qPCR and ddPCR, would conclude the remaining candidates on their potential association with the E8 maturity locus.

Chapter 2: Materials and methods

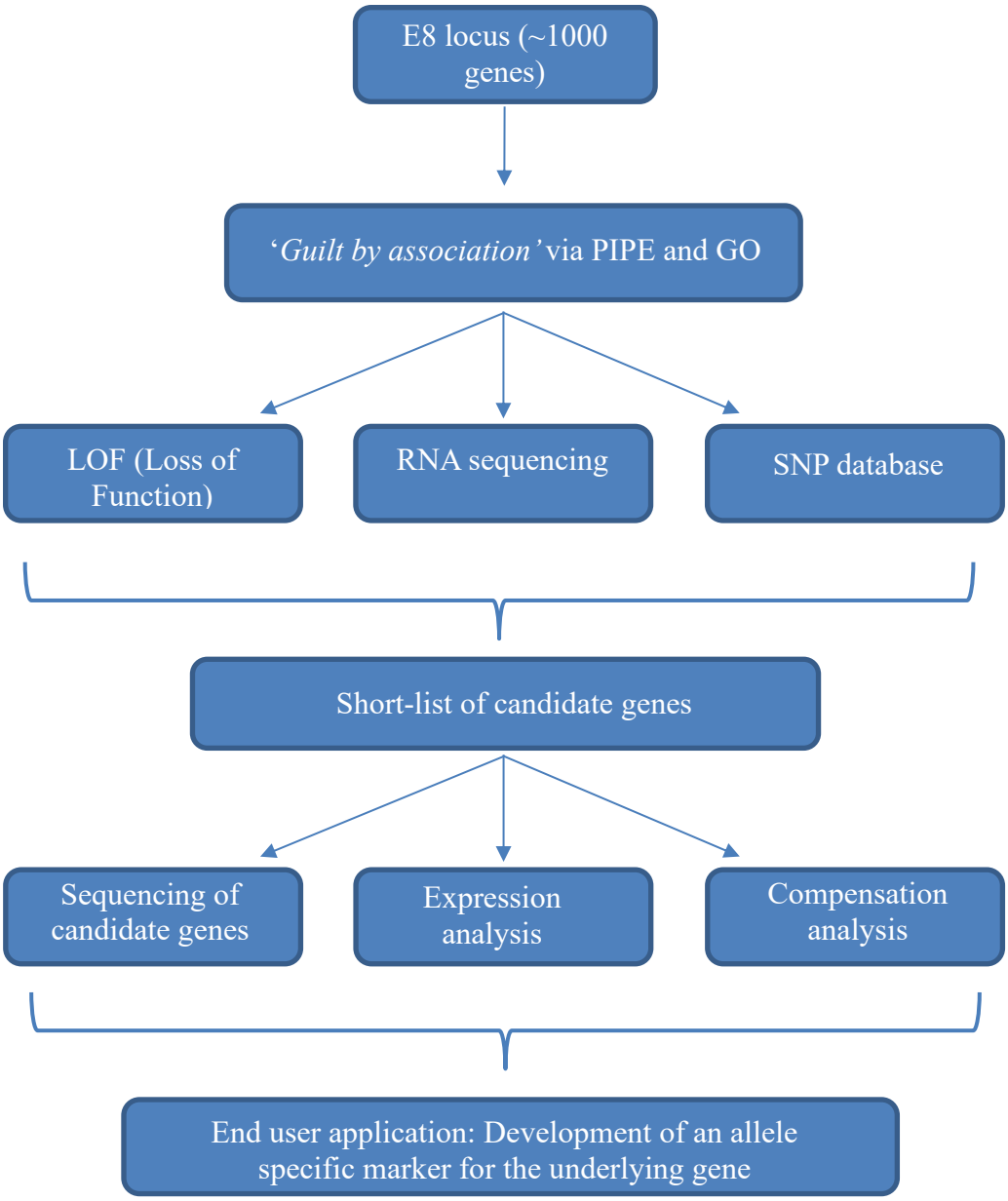


Figure 12: Experimental workflow for the identification of the candidate gene for the E8 maturity locus in soybean.

2.1 Computational analysis

With almost 1000 genes physically located within the E8 region traditional approaches would be far too time consuming and costly. For this reason, a computational approach was first employed to significantly short-list the potential candidate genes into a much more manageable list. This approach consisted of analyzing available genomic, transcriptomic and proteomic data such as PIPE, gene ontology (GO), identifying LOF mutants, the presence of any SNP's, and analyzing available RNA-sequence data for changes in expression between growth stages.

2.1.1 PIPE & Gene Ontology (GO)

PIPE (Protein-protein Interaction Prediction Engine) is a computational tool that predicts protein-protein interactions based on short reoccurring amino acid sequences and a known interaction database [16]. Briefly as presented in Figure 6, within the known interaction database the PPIs among V, W, X, Y and Z are known, and we would like to investigate whether or not A and B are interacting. PIPE investigates the similarity, similarity matrix, of a window of twenty amino acids from protein A across all known databases, shifting over one amino acid after each completion until the end of the sequence is reached. Say at a given position in protein A, there are similarities in V and W (from the known database), PIPE then selects all the known interacting partners with V and W, which in this case are X, Y and Z [these are selected and categorized as a separate list (data set) called "R"]. PIPE now investigates the similarity, similarity matrix, of a twenty amino acids long window from protein B, across all interacting partners located on list R. Then, based on a number of computational calculations including similarity matrix indexes, PAM

scores and so on, PIPE will confirm whether or not A and B are interacting or not [at specific sensitivity (23%) and specificity (99.9%)]. The most updated and advanced version of PIPE, version 4, has been used [68].

Together with PIPE, GO (Gene Ontology) can be a very powerful resource in identifying/annotating genes with unknown function. Thus, attempting to identify interacting partners using PIPE and annotated these partners using GO, can help identify genes that are involved in time of flowering and maturity.

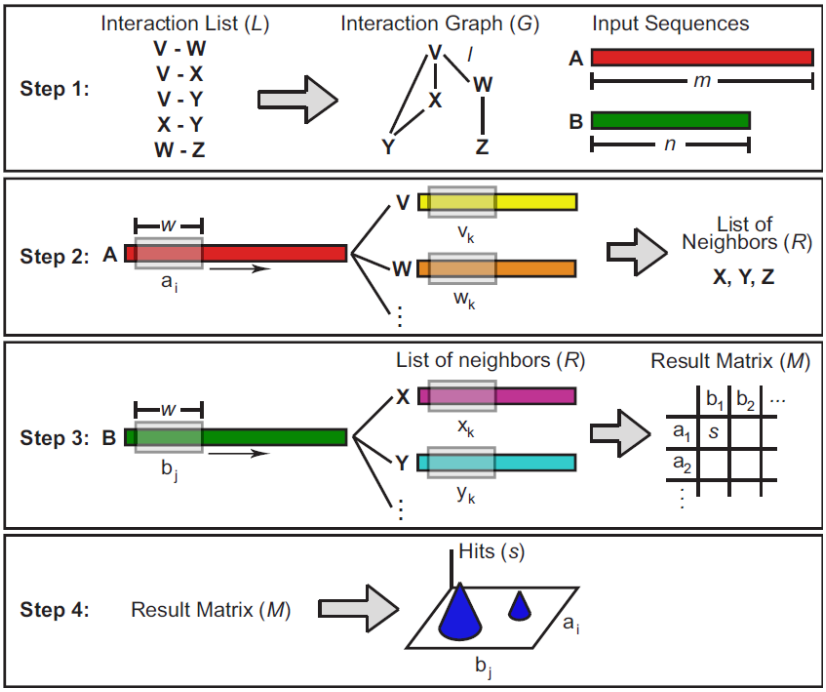


Figure 7: A simplified algorithm for PIPE workflow, adapted from Pitre et al., 2006 [16]. Copyright permission has been obtained.

2.1.2 Loss of function (LOF) analysis

LOF mutations have a large functional impact on gene expression by introducing a premature stop codon or shifting the reading frame to alter the function of a given gene. A database of 1,007 *Glycine max* accessions, including Canadian and international soybean

cultivars, containing LOF mutations was provided by Torkamaneh et al., 2019 [69]. In this study 18,031 putative LOF mutations in 10,662 genes were predicted, with the assumption that some may be relevant to candidate genes among the E8 region.

2.1.3 **SNP database investigation**

Genome wide nucleotide and structural variations were identified in 530 Canadian and international soybean accessions using a combined GBS (genotype-by-sequencing) and WGS (whole genome sequencing) approach by Torkamaneh et al., 2018 [70]. Accessions included in this study contained three E8 lines (OT94-47, Harosoy, and Maple Presto), and two e8 lines (840-7-3, and OT02-18). Presuming a rare e8 allele, SNPs were only recorded if they occurred below 15% of the population in e8 lines. The occurrence of these SNPs was then analyzed for their location within a gene, if they were found within the intron, exon and whether they were upstream or downstream of the gene.

2.1.4 **RNA-sequencing (expression) database analysis**

RNA sequence data adopted from Severin et al., 2010 [71] was also utilized to identify expression differences in tissues, specifically leaf, first flower, etc., and is presented in Appendix 2.

2.2 Plant material

Soybean lines provided by AAFC-Ottawa RDC soybean breeder Dr. Elroy Cober, Table 1, with contrasting E8/e8 genotypes were grown in the greenhouse at AAFC. Seeds were first germinated in vermiculite for one week before being transferred to commercial

peat based potting soil. The week old plants were then grown in a 13 hour photoperiod with day temperatures of 25ºC - 29 ºC, and night temperatures of 20 ºC - 25 ºC. The early maturing e8 allele (5-8 days earlier) was presumed to come from a variety of crosses using isolines derived from Maple Presto and Harosoy with 840-7-3 (source of early maturity and high protein).

Table 1: List of soybean lines used in this study with their presumed E8 locus status

Line	***Genotype***
Harosoy	E8
OT94-47	
Maple Presto	
X4627-1-16-1	
840-7-3	e8
OT02-18	
OT98-17	
OT04-12	

2.3 DNA extraction

Young trifoliate leaves for each soybean line were collected around the V4-V5 stage and immediately placed in liquid nitrogen. Genomic DNA was extracted from frozen leaves using a modified urea extraction buffer method which included 100 mmol Tris-HCl/L (pH 8), 20 mmol EDTA/L (pH 8), 700 mmol NaCl/L, 8 mol urea/L, 3% w/v sarcosyl, and 1% w/v SDS. Each sample was crushed into a fine powder using a pestle and liquid nitrogen, then incubated at room temperature in 500 µL of the urea buffer. Following this, 500 µL of phenol-chloroform-isoamyl alcohol (ratio of 25:24:1), was added, and mixed for an additional 5 min before centrifugation at 13,000g for 12 minutes. The upper supernatant was extracted using wide bore 200 µL pipette tips and added to a new 1.7 mL micro centrifuge tube containing 500 µL chloroform-isoamyl alcohol (ratio of 24:1). New

tubes were then placed in the centrifuge at 13,000g for 12 minutes, after which the upper aqueous phase transferred to new tubes, using 200 µL wide bore tips, containing 250 µL of 5 mol NaCl/L. Precipitation of DNA was then done using one volume of isopropanol incubated at -20ºC for 10 minutes. Following this the samples were centrifuged at 10,000g for 2 minutes, and washed twice using 750 µL of 70% ethanol. Between washes the solution was sucked out using a vacuum without disturbing the pellet. After the final wash DNA pellets were placed in a Speedvac (Eppendorf Vacufuge) with heat to remove the leftover ethanol. Pellets were then dissolved in 100 µL TE buffer (10 mM Tris, 1 mM EDTA, (pH 8.0)) for 30 minutes in a 37ºC hot bath. The DNA concentration was then measured by optical density using a NanoDrop 2000 spectrophotometer by ThermoFisher. DNA concentration was constantly monitored to ensure above 0.5 µg/µL and an A260/280 ratio above 1.8. DNA was then tested on 1% w/v agarose gel to confirm quality.

2.4 PCR and sequencing

Genomic sequences for each candidate gene were extracted from soybase.org and primers were designed using PRIMER3 web software, listed in Appendix 4 [72]. A PCR gradient for each primer pair was performed to first identify the optimal annealing temperatures using TaKaRa Ex Taq™ with the TaKaRa recommended reagents and cycling conditions. The primer pairs were then testing against lines presented in Table 1 to ensure proper amplification of a single product. Samples were then purified using ExoSAP-IT™ before performing the sequencing reaction using BigDye™ Terminator v3.1 with ThermoFishers recommended protocol. These samples were then sent to an in-house sequencing facility at AAFC-Ottawa RDC. Chromatograms were analyzed using

Lasergene suites SeqMan Pro and Megaalign programs for identification of contrasting sequence variation in contrasting lines for E8.

2.5 Identifying conserved domains

Proteins of unknown function can be annotated based on similarities in their sequence to that of proteins with known functions. Similarities between key sequences throughout many proteins with the same or similar function can be deemed a conserved domain (a functional or structural unit of a protein). Additionally, analyzing protein sequences for where this conserved domain lies can aid in predicting the potential functional effect of amino acid variation [58].

For this reason the conserved domain search from NCBI (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) was used to identify where the conserved domains lied within our candidate proteins which would aid in confirming a functional domain, and aiding in predicting the functional impact of an amino acid substitution.

2.6 2D RNA structure analysis

2D-RNA structures were predicted using the RNAfold WebFold server from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi*)* [73], to determine if there were any significant structural changes among genotypes [30].

2.7 Predicting functional effects of amino acid substitutions

PROVEAN, (Protein Variation Effect Analyzer) is an algorithm which predicts the functional effect of protein sequence variations with an accuracy of 77.9% [74]. PROVEAN works by analyzing the input sequence similarity among functional protein homologs and then measures the differences in protein sequence similarity after the amino acid variation. This provides a difference in alignment score, called the delta alignment score, and is used to measure the effect of a variation [74].

2.8 Expression analysis

Soybean lines contrasting for E8 genotypes, Table 1, were grown in the greenhouse in both 20 hour and 12-hour photoperiods to identify any expression differences between the wild type and mutant genotypes. Samples were collected at 10 AM every morning. Young trifoliate leaves were collected before flowering (at the V4-V5 stage) during flowering, and 1 week after flowering, then immediately placed in liquid nitrogen and then into -80 ºC freezer. RNA was extracted with TRIzol™ reagent from Thermofisher with 100mg of tissue following their recommended protocol. Pellets were resuspended in 20 µL of RNase-free water and concentrations were measured using the NanoDrop 2000 spectrophotometer by ThermoFisher, ensuring an A260/A280 ratio of ~2. Following RNA extraction, the High-Capacity cDNA reverse transcription kit with RNase inhibitor from ThermoFisher was used for cDNA synthesis, using 1 µg of RNA. To ensure proper cDNA synthesis samples were amplified using traditional PCR and ran on gel electrophoresis to ensure a single specific product was formed.

Commonly used housekeeping genes (*Tubulin, Actin,* etc.) from previous studies (included in Appendix 3) [75], [76] were tested based on their stability throughout developmental stages in soybean. Primers were designed in the same way as for sequence; however, template size was much smaller at 80-150 bp (Appendix 4). SYBR Green™ PCR master mix from ThermoFisher was used to quantify gene expression using the recommended protocol. The standard curve method using four dilution points (1, 0.5, 0.25, and 0.125) was employed for each candidate and housekeeping gene to ensure amplification efficiency between 90% and 110%. Efficiency outside of this range was optimized by changing solution concentrations and dilutions. Mutant (e8) samples were tested against a reference (E8, wildtype) using three biological and technical replicates along with *TUB4* and *ACT11* housekeeping genes. Controls included; a no template control (NTC), which contains zero DNA or RNA to identify if there is any nucleic acid contamination, a no reverse transcriptase control (NRT), which contained RNA instead of cDNA to identify contamination of DNA, and a no amplification control (NAC), which contained no fluorescent dye (SYBR Green) to measure background fluorescence (amplification). This was completed across 8 conditions: 2 genotypes, wild type (E8) and mutant (e8), both grown in long and short-day conditions, each sampled before and after flowering. The Michael Pfaffl method ($2^{-\Delta\Delta C}$)[77], was used to normalize the difference in expression levels among samples.

2.8.1 **Digital Droplet PCR (ddPCR)**

Another approach used for quantifying gene expression was ddPCR, which similar to qPCR, can quantify copies of target DNA [78]. There are two major differences between

ddPCR and qPCR. First that ddPCR partitions the sample into thousands of individual droplets, each droplet with a single target molecule, and can thus be used to produce reliable results even with low low-copy number genes [79]. Second, ddPCR eliminates the need for calibration curve or Ct values, rather than taking a reading after each cycle like qPCR, it simply records fluorescence at the end point and therefore eliminates or reduces the impact of any inhibitors in the PCR reaction [79]. Therefore, ddPCR was chosen for this project to potentially identify larger differences in gene expression among contrasting genotypes (E8/e8), considering the greater detection capabilities of ddPCR along with a significantly lower dependence on Taq polymerase inhibitors.

The same primers and cDNA used for qPCR, was used for the ddPCR reaction, however lower cDNA concentrations were used. EvaGreen Supermix from BIORAD was used to run PCR reactions for digital PCR analysis, according to the recommended set up by BIORAD. Template cDNA was run across a temperature gradient to find the best annealing temperature. Following PCR, samples (in triplicates were formed into droplets using the QX200 droplet generator from BIORAD and then simply ran on digital PCR, with a blank. Only one gene was analyzed using this method as a preliminary experiment, however more results are to come.

2.9 Transformation of candidates into *Arabidopsis*

Candidate genes are currently in the process of functional characterization using *Arabidopsis*. To date, transcript sequences for two candidate genes, *Glyma.04G101500* and *Glyma.04G124600* have been amplified, using Platinum® *Pfx* taq from ThermoFisher, and are ready to be cloned into the Zero Blunt™ TOPO™ cloning vector from ThermoFisher

(Figure 7**)**. This vector contains restriction enzymes that cut blunt ends, both strands at the same spot, directly inserting the blunt-end PCR product into the plasmid vector. The cloning reaction will then be transformed into chemically competent cells, and successful transformants selected using kanamycin.

However, these vectors will first be sequenced to confirm the presence of the premature stop codons. Once confirmed the selected candidates were to be transformed into *Arabidopsis* to analyze if the transformed mutant genotypes affect time of flowering and maturity; more outlined in future direction section.

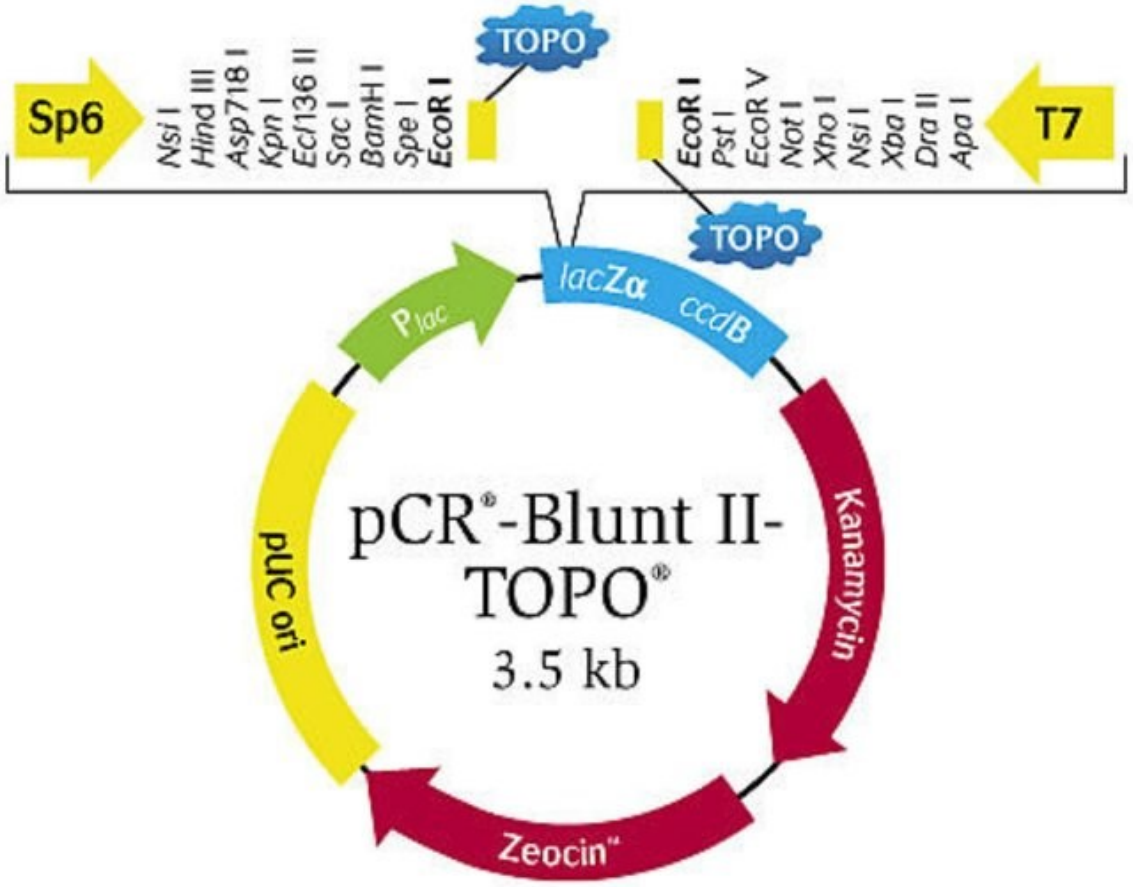


Figure 8: Image of the Zero Blunt™ TOPO™ cloning vector from ThermoFisher. Included are the potential restriction enzyme sites, SP6 promoter/primer for RNA transcription and sequencing, and antibiotics (Kanamycin or Zeocin) for selection in *Escherichia coli* (https://www.thermofisher.com/order/catalog/product/450245#/450245).

Chapter 3: Results

3.1 Identification of candidate genes involved in time of flowering and maturity

Previously in our lab the E8 locus was identified using classical breeding practices and genome wide SSR marker analysis [28]. Recent studies have also been published on the location of the E8 locus, via genome wide resequencing using INDEL (insertion/deletion) markers and QTL mapping using NGS [28], [31], [32]. Collectively, the candidate region among the three studies was used for computational analysis for the E8 region located on chromosome 4, between physical positions of 7,000,000 bp – 44,000,000 bp, containing approximately 1000 genes. The most updated version of soybean-PIPE (version 4) [68], was used to identify the top 200 interacting partners for each potential candidate gene. These top 200 interacting partners were analyzed using GO terms, identifying which, or how many, of these 200 interacting partners were involved in time of flowering and maturity. This list was then ranked based on genes with the most interacting partners involved in time of flowering and maturity, providing a short-list of 100 potential candidate genes.

These 100 candidates were then refined further using the combination of our other computational tools, LOF, SNP and RNA sequence data, outlined in Table 2. Observing the top 100 candidate genes the list was further reduced by looking at both LOF data and annotations for *Arabidopsis* and soybean. Any candidate gene that was clearly not involved in time of flowering and maturity for both *Arabidopsis* and soybean and did not contain a LOF mutation was removed. This process continued with analyzing rare SNPs within our

potential candidate genes. Finally, RNA sequence data was analyzed for expression differences to further refine the list and filter out any potential candidates.

This computational approach thus resulted in a manageable list of 18 candidates (process visualized in Figure 8). The computational approach predicted five candidates all among the FAR1/FHY3 family. Additionally, based on their annotation in soybean and *Arabidopsis* all other candidates were found to be heavily involved in time of flowering and maturity.

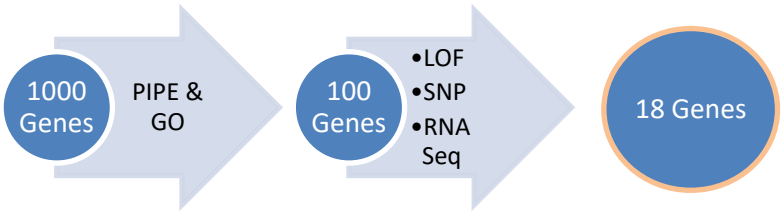


Figure 9: Process of funneling the candidate genes down into the current list of 18 potential E8 candidates.

Table 2: Short-list of candidate genes selected by PIPE, LOF, SNP and RNA sequence databases, each with their annotation in both *Arabidopsis* and soybean, presence of LOF mutations, SNPs, and any available RNA sequence data (available in Appendix 2).

Gene	*Annotation (AT)*	*Annotation (Soy)*	*LOF*	*SNP*	*RNA seq data*
Glyma.04G111200	Far-red impaired responsive	FAR1 DNA-binding domain	NO	N/A	YES
Glyma.04G124300	Far-red elongated hypocotyls 3	SWIM zinc finger	NO	INTRON	YES
Glyma.04G124600	FAR1-related sequence 5	MULE transposase domain	YES	N/A	NO
Glyma.04G101500	Cryptochrome	Cryptochrome	NO	N/A	YES
Glyma.04G126000	Protein Kinase superfamily	Leucine-rich repeat receptor-like kinase	YES	EXON	NO
Glyma.04G138900	Scarecrowe-like-3	GRAS domain family	NO	N/A	YES
*Glyma.04G146000**	Transcription initiation factor	Transcription initiation factor	YES	N/A	YES
Glyma.04G140000	FAR1-related sequence 5	MULE transposase domain	YES	INTRON	NO
Glyma.04G093900	Agamous-like-24	MADS box protein	NO	N/A	NO
Glyma.04G110400	Cytochrome P450	Cytochrome P450 CYP2 subfamily	NO	58 Bp upstream	YES
Glyma.04G120900	N/A	N/A	NO	N/A	NO
Glyma.04G143300	AP2/B3-like transcriptional factor	N/A	YES	N/A	NO
Glyma.04G156400	AP2/B3-like transcriptional factor	N/A	NO	N/A	NO
Glyma.04G099600	VQ Motif	VQ Motif	NO	EXON	YES
Glyma.04G144700	Far-red impaired responsive	FAR1 DNA-binding domain	NO	N/A	NO
Glyma.04G150800	FAR1-related sequence 5	FAR1 DNA-binding domain	NO	N/A	NO
Glyma.04G147500	Integrase-type DNA-binding superfamily	AP2 domain	NO	N/A	YES
*Glyma.04G147300**	DNA helicase PIF1/RRM3	PIF1 helicase	NO	N/A	NO

3.2 Sequencing candidate genes in contrasting lines for E8

Upon sequencing candidate genes from Table 2, 7 of the 18 potential candidates exhibited variations in their coding regions causing amino acid changes. Two genes within this list are still under investigation, *Glyma.04G146000,* due to a very large size of nearly 20 kb, and *Glyma.04G147300*. Table 3 summarizes variation contrasting our E8/e8 lines within each the potential candidates along with any changes to amino acid structure. *Glyma.04G124600* contained a very large number of SNPs along with the presence of a premature stop codon. *Glyma.04G101500* and *Glyma.04G140000* were the only two other genes with premature stop codon, the rest contained amino acid substitutions. There were three genes (*Glyma.04G110400, Glyma.04G156400,* and *Glyma.*04G099600) that presented SNPs however did not account for any amino acid change, therefore removed from further analysis.

Candidates with confirmed variation contrasting E8/e8 lines outlined in Table 3, were then searched for conserved domains using NCBI's conserved domain search (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). This indicates whether or not the variation that was identified resides in an important functional domain. Based on the search, three candidates *Glyma.04G111200, Glyma.04G1243000,* and *Glyma.04G138900* contain amino acid substitutions within their respective functional domains, and three *Glyma.04G123600, Glyma.04G140000* and *Glyma.04G101500* contain premature stop codons on or before the functional domain (Table 4).

Table 3: Summary of sequence variation identified among candidate genes along with resulting amino acid changes, location within intron or exon, and type of variation (SNP, INS, DEL). **Glyma.04G146000* and **Glyma.04G147300* have yet to be sequenced.

Gene	*Intron Variation*	*Exon Variation*	*Amino Acid Change*
Glyma.04G111200	6 SNPs, 3 DEL, 2 INS	1 SNPs	Serine>Proline
Glyma.04G124300	26 SNPs, 2 DEL, 3 INS	4 SNPs	Valine>Isoleucine
Glyma.04G124600	29 SNPs, 2 DEL, 2 INS	51 SNPs, 1 DEL, 1 INS	Premature Stop Codon
Glyma.04G101500	6 SNPs, 3 DEL	2 SNPs, 4 DEL	Premature Stop Codon
Glyma.04G126000	6 SNPs	4 SNPs	Serine>Leucine
Glyma.04G138900	No Variation	6 SNPs	Serine>Tyrosine, Arginine>Glutamine, Lysine>Glutamic Acid, Valine>Isoleucine
*Glyma.04G146000**	TBD	TBD	TBD
Glyma.04G140000	2 SNPs	8 SNPs, 1 INS	Alanine>Serine, Serine>Arginine, Valine>Phenylalanine, Threonine>Alanine, Premature Stop Codon
Glyma.04G093900	No Variation	No Variation	No Variation
Glyma.04G110400	No Variation	1 SNP	No Variation
Glyma.04G120900	3 SNPs, 1 INS	No Variation	No Variation
Glyma.04G143300	No Variation	No Variation	No Variation
Glyma.04G156400	No Variation	1 SNP	No Variation
Glyma.04G099600	8 SNPs, 1 INS	3 SNPs	No Variation
Glyma.04G144700	No Variation	No Variation	No Variation
Glyma.04G150800	No Variation	No Variation	No Variation
Glyma.04G147500	No Variation	No Variation	No Variation
*Glyma.04G147300**	TBD	TBD	TBD

Table 4: Conserved domains associated with each candidate gene. Variation within each functional domain is also shown along with the size of the domain in bp.

Soy Name	*Conserved domain*	*Size of domain (Bp)*	*Variation within domain*
Glyma.04G111200	FAR1 Superfamily	229-495	Serine > Proline AA substitution
Glyma.04G124300	FAR-RED ELONGATED HYPOCOTYL3	1-2529	Valine > Isoleucine AA substitution
Glyma.04G124600	FAR-RED ELONGATED HYPOCOTYL3	4-1260	Premature stop codon early
Glyma.04G140000	FAR-RED ELONGATED HYPOCOTYL3	1-345	Premature stop codon in this domain
Glyma.04G101500	Deoxyribopyrimidine photolyase	22-1446	Many 1 nt. DEL
Glyma.04G101500	Blue/Ultraviolet sensing protein C	1531-1878	Premature stop codon before this domain
Glyma.04G126000	Protein Kinases	975-1752	Serine > Leucine AA substitution before this domain
Glyma.04G138900	GRAS domain family	90-1317	4 AA substitutions within this domain

3.3 2D RNA structure analysis

RNA sequence contains important information for biological function, however RNA molecules must first fold into two dimensional structures (2D) and then three-dimensional structures (3D) to execute their function. Since SNP's could have an effect on RNA structure, analyzing the differences in 2D RNA structure prediction can be used to generate an idea for the differences in gene expression between two genotypes [30]. 2D RNA structure differences were identified among candidates with amino acid variation (Table 3). Considering SNPs, and from this amino acid substitutions, are known to have a functional effect and cause potential changes in 2D RNA folding, the differences in 2D RNA structure was analyzed (Figures 9-14). Most of the candidates had major structural

changes apart from *Glyma.04G111200, Glyma.04G126000* and *Glyma.04G138900*, which were minor (not significant). The differences in structural changes among genotypes for each candidate corresponded to the SNPs within the coding sequence along with 5' and 3' UTR (UnTranslated Region). The more variation presents the more the structure is predicted to change. The minimum free energy for each 2D prediction was therefore recorded to determine if there were any major differences in thermodynamic stability among genotypes, these are indicated in Figures 8-13. Although *Glyma.04G111200* only contained one SNP in the coding region and showed minor structural differences, the MFE differences were among the highest. Interestingly, *Glyma.04G101500*, which contained many deletions had the lowest change in MFE.

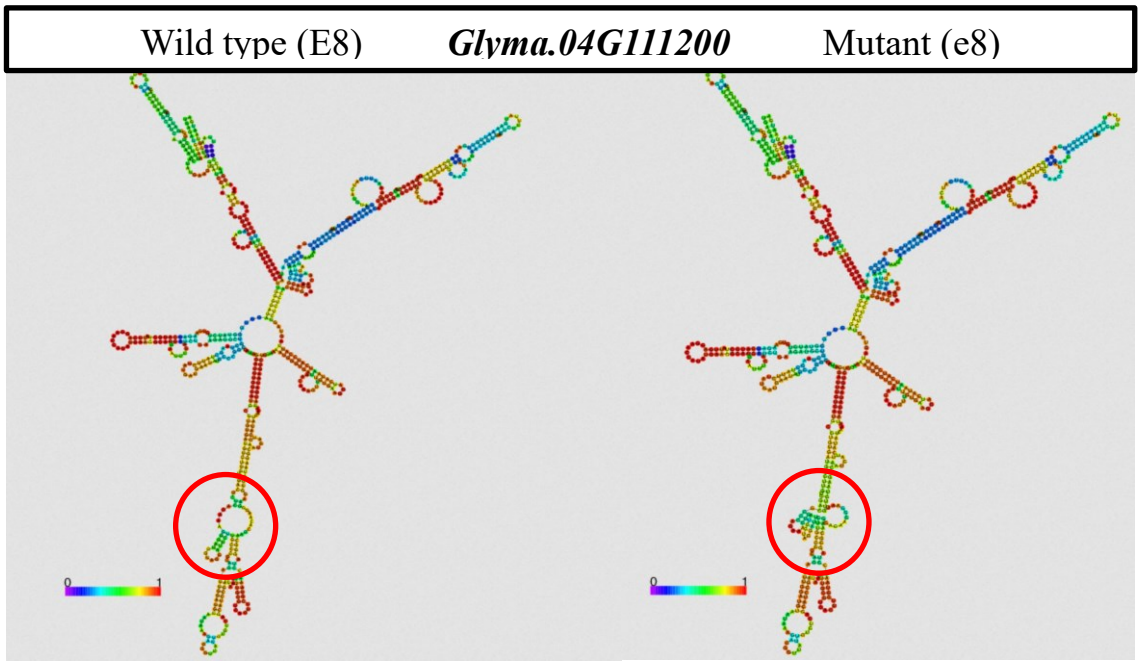


Figure 10: 2D RNA structure prediction of Glyma.04G111200 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -172.90 kcal/mol, and e8 calculated at -175.40 kcal/mol. The only structural change is present from the 1 SNP located in the exon.

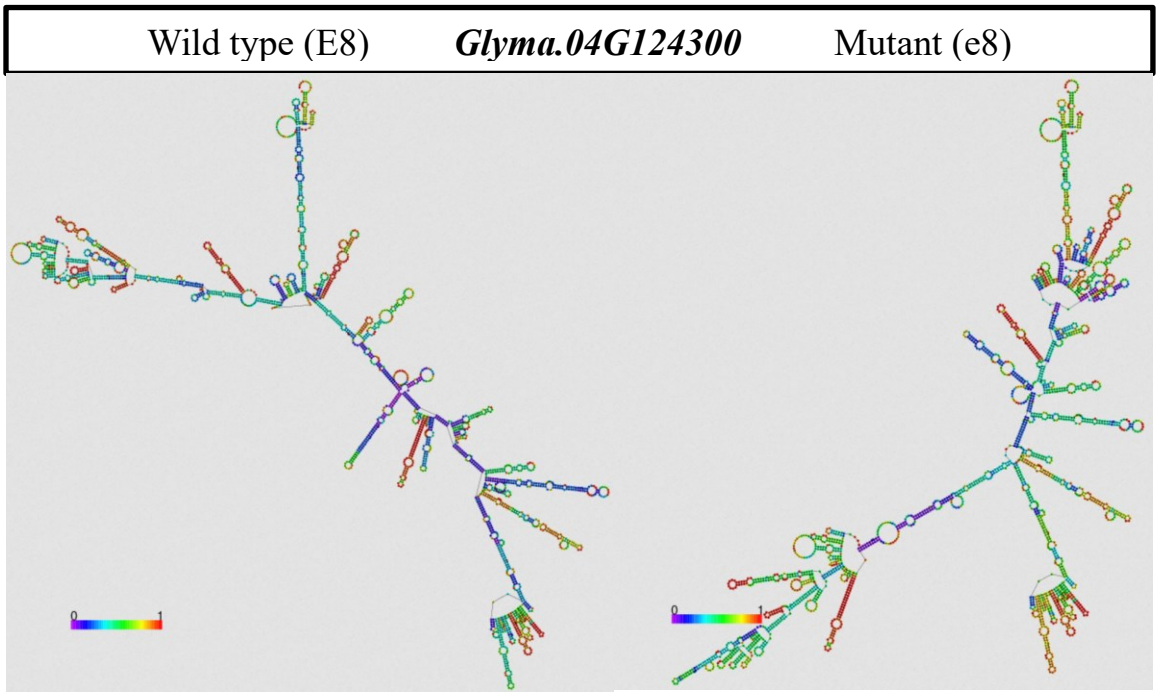


Figure 11: 2D RNA structure prediction of Glyma.04G124300 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -677.10 kcal/mol, and e8 calculated at -680.00 kcal/mol.

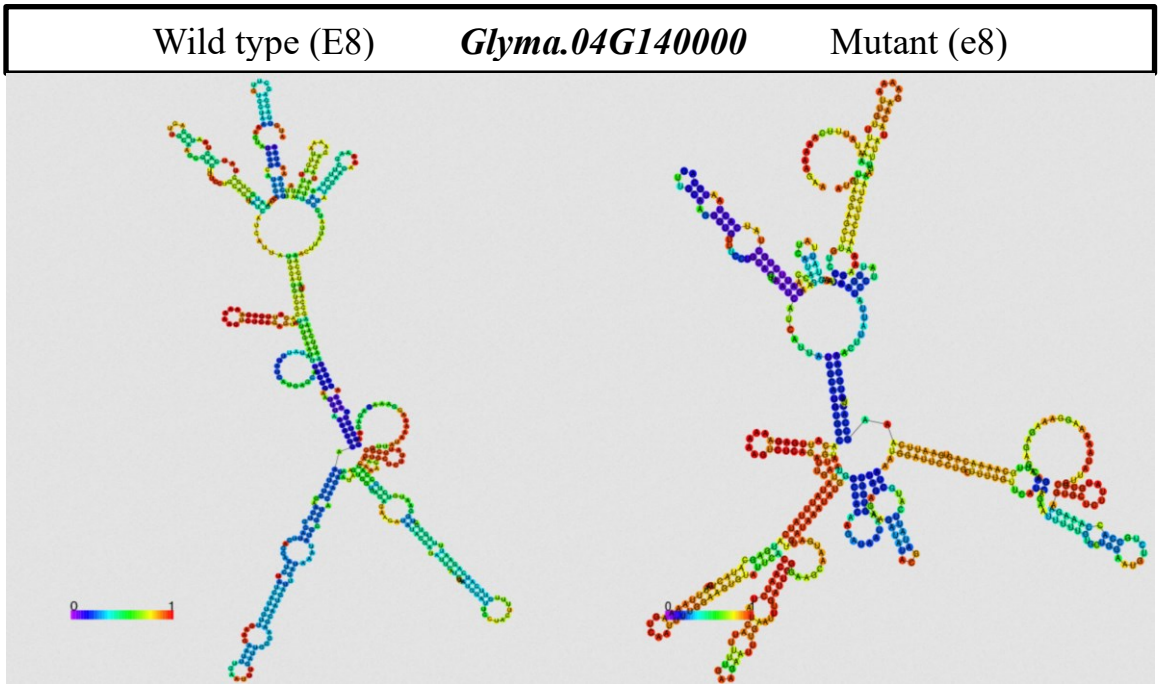


Figure 12: 2D RNA structure prediction of Glyma.04G14000 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -99.20 kcal/mol, and e8 calculated at -96.90 kcal/mol.

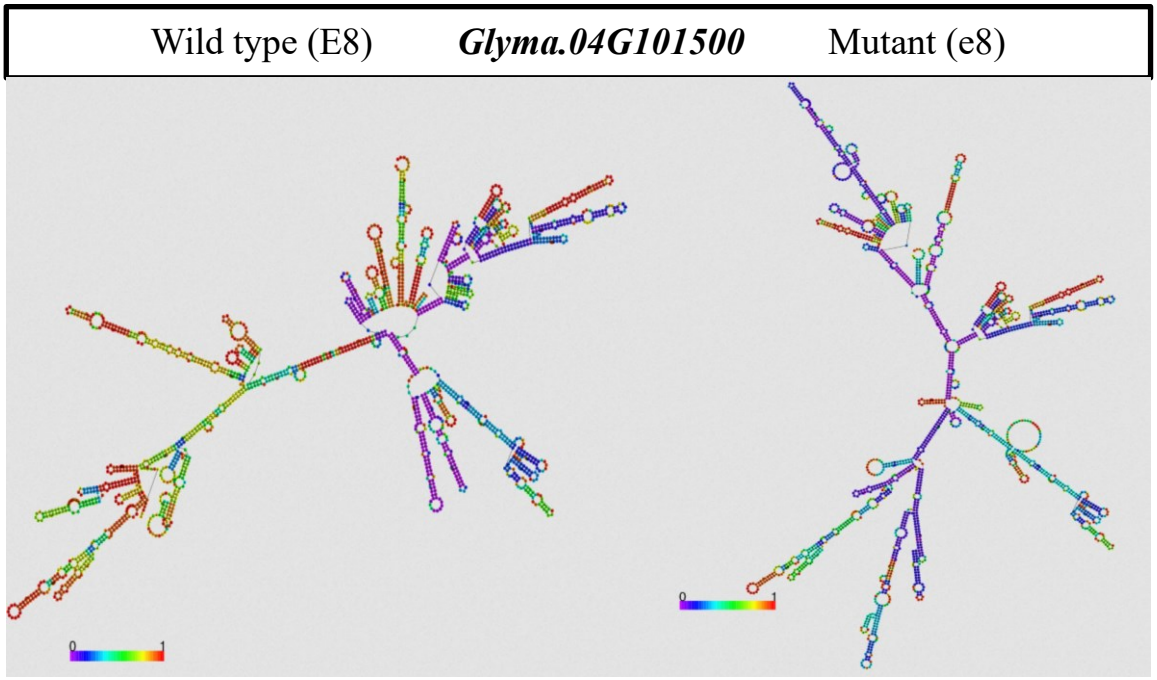


Figure 13: 2D RNA structure prediction of Glyma.04G101500 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -620.70, and e8 calculated at -620.60 kcal/mol.

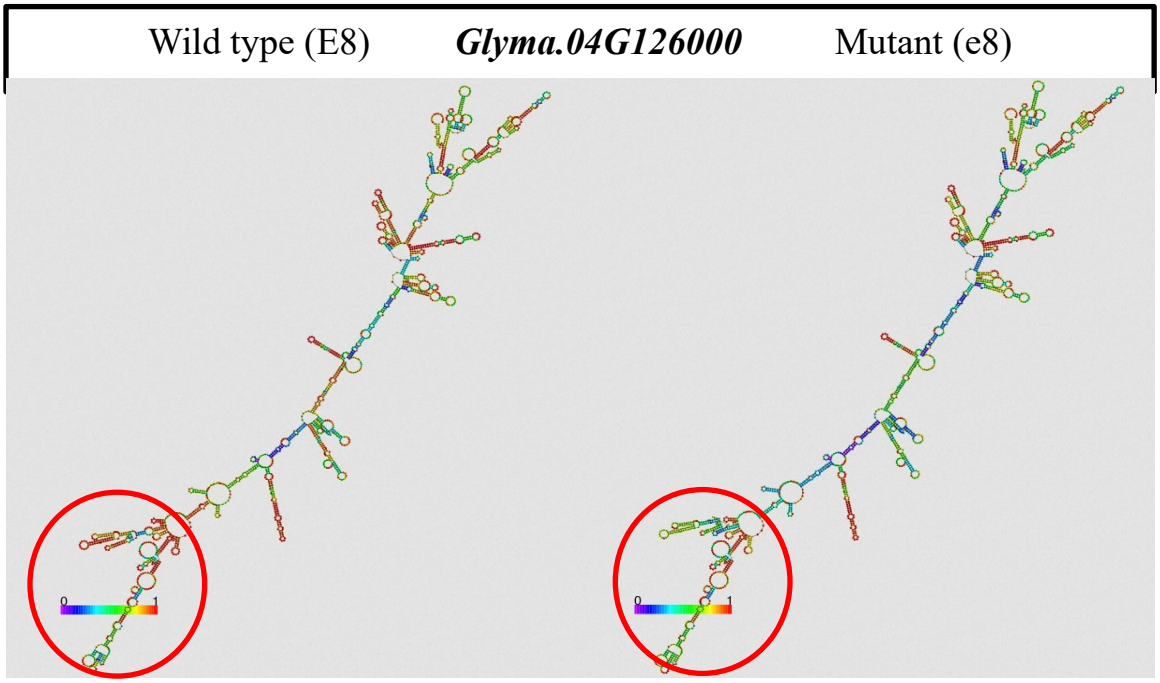


Figure 14: 2D RNA structure prediction of Glyma.04G126000 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -542.20, and e8 calculated at -543.20 kcal/mol.

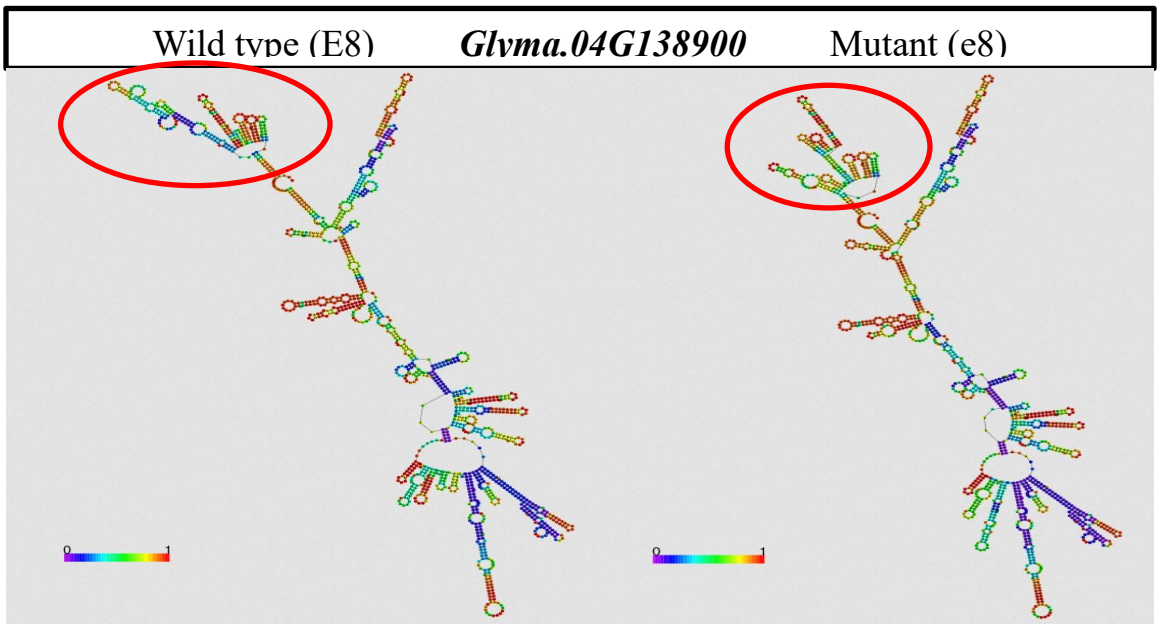


Figure 15: 2D RNA structure prediction of Glyma.04G14000 generated using the RNAfold WebServer from Vienna University (www.rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi). Minimum free energy (MFE) structure was generated with base pairing probabilities, 0 (purple) to 1 (red). MFE for E8 calculated at -378.20 kcal/mol, and e8 calculated at -376.00 kcal/mol.

3.4 Predicting functional effects of amino acid substitutions

The candidate genes shortlisted from Table 3 have been analyzed using PROVEAN, and their function was predicted as either deleterious or neutral, outlined in Table 5. A deleterious prediction indicates a PROVEAN score, or delta alignment score, of less than or equal to -2.5, any value above this threshold and the variation is predicted as neutral [74]. Only two amino acid substitutions were predicted to have a large functional, *Glyma.04G111200* and *Glyma.04G126000*. Candidates, *Glyma.04G124600* and *Glyma.04G101500* contained premature stop codons before any amino changes therefore were not analyzed. Only *Glyma.04G140000* contained amino acid changes before the premature stop codon, however they were not predicted to have a functional effect.

Table 5: PROVEAN prediction of the functional effect from amino acid substitutions. Only two amino acid substitutions were predicted to have a large functional effect, Glyma.04G111200 and Glyma.04G126000.

Gene	*Variation*	*PROVEAN score*
Glyma.04G111200	Serine>Proline	-4.750, **Deleterious**
Glyma.04G124300	Isoleucine>Valine	-0.8840, **Neutral**
Glyma.04G124600	Premature Stop Codon	N/A
Glyma.04G126000	Serine>Leucine	-4.123, **Deleterious**
Glyma.04G101500	Premature Stop Codon	N/A
Glyma.04G138900	Serine>Tyrosine	2.635, **Neutral**
	Arginine>Glutamine	1.124, **Neutral**
	Lysine>Glutamic Acid	-1.616, **Neutral**
	Valine>Isoleucine	-0.181, **Neutral**
Glyma.04G140000	Alanine>Serine	0.230, **Neutral**
	Serine>Arginine	1.504, **Neutral**
	Valine>Phenylalanine	-2.150, **Neutral**
	Threonine>Alanine	0.5330, **Neutral**
	Premature Stop Codon	N/A

3.5 STRING interaction analysis

Additional protein interaction analysis was conducted using the STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database. STRING utilizes publicly available sources of protein-protein interaction information and complements them with computational predictions [80]. In this case STRING was used to identify any experimentally known protein-protein interactions for our candidate's genes. Interestingly only two candidates, *Glyma.04G124300* and *Glyma.04G101500*, have been found to interact with E3 and E4 maturity genes in soybean. STRING showed no available data for *Glyma.04G140000* and *Glyma.04G124600*, with the rest of the candidates only interacting with uncharacterized proteins.

3.6 Expression analysis with qRT-PCR & ddPCR

Based on the results above, 7 candidate genes out of the 18 were chosen for mRNA transcript level analysis (mRNA content); *Glyma.04G111200, Glyma.04G124300, Glyma.04G124600, Glyma.04G101500* and *Glyma.04G126000*, *Glyma.04G140000* and *Glyma.04G138900* to identify differences in gene expression, Figure 15. However, *Glyma.04G140000* and *Glyma.04G138900* are still in the process of completion. Considering the involvement of E8 in photoperiod flowering response, our candidates should exhibit differences in gene expression among samples grown in SD and LD conditions. To extend the identification of this expression difference, samples were also analyzed before and after flowering. Among the 5 candidates, *Glyma.04G124600* and *Glyma.04G126000* revealed the highest expression fold change in samples collected before flowering in short day conditions. Additionally, the most change that appears to occur among these candidates is after the flowering stage. Among these candidates *Glyma.04G124300* indicated a small change in expression.

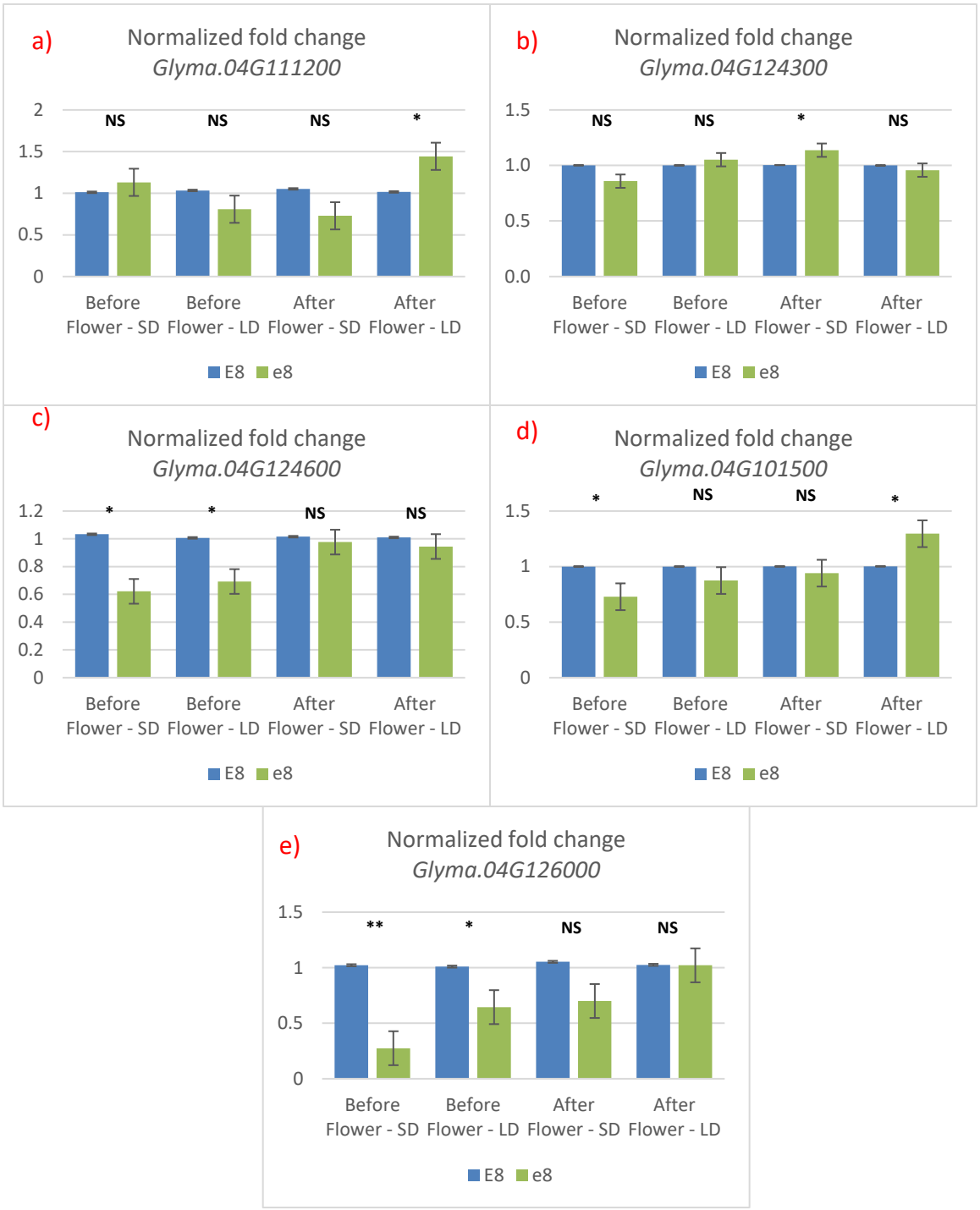


Figure 16: Relative (normalized) fold change for 5 of the candidate genes across 8 different conditions, three biological replicates, using the Pfaffl method with TUBB4 and UKN1 housekeeping genes. The most significant changes occur in the samples collected before flowering in short day photoperiods from Glyma.04G124600, Glyma.04G101500 and Glyma.04G126000. P-value ≤0.05*, P-value ≤0.01, Not significant denoted as NS.**

Additionally, another recent technology that can measure gene expression is Droplet Digital PCR (ddPCR). The requirements for accurate qPCR results involved generating a standard curve to identify efficient amplification between 90 and 110%, which ultimately works based on Cq values generated after each cycle of amplification. Alternatively, ddPCR removes the need to construct a calibration curve and does not measure readings after each cycle, it simply reads the total number of targets generated at the end-product of a standard PCR reaction. This is performed with template, primers, and an intercalating dye such as EvaGreen. These samples are then partitioned into thousands of droplets, each containing at least one copy of the target DNA or zero copies of target DNA. The digital PCR machine then reads how many droplets within a reaction contain the target DNA, allowing for quantification of very low target template and observable differences of expression, which may be low, or highly variable with qPCR. Given the low expression level differences of E8 candidates, ddPCR may be able to highlight more information about expression patterns between samples.

Currently there is only preliminary data from this instrument for one of the candidate genes, *Glyma.04G101500*, Figure 16. The ddPCR results indicate a significant difference in transcript abundance for samples collected after flowering in short day conditions, following the trend of qPCR results, however additional testing will have to be completed along with analysis for our other candidate genes.

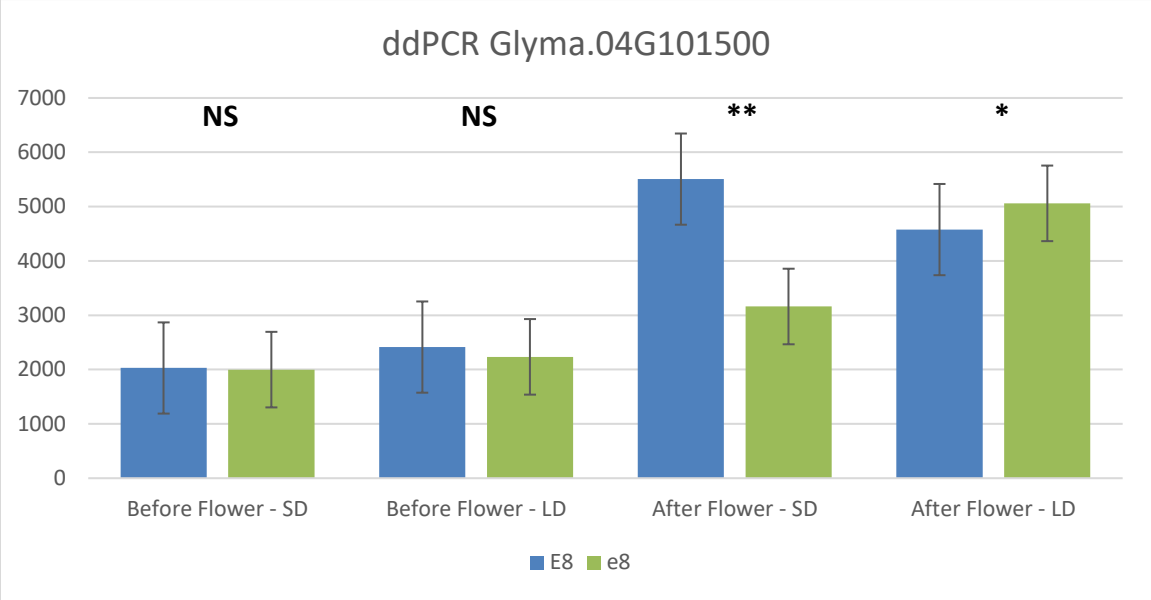


Figure 17: Preliminary ddPCR data for one of the candidate genes, Glyma.04G101500. The number of events correlate to the number of droplets (out of ~10,000) that were positive for the gene of interest. In this case there appears to be a significant change in copy number count for samples collected after flowering in short day conditions. P-value ≤0.05*, P-value ≤0.01, Not significant denoted as NS.**

3.7 Candidate gene summary

Glyma.04G111200 is a 7022 bp gene with a transcript size of 2630 bp spanning 4 introns and 3 exons, visualized below in Figure 17a**.** Figure 17b shows the location of the FAR1 Superfamily functional domain, in respect to the full gene. Sequence data of *Glyma.04G111200* shows that there are several mutations within the intronic regions along with 1 very large 437 bp deletion. In the transcript region there is 1 SNP (T>C) found in the 2nd exon, resulting in a serine to proline amino acid change that occurs within the FAR1 superfamily conserved domain. 2D RNA structure analysis indicates a minor structural change while PROVEAN predicts a deleterious functional effect.

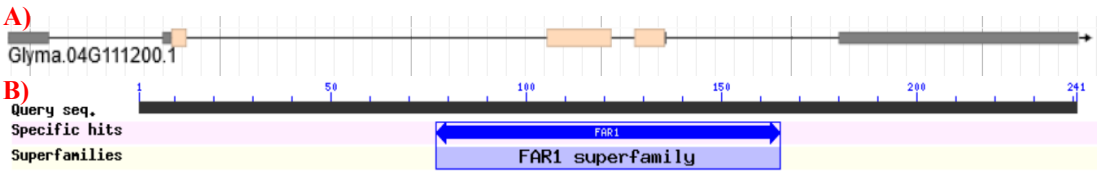


Figure 18: A) Glyma.04G111200 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) FAR1 superfamily conserved domain within Glyma.04G111200 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Glyma.04G124300 is a 7007 bp gene with a transcript size of 3217 bp spanning 8 introns and 7 exons, visualized below in Figure 18a, along with the location of the FHY3 conserved domain which spans the entirety of the gene in Figure 18bSequence data of *Glyma.04G124300* shows a total of 4 SNPs within the CDS region and one in the 3'UTR. One of the SNPs (G>A) in the first CDS region accounts for a valine to isoleucine amino acid substitution which occurs within the FAR-RED ELONGATED HYPOCOTYL 3 conserved domain. Although 2D RNA structure analysis indicates a major structural change, PROVEAN prediction has detected no significant functional changes as a result of the amino acid change. Compared to the other candidates analyzed for expression changes, *Glyma.04G124300* shows very minor differences across samples.

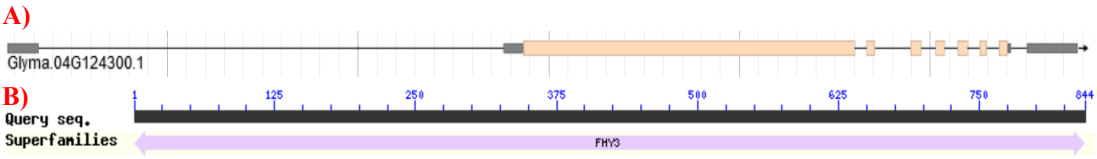


Figure 19: A) Glyma.04G124300 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) FAR1 superfamily conserved domain within Glyma.04G124300 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Glyma.04G124600 is a 2916 bp gene with a transcript size of 1650 bp spanning 7 introns and 6 exons, visualized below in Figure 19a along with the location of the FHY3 conserved domain, Figure 19b, in respect to the full gene. Sequence data of *Glyma.04G124600* show a large amount of variation within the mutant lines with one, 25bp DEL in the first CDS along with 51 SNPs across the other CDS regions alone. This results in a premature stop codon very early in the amino acid sequence (19 amino acids in) due to the 25bp DEL. The premature stop codon occurs within the FAR-RED ELONGATED HYPOCOTYL3 domain and for this reason a functional prediction was not tested, however 2D RNA structure has indicated a significant change. Differences in gene expression are only present between samples collected before and after flowering.

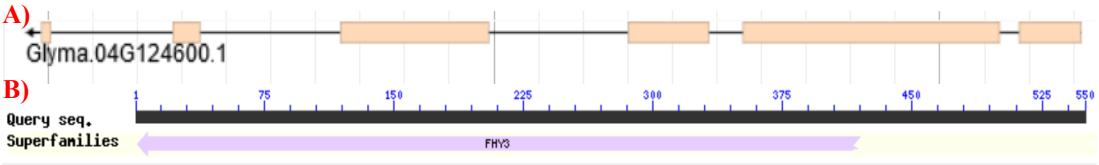


Figure 20: A) Glyma.04G124600 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) FHY3 conserved domain within Glyma.04G124600 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Glyma.04G140000 is a 582 bp gene with a transcript size of 501bp spanning 1 intron and 2 exons, visualized below in Figure 20a, with the location of the FHY3 superfamily, Figure 20b, in respect to the full gene. Sequence data for *Glyma.04G140000* shows 7 SNP's and one insertion (T) in the CDS region. There are 4 SNPs that result in amino acid substitutions (alanine > serine, serine > arginine, valine > phenylalanine, and threonine > alanine) until the eventual premature stop codon after the T insertion. 2D RNA

structure is entirely different between genotypes and the functional effects of the amino acids before the premature stop codon appear minor.

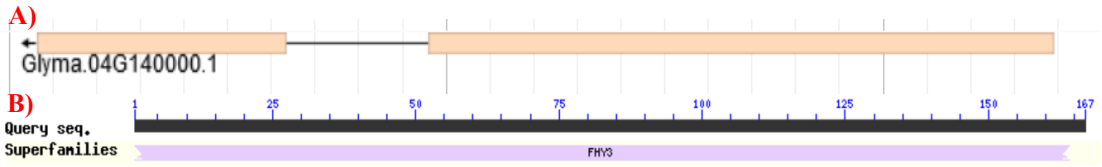


Figure 21: A) Glyma.04G140000 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) FHY3 conserved domain within Glyma.04G140000 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Glyma.04G101500 is a 4518 bp gene with a transcript size of 3001 bp spanning 4 introns and 4 exons visualized below in Figure 21a, with the PhrB and cryptochrome C superfamilies, Figure 21b, in respect to the whole gene. Sequence data for *Glyma.04G101500* revealed few SNP's that do not result in any amino acid substitutions, however there is a series of 3 single nucleotide DEL that result in a premature stop codon after the 139th amino acid that occurs before the blue/ultraviolet sensing protein domain. 2D RNA structure between genotypes for this gene is very different and no functional prediction was generate due to the premature stop codon. Expression analysis shows an increase with the mutant phenotype in only the samples collected after flowering in LD and ultimately stays consistent with the ddPCR results. However, ddPCR was able to detect a large reduction in expression of the mutant genotype in samples collected after flowering under SD.

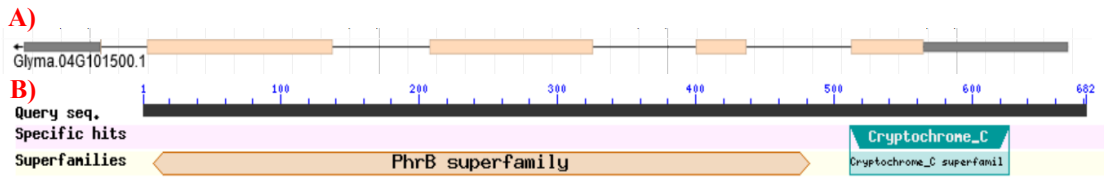


Figure 22: A) Glyma.04G1010500 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) PHrB and Crpytochrome C superfamily conserved domain within Glyma.04G101500 in respect to the whole gene. Figure taken from NCBI's conserved domain search

Glyma.04G126000 is a 6067 bp gene with a transcript size of 2097 bp spanning 8 introns and 9 exons, visualized in Figure 22a, with the location of the PKc-like superfamily domain, Figure 22b, in respect to the full gene. Sequence data for *Glyma.04G126000* shows 4 SNPs in the CDS region, two of which are directly beside each other (T>C and a C>T) in the 5th CDS region. This results in a serine to leucine amino acid substitution. This substitution is not predicted to have a large functional effect, which is supported by 2D RNA structure prediction showing minimal to no changes. However, there is a very large expression fold change between genotypes from samples collected before flowering under SD and LD, and after flowering under SD.

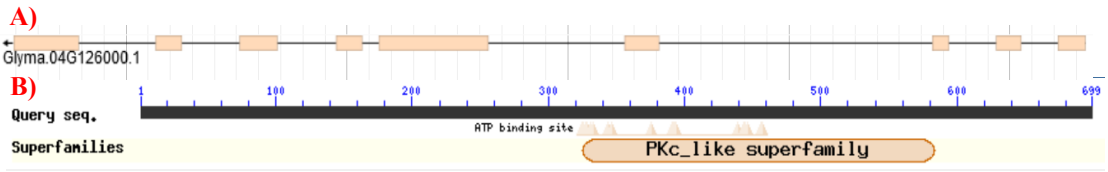


Figure 23: A) Glyma.04G126000 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) PKc like superfamily conserved domain within Glyma.04G126000 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Glyma.04G13900 is a 1332 bp gene with no introns or UTRs, it can be visualized below in Figure 23a, with the location of the GRAS superfamily domain, Figure 23b, in respect to the full gene. Sequence data for *Glyma.04G138900* shows 6 SNPs in the CDS region, 4 of which result in amino acid substitutions (tyrosine > serine, glutamine > arginine, glutamic acid > lysine, and isoleucine > valine). 2D RNA structure analysis detected minor changes among genotypes and no major functional effect was predicted as a result of these amino acid changes.

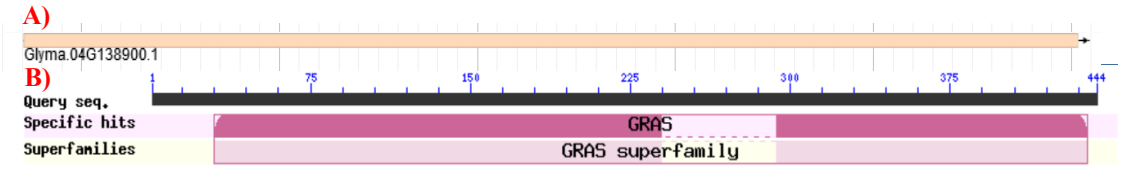


Figure 24: A) Glyma.04G138900 intron(blank spaces)/exon(beige) map with UTR's (grey). Figure taken from Phytozome.org B) GRAS superfamily conserved domain within Glyma.04G138900 in respect to the whole gene. Adopted from NCBI's conserved domain search (March 2020).

Chapter 4: Discussion

Soybean is an economically important crop for its wide array of uses and nitrogen fixating capabilities. The northern regions of western Canada can benefit greatly from this however soybean struggles to grow in the higher latitudes of these regions for several reasons. Adapting soybean for growth in Northern Canada requires extensive knowledge of the genetics underlying various pathways. More specifically in this case genetic pathways responsible for time of flowering and maturity. With 11 loci known to have substantial effect on time of flowering and maturity in soybean, the molecular basis for E6, E7, E8 and E11 are still unknown. Additionally, there are many QTLs known to play a function in soybean growth and development, however, most remain uncharacterized. Although E8 is just one of these loci, identifying the underlying gene here is among the first steps in solving a highly complex puzzle.

4.1 Analysis of candidate genes

The search for the underlying gene responsible for the E8 loci began with a thorough computational investigation that highlighted 18 potential candidate genes out of our initial 1000 candidate gene pool. Upon sequencing these candidate's half were eliminated from the discussion as they did not contain any amino acid variation among contrasting genotypes (E8/e8). Ultimately, 7 candidates remain, with strong possibility that our gene of interest lies among *Glyma.04G124600, Glyma.04G140000* and *Glyma.04G101500*. Considering the majority of the maturity loci contain recessive dysfunctional alleles that promote flowering, these were immediately placed near the top

of our list for further compensation analysis. However, the other 4 candidates are not eliminated from the discussion and will be further investigated if necessary. Additionally, two genes (*Glyma.04G146000* and *Glyma.04G147300*) have yet to be completely sequenced and therefore will also remain as backup candidates.

4.1.1 *FAR1/FHY3* family of genes

The computational analysis identified an interesting family of proteins, the *FAR-RED IMPAIRED RESPONSE 1 (FAR1)* and *FAR-RED ELONGATED HYPOCOTYL 3 (FHY3)* family. These include *Glyma.04G111200, Glyma.04G124300, Glyma.04G124600* and *Glyma.04G140000*. *FAR1/FHY3* have previously been identified to have crucial functions in plant growth and development in *Arabidopsis* with involvement in light signal transduction, photomorphogenesis, circadian clock and flowering time regulation, shoot meristem and floral development [81]. *FHY3* and *FAR1* are transcription factors derived from *Mutator-like element (MULE)* transposases and also contain a SWIM zinc finger domain, both of which are required for transcriptional activation of their target genes including; *EARLY-FLOWERING4 (ELF4), COP1, FAR-RED ELONGATED HYPOCOTYLS1(FHY1), FHY1-LIKE(FHL). FHY3* and *FAR1* activate transcription of *FHY1* and *FHL*, which are regulators of *PHYA* translocation under far-red light, therefore indirectly regulate *PHYA*. Additionally, loss-of-function mutants *fhy3* and *far1* were shown to alter expression of circadian clock-related genes [81]. Altogether, previous studies conducted on the *FAR1/FHY3* family in *Arabidopsis* has suggested a significant role in growth and development and therefore support the possibility of their potential involvement of E8 in soybean.

Among the FAR1/FHY3 family, *Glyma.04G124600* and *Glyma.04G14000* contained premature stop codons among or before their conserved domains along with identified LOF mutants within our database, suggesting potential altered function. *Glyma.04G124600* and *Glyma.04G140000* both are annotated as *FAR1 RELATED SEQUENCE 5* in *Arabidopsis* and contain a MULE transposase domain in soybean thus suggesting they may both have an involvement in the early E8 flowering phenotype. *Glyma.04G124600* revealed a substantial expression fold change in samples collected before flowering which then seemed to normalize once flowering had occurred. This suggests that this gene may be involved in pre flowering response and *FAR1* may be indirectly downregulating *PHYA* expression, thus allowing early flowering of soybean by further increasing photoperiod insensitivity. Ouyang et al., 2011 found that FAR1/FHY3 have more than a thousand putative targets in *Arabidopsis* suggesting a very broad function in plant growth and development [82], [83]. Considering *e8* was found to respond to FR-enriched LD conditions promoting flowering, the lower expression levels of *Glyma.04G124600* in our before flowering mutant samples may indicate a role in photomorphogenesis until flowering occurs. Although expression fold change is not yet available for *Glyma.04G140000*, its mode of action may be similar to that of *Glyma.04G124600* and therefore also presents a very likely candidate for E8.

The other FAR1/FHY3 candidates, *Glyma.04G111200* and *Glyma.04G124300* only contained a single amino acid substitution and although they show significant changes in 2D RNA structure, only the amino acid variation for *Glyma.04G111200* is predicted to have a functional effect. Additionally, expression changes among daylength and growth stages for *Glyma.04G124300* were minor, while *Glyma.04G111200* indicated more of a

difference with a flip in expression profile among SD and LD conditions between growth stages. The increase in expression fold change after flowering in LD conditions among the mutant suggests that *Glyma.04G111200* may also have an influence on growth and development post flowering under LD and therefore promote an early maturity.

Protein interaction prediction using STRING indicated that *Glyma.04G124300* has been identified to interact with E3 and E4 maturity genes thus suggesting the other *FAR1/FHY3* candidates may also be interacting with E3 and E4 due to their sequence homology. This coincides with studies done in *Arabidopsis* where *FAR1/FHY3* positively regulate *PHYA* signaling, along with a known involvement of e8 in photoperiod response, suggesting that the candidate for E8 may lie among these 4 genes. However, based on sequence results showing a premature stop codon in *Glyma.04G124600* and its interesting expression data, it has been chosen as a candidate for further analysis. *Glyma.04G140000* also contained a premature stop codon, however until expression results are obtained and compensation analysis does not indicate any major changes in *Arabidopsis* for *Glyma.04G124600*, *Glyma.04G140000* will be placed on hold. Ultimately, *Glyma.04G124600* will be further analyzed for compensation in *Arabidopsis* until/if promising expression results are obtained for *Glyma.04G140000*.

4.1.2 **Cryptochrome**

Glyma.04G101500 encodes *Cryptochrome*, a blue-light photosensory receptor that mediates non-photosynthetic light responses including growth, development, and the circadian clock [84]. In Arabidopsis cryptochromes mediate blue light inhibition of hypocotyl elongation, floral initiation, circadian rhythms and acts to stabilize *CO*

abundance by interacting with *COP1/SPA* to prevent their repression [19]. The soybean genome encodes at least 7 cryptochromes, four CRY1 (CRY1a, CRY1b, CRY1c, and CRY1d), and three CRY2 (CRY2a, CRY2b and CRY2c) proteins [85]. Zhang et al., 2008 [86], found that *GmCRY1a (Glyma.04G101500)* affected blue light inhibition of cell elongation and its abundance was associated with days to flowering. It was also identified that transgenic *Arabidopsis* plants expressing *GmCRY1a* accelerated flowering by stimulating mRNA expression of *FT* and that *GmCRY1a* mRNA and protein expression were found to associate with circadian rhythms in soybean. He also identified that *GmCRY1a* rescued a late-flowering *cry2* mutant in *Arabidopsis* thus suggesting *GmCRY1a* promotes floral initiation. Although Zhang et al., 2008 [86] discovered that *GmCRY1a* had a large impact on floral initiation that oscillated with circadian rhythm and changed expression based on photoperiod, there was no QTL associated with this gene at the time and the cultivars used may not have contained the *e8* allele.

Glyma.04G101500 was one of the three final candidates with a premature stop codon, which occurred before an important blue/ultraviolet sensing protein C, suggesting truncated activity involved in blue light response. Additionally, 2D RNA structure analysis had indicated a significant change in structure, which was most likely due to the amount of variation present throughout the gene. *Glyma.04G101500* was also found to interact with the E3 and E4 maturity loci which may be because of a previous study suggesting the involvement of phytochromes in posttranslational regulation of *GmCRY1a* protein expression [86]. Expression results from qPCR for *Glyma.04G101500* appear to flip expression profiles among genotypes after flowering has occurred. These results indicate that *Glyma.04G101500* is affected by changes in photoperiod among genotypes, matching

the results collected by Zhang et al., 2008 [86], thus strongly supporting the probability of this gene as the underlying candidate for the E8 locus. That being said, the preliminary ddPCR results may suggest a different conclusion. Expression results produced by ddPCR indicate a repressed activity of the *e8* allele in SD conditions after flowering, while all other growth stages and photoperiod conditions remain relatively even. This suggests that *Glyma.04G101500* may have a larger involvement in post flowering response under SD conditions, however further testing will be required.

4.1.3 **Receptor like kinase (RLK)**

Glyma.04G126000 is a receptor-like kinase (RLK) that mediates many cellular signals in plants. One of the largest of these RLK family's is the leucine-rich repeat receptor like kinase family (LRR-RLK), which is what *Glyma.04G126000* is annotated as in Arabidopsis [87]. LRR-RLKs are involved in numerous plant processes including plant growth and development and abiotic and biotic stress response. LRR-RLKs thus, play an essential role in adaptive mechanisms during environmental stress. However the majority of these genes functions are still largely unknown [87]. Considering 467 genes have been identified in soybean encoding RLKs to date, predicting the function of *Glyma.04G126000* becomes much more difficult as its functional annotation is so broad [88]. However, there have recently been studies done on RLKs and LRR-RLKs in soybean that identified 13.1%, 11.3%, 20.3%, 18.6%, and 10.9% of all RLKs are expressed in nodules, leaves, stems, SAM, and flowers, respectively, suggesting a strong association with plant growth and development [88]. A previous study reported by Lee at al., 2016 [89], identified mutated flowering genes in soybean shown to affect *GmFT2a* (florigen initiating hormone)

expression resulting in early flowering. One of these genes in the study, *Glyma.14G105300* (*ELF3*), also exhibited a serine to leucine amino acid change and therefore suggests the possibility of this same mutation in *Glyma.04G126000* being responsible for the early flowering phenotype in the E8 mutant lines

This gene was chosen as a potential candidate for the underlying gene for E8 based on having on PIPE interaction results with many genes involved in time of flowering and maturity, which may be due to the large RLK family within the soybean genome. However, our LOF analysis also indicated a LOF mutant was present within this gene, which was not identified after sequencing contrasting lines. Additionally, a SNP was identified in our SNP database within an exon of this gene, however this was also found to be a different SNP after sequencing. Structurally, the wild type and mutants of this gene were very similar upon 2D RNA structure analysis and the resulting amino acid substitution was not found to have an effect on overall gene function. However, expression results here indicated a significant fold change among genotypes in all samples apart from the ones collected after flowering in LD conditions. Expression of the mutant appeared to be significantly reduced in sampled collected before flowering in SD conditions suggesting there may be some involvement in the early flowering phenotype. Ultimately, it remains difficult to draw a conclusion on the function of this gene due to the broad range of roles associated with soybean RLKs.

4.1.4 **GRAS family domain**

Glyma.04G138900 is annotated as *SCARECROW-LIKE 3 (SCL3)* in *Arabidopsis* and part of the *GRAS* domain family in soybean, acting as a positive regulator of

gibberellins, a class of phytohormones that regulate plant growth and development. *SCL3* is also known as part of *GRAS [GIBBERELLIC ACID INSENSITIVE (GAI), REPRESSOR OF GA1-3 (RGA), and SCARECROW (SCR)]* transcription factor which overall play diverse roles in in growth and development [90]. *Glyma.04G138900* is annotated as *SCL3* in Arabidopsis and contains a *GRAS* conserved domain suggesting that this gene has some form of role in flowering and development. Previously, *GRAS* transcription factors have been found to play a role in phytochrome A signal transduction in *Arabidopsis*. Additionally, the majority of *GRAS* family members in soybean were found to have a high expression in aerial tissues such young lead and flowers. Within the *GRAS* family are *GAI* and *RGA*, both of which have been found to act as major repressors of vegetative growth and floral induction [90]. The glutamine to arginine mutation found within this gene has previously been identified in human cytomegalovirus (HCMV) by Burgdorf et al., 2011 [91], where it resulted in severely debilitating the virus, impacting growth rate. The glutamic acid to lysine mutation was previously reported in *Arabidopsis* within a ribosomal export protein (MDN1/Rea1) by Li et al., 2016 [92]. This mutation was shown to be extremely conserved among organisms within MDN1 where it impaired expression of genes relate to plant growth and development in *Arabidopsis.*

Although these amino acid substitutions have been previously found in literature to significantly impact gene function, there was no functional effect predicted by PROVEAN. Additionally, the 2D RNA structure comparison between the two genotypes only contained a very minor structural variation. Without further expression results, conclusions cannot be made on the function of this gene in soybean and its potential involvement in E8. Originally this gene was selected as a candidate due to interesting PIPE results, and although no LOF

or SNP data was available, the RNA sequence results, Appendix 2, indicated a significant drop in expression after flowering. Considering the strong involvement of GRAS transcription factors on vegetative growth and floral induction, this gene remains a strong candidate for further analysis.

4.2 Literature curated candidate genes

Of the four genes that were identified through literature (*Glyma.04G143300*, *Glyma.04G156400*, *Glyma.04G147500* and *Glyma.04G147300*), one requires a further investigation and cannot yet be ruled out. *Glyma.04G143300* and *Glyma.04G156400,* which correspond to *E1Lb* and *E1La*, respectively, and are annotated as AP2/B3-like transcription factors in soybean, were suggested as two potential E8 candidates [35]. Both of these genes were located within the E8 region and are homologous to the E1 maturity gene. Additionally, these two genes (E1La & E1Lb) were found to downregulate *GmFT*2a and *GmFT5a* expression, thus delaying flowering under LD. However, upon sequencing E1La and E1Lb accross contrasting E8 lines, no sequence variation was detected and therefore ruled out as potential E8 candidates.

The other two literature curated genes, *Glyma.04G147500* and *Glyma.04G147300*, were recently identified in close proximity to a SNP associated with early time of flowering and maturity from GWAS data provided by Bruce et al., 2020 [93]. Although no genes were present around the SNP, the two closest candidates (~100 kb away) were selected for additional analysis as they were found to play a role in flowering in *Arabidopsis.* Currently, only *Glyma.04G147500* has been sequenced, which showed no sequence variation and therefore has been ruled out. However, *Glyma.04G147300,* which is annotated as *PIF1*

helicase, in *Arabidopsis* is still under investigation and has yet to be sequenced. *PIF1* (*PHYOCHROME-INTERACTING FACTOR1*), negatively regulates photomorphogenesis and chlorophyll biosynthesis in *Arabidopsis* and therefore remains an interesting candidate for E8. Overall, one literature curated gene remains (*Glyma.04G147300*) as a recommended candidate (with a very low confidence) and may be investigated further if necessary.

4.3 Final list for further analysis

Altogether the data has provided a list of 3 most likely candidates for the E8 locus, *Glyma.04G124600, Glyma.04G140000* and *Glyma.04G101500*. However, until expression analysis is completed for *Glyma.04G140000, Glyma.04G124600* and *Glyma.04G101500* =will continue with further analysis, outlined in future directions. The other four candidate genes, *Glyma.04G111200, Glyma.04G124300, Glyma.04G126000,* and *Glyma.04G138900* will be placed among the next potential candidates if required.

Chapter 5: Conclusion and future direction

5.1 Conclusion

Altogether PIPE has proven to successfully identify candidates with distinct involvement in time of flowering and maturity, and has previously established itself successful in the identification of the underlying gene for E10 [30]. With the vast availability of functional genomics data and the continuously increasing capabilities of computational tools it has become essential to use these resources in modern biology. The combination of computational tools and resources were key components in continuously refining and short listing the potential candidate genes underlying E8.

Based on the sequence data all seven candidates could serve as potential underlying genes for E8 due to the presence of amino acid variation. The *FAR1/FHY3* family was the largest family identified by the computational approach, with six original candidates out of 18. Based on the collective data from sequencing, prediction tools and expression analysis, *Glyma.04G124600* and *Glyma.04G140000* are the leading candidates from this family of genes simply due to the presence of premature stop codons. *Glyma.04G101500,* the cryptochrome gene, *GmCRY1a*, also contained a premature stop codon and with the previous results published by Zhang et al., 2008 [86] remains a highly interesting candidate. Altogether, *Glyma.04G124600,* and *Glyma.04G101500* are moving forward with compensation analysis while expression results await for *Glyma.04G140000*. The data regarding the other candidate genes, *Glyma.04G111200*, *Glyma.04G124300, Glyma.04G126000* and *Glyma.04G138900* suggest they may also be likely candidates and will remain as potential backup candidate genes.

5.2 Future direction

5.2.1 Blue/Red light experiment

Glyma.04G124600 and *Glyma.04G101500* encode *FAR1/FHY3* and *CRY1a*, respectively. Based on previous literature loss of function *Arabidopsis* mutants at these genes were found to have an effect on hypocotyl elongation [94], [95]. When grown under low R:FR light ratios (shade), *far1* and *fhy3 Arabidopsis* mutants developed an elongated hypocotyl. Similarly, *cry1 Arabidopsis* mutants grown under continuous blue light also exhibited the same hypocotyl elongation. Thus, future experiments are under development to grow E8 and e8 soybean lines in both R:FR light and continuous blue light and assess their hypocotyl length. If there is a distinct difference in hypocotyl length of the mutant (e8) lines, under either light source then it can be confirmed that the underlying gene for the E8 locus is either *Glyma.04G1246000* or *Glyma.04G101500*.

5.2.2 Transformation into *Arabidopsis*

In parallel with the blue/red light experiment, the process to transform *Glyma.04G101500* and *Glyma.04G124600* into *Arabidopsis* to confirm their involvement on time of flowering and maturity has begun. The transcript sequences are in the process of being cloned into the Zero Blunt™ TOPO™ cloning vector from ThermoFisher for sequencing.

5.2.3 **Sequencing and expression analysis**

To be thorough, sequencing the remaining candidate genes will continue to detect for any amino acid variation. These may not be top priority candidates; however, they cannot be disregarded in the event *Glyma.04G124600, Glyma.04G140000* and *Glyma.04G101500* are proven to not be the candidates for E8. Additionally, qPCR will be performed on the remaining candidate genes. Since leaf samples were also collected during flowering, adding another growth stage to the expression results may further reveal the expression profiles of the candidate genes. The ddPCR was largely different over the qPCR data, most likely due to Taq polymerase inhibition or lower detection capabilities in qPCR. Ultimately ddPCR has proven to be more accurate in detecting minor expression differences, performing an additional step using this method may further aid in predicting the function of the candidate genes.

References

[1] M. C. Pagano and M. Miransari, "The importance of soybean production worldwide" *Abiotic and Biotic Stresses in Soybean Production*, pp. 1-26, 2016

[2] Soy Story: A Short History of Glycine max in Canada, *Statistics Canada*, Catalogue No. 21-004-x2017001, March 3rd 2017. [Accessed: 2019-04-19]

[3] Agricultrue and Agri-Food Canada, Canada: Outlook for Principal Field Crops 2019-10-18, [Online] Available: https://www.agr.gc.ca/resources/prod/doc/misb/mag-gam/fco-ppc/fco-ppc_20191018-eng.pdf. [Accessed: 2019-12-06]

[4] Statistics Canada. Agriculture Division, "Table 32-10-0359-01 Estimated areas, yield, production, average farm price and total farm value of principal field crops, in metric and imperial units," 2019. [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=3210035901. [Accessed: 2020-06-23].

[5] G. L. Hartman, E. D. West, and T. K. Herman, "Crops that feed the World 2. Soybean-worldwide production, use, and constraints caused by pathogens and pests," *Food Security.*, vol. 3, no. 1, pp. 5–17, 2011.

[6] Statistics Canada. Agriculture Division, "Principal field crop areas , June 2019," 2019. [Online]. Available: https://www150.statcan.gc.ca/n1/en/daily-quotidien/190626/dq190626b-eng.pdf?st=qrJQD5lX. [Accessed: 2020-05-21]

[7] SoyCanada, "10 Million Acres of Opportunity is poised for explosive growth .," [Online] Available: http://soycanada.ca/wp-content/uploads/2017/04/10-Million-Acres-of-Opportunity-Discussion-Paper.pdf. [Accessed: 2019-11-16]

[8] M. Gabruch and R. Gietz, "The Potential for Soybeans in Alberta," 2014.[Online] Available: https://www1.agric.gov.ab.ca/$department/deptdocs.nsf/all/bus15100/$file/soybeans-1.pdf?OpenElement#:~:text=Overall%2C%20soybeans%20are%20a%20crop,slower%20than%20for%20grain%20corn.&text=Soybeans%20are%20sometimes%20referred%20to,response%20to%20shortened%20day%20length.. [Accessed: 2019-11-16]

[9] G. Wolfgang and Y. qiang C. An, "Genetic separation of southern and northern soybean breeding programs in North America and their associated allelic variation at four maturity loci," *Molecular Breeding.*, vol. 37, no. 1, 2017.

[10] G. S. Golembeski, H. A. Kinmonth-Schultz, Y. H. Song, and T. Imaizumi, "Photoperiodic Flowering Regulation in Arabidopsis thaliana", *The Molecular Genetics of Floral Transition and Flower Development Advances in Botanical Research,* pp. 1-28, 2014.

[11] S. Kumudini, "Soybean Growth and Development," *The soybean: botany, production, and*

uses, pp. 48–73, 2010.

[12] W. R. Fehr and C. E. Caviness, “Stages of Soybean Development,”*, Special Report.* 87, 1977

[13] X. Zhang, H. Zhai, Y. Wang, X. Tian, Y. Zhang, H. Wu, S. Lu, G. Yang, Y. Li, L. Wang, B. Hu, Q. Bu, and Z. Xia, “Functional conservation and diversification of the soybean maturity gene E1 and its homologs in legumes,*″ Scientific Reports*, vol. 6, no. 1, 2016,

[14] Z. Xia, H. Zhai, and B. Liu, F. Kong, X. Yuan, H. Wu, E. R. Cober, and K. Harada, “Molecular identification of genes controlling flowering time , maturity , and photoperiod response in soybean,” *Plant Systematics and Evolution,* vol. 298, no. 7, pp. 1217–1227, 2012.

[15] W. A. Rensink and C. R. Buell, “Arabidopsis to rice. Applying knowledge from a weed to enhance our understanding of a crop species,” *Plant Physiology*, vol. 135, no. 2, pp. 622–629, 2004.

[16] R. Shrestha, J. Gomez-Ariza, V. Brambilla, F. Fornara, “Molecular control of seasonal flowering in rice, arabidopsis and temperate cereals,” *Annals of Botany,* vol. 114, no. 7, pp. 1445–1458, 2014.

[17] M. Blümel, N. Dally, and C. Jung, “Flowering time regulation in crops — what did we learn from Arabidopsis?,” *Current Opinion in Biotechnoology*, vol. 32, pp. 121–129, 2015.

[18] H. Tsuji and K.-I. Taoka, "Florigen signaling", *Signaling Pathways in Plants the Enzymes,* pp. 113-144, 2014.

[19] D.-H. Kim, “Current understanding of flowering pathways in plants : focusing on the vernalization pathway in Arabidopsis and several vegetable crop plants,” *Horticulture, Enviornment, and Biotechnology,* vol. 71, pp. 209–227, 2020.

[20] Q. Li, C. Fang, Z. Duan, Y. Liu, H. Qin, J. Zhang, P. Sun, W. Li, G. Wang, and Z. Tian, “Functional conservation and divergence of GmCHLI genes in polyploid soybean,” *The Plant Journal,* vol. 88, no. 4, pp. 584–596, 2016.

[21] R. Takeshima, T. Hayashi, J. Zhu, C. Zhao, M. Xu, N. Yamaguchi, T. Sayama, M. Ishimoto, L. Kong, X. Shi, B. Liu, Z. Tian, T. Yamada, F. Kong, and J. Abe, “A soybean quantitative trait locus that promotes flowering under long days is identified as FT5a , a FLOWERING LOCUS T ortholog,” *Journal of Experimental Botany,* vol. 67, no. 17, pp. 5247–5258, 2016.

[22] S. Chu, J. Wang, Y. Zhu, S. Liu, X. Zhou, H. Zhang, C.-E. Wang, W. Yang, Z. Tian, H. Cheng, and D. Yu, “An R2R3-type MYB transcription factor , GmMYB29 , regulates isoflavone biosynthesis in soybean,” *PLOS Genetics,* vol. 13, no. 5, 2017.

[23] X. Li, L. Huang, J. Lu, Y. Cheng, Q. You, L. Wang, X. Song, X. Zhou, and Y. Jiao, "Large-Scale Investigation of Soybean Gene Functions by Overexpressing a Full-Length Soybean cDNA Library in Arabidopsis," *Frontiers in Plant Science,* vol. 9, 2018.

[24] B.A. Mcblain, and R.L. Bernard, "A new gene affecting the time of flowering and maturity in soybeans," *Journal of Heredity*, vol. 78, no. 3, pp. 160–162, 1987.

[25] E. R. Bonato and N. A. Vello, "E6, a dominant gene conditioning early flowering and maturity in soybeans," *Genetics and Molecular Biology*, vol. 22, no. 2, pp. 229–232, 1999.

[26] E. R. Cober and H. D. Voldeng, "A new soybean maturity and photoperiod-sensitivity locus linked to E1 and T," *Crop Science.*, vol. 41, no. 3, pp. 698–701, 2001.

[27] B. Liu, A. Kanazawa, H. Matsumura, R. Takahashi, K. Harada, and J. Abe, "Genetic Redundancy in Soybean Photoresponses Associated With Duplication of the Phytochrome A Gene," *Genetics*, vol. 180, no. 2, pp. 995–1007, 2008.

[28] S. Watanabe, R. Hideshima, Z. Xia, Y. Tsubokura, S. Sato, Y. Nakamoto, N. Yamanaka, R. Takahashi, M. Ishimoto, T. Anai, S. Tabata, and K. Harada, "Map-Based Cloning of the Gene Associated With the Soybean Maturity Locus E3," *Genetics*, vol. 182, no. 4, pp. 1251–1262, 2009.

[29] Z. Xia, S. Watanabe, T. Yamada, Y. Tsubokura, H. Nakashima, H. Zhai, T. Anai, S. Sato, T. Yamazaki, S. Lu, H. Wu, S. Tabata, and K. Harada, "Positional cloning and characterization reveal the molecular basis for soybean maturity locus E1 that regulates photoperiodic flowering," *Proceedings of the National Academy of Sciences*, vol. 109, no. 32, 2012.

[30] B. Samanfar, S. J. Molnar, M. Charette, A. Schoenrock, F. Dehne, A. Golshani, F. Belzile, and E. R. Cober, "Mapping and identification of a potential candidate gene for a novel maturity locus, E10, in soybean," *Theoretical and Applied Genetics*, vol. 130, no. 2, pp. 377–390, 2016.

[31] C. Zhao, R. Takeshima, J. Zhu, M. Xu, M. Sato, S. Watanabe, A. Kanazawa, B. Liu, F. Kong, T. Yamada, and J. Abe, "A recessive allele for delayed flowering at the soybean maturity locus E9 is a leaky allele of FT2a, a FLOWERING LOCUS T ortholog," *BMC Plant Biology*, vol. 16, no. 1, 2016.

[32] F. Wang, H. Nan, L. Chen, C. Fang, H. Zhang, T. Su, S. Li, Q. Cheng, L. Dong, B. Liu, F. Kong, and S. Lu, "A new dominant locus, E11, controls early flowering time and maturity in soybean," *Molecular Breeding*, vol. 39, no. 5, 2019.

[33] A. Dissanayaka, T. O. Rodriguez, S. Di, F. Yan, S. M. Githiri, F. R. Rodas, J. Abe, and R. Takahashi, "Quantitative trait locus mapping of soybean maturity gene *E5*," *Breeding Science*, vol. 66, no. 3, pp. 407–415, 2016.

[34] J. Miladinović, M. Ćeran, V. Đorđević, S. Balešević-Tubić, K. Petrović, V. Đukić, and D. Miladinović, "Allelic Variation and Distribution of the Major Maturity Genes in Different Soybean Collections," *Frontiers in Plant Science*, vol. 9, 2018.

[35] D. Cao, R. Takeshima, C. Zhao, B. Liu, A. Jun, and F. Kong, "Molecular mechanisms of flowering under long days and stem growth habit in soybean," *Journal of Experimental Botany*, 2016.

[36] J. L. Weller and R. Ortega, "Genetic control of flowering time in legumes," *Frontiers in Plant Science*, vol. 6, 2015.

[37] R. L. Bernard, "Two Major Genes for Time of Flowering and Maturity in Soybeans 1," *Crop Science*, vol. 11, no. 2, pp. 242–244, 1971.

[38] T. Langewisch, J. Lenis, G.-L. Jiang, D. Wang, V. Pantalone, and K. Bilyeu, "The development and use of a molecular model for soybean maturity groups," *BMC Plant Biology*, vol. 17, no. 1, 2017.

[39] R. I. Buzzell, "Inheritance Of A Soybean Flowering Response To Fluorescent-Daylength Conditions," *Canadian Journal of Genetics and Cytology*, vol. 13, no. 4, pp. 703–707, 1971.

[40] S. Watanabe, Z. Xia, R. Hideshima, Y. Tsubokura, S. Sato, N. Yamanaka, R. Takahashi, T. Anai, S. Tabata, K. Kitamura, and K. Harada, "A Map-Based Cloning Strategy Employing a Residual Heterozygous Line Reveals that theGIGANTEAGene Is Involved in Soybean Maturity and Flowering," *Genetics*, vol. 188, no. 2, pp. 395–407, 2011.

[41] L. Liu, W. Song, L. Wang, X. Sun, Y. Qi, T. Wu, S. Sun, B. Jiang, C. Wu, W. Hou, Z. Ni, and T. Han, "Allele combinations of maturity genes E1-E4 affect adaptation of soybean to diverse geographic regions and farming systems in China," *Plos One*, vol. 15, no. 7, 2020.

[42] H. D. Buzzell, R I Voldeng, "Research Notes : Inheritance of insensitivity to long daylength," *Soybean Genetics. Newsletter*, vol. 7, no. 7, 1980.

[43] S. Lu, X. Zhao, Y. Hu, S. Liu, H. Nan, X. Li, C. Fang, D. Cao, X. Shi, L. Kong, T. Su, F. Zhang, S. Li, Z. Wang, X. Yuan, E. R. Cober, J. L. Weller, B. Liu, X. Hou, Z. Tian, and F. Kong, "Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield," *Nature Genetics*, vol. 49, no. 5, pp. 773–779, 2017.

[44] X. Li, C. Fang, M. Xu, F. Zhang, S. Lu, H. Nan, T. Su, S. Li, X. Zhao, L. Kong, X. Yuan, B. Liu, J. Abe, E. R. Cober, and F. Kong, "Quantitative Trait Locus Mapping of Soybean Maturity Gene E6," *Crop Science*, vol. 57, no. 5, pp. 2547–2554, 2017.

[45] J. D. Ray, K. Hinson, J. E. B. Mankono, and M. F. Malo, "Genetic Control of a Long-Juvenile Trait in Soybean," *Crop Science*, vol. 35, no. 4, pp. 1001–1006, 1995.

[46] S. J. Molnar, S. Rai, M. Charette, and E. R. Cober, “Simple sequence repeat (SSR) markers linked to E1, E3, E4, and E7 maturity genes in soybean,” *Genome*, vol. 46, no. 6, pp. 1024–1036, 2003.

[47] L. Kong, S. Lu, Y. Wang, C. Fang, F. Wang, H. Nan, T. Su, S. Li, F. Zhang, X. Li, X. Zhao, X. Yuan, B. Liu, and F. Kong, “Quantitative Trait Locus Mapping of Flowering Time and Maturity in Soybean Using Next-Generation Sequencing-Based Analysis,” *Frontiers in Plant Science*, vol. 9, 2018.

[48] E. R. Cober, S. J. Molnar, M. Charette, and H. D. Voldeng, “A New Locus for Early Maturity in Soybean,” *Crop Science*, vol. 50, no. 2, pp. 524–527, 2010.

[49] J. Wang, L. Kong, K. Yu, F. Zhang, X. Shi, Y. Wang, H. Nan, X. Zhao, S. Lu, D. Cao, X. Li, C. Fang, F. Wang, T. Su, S. Li, X. Yuan, B. Liu, and F. Kong, “Development and validation of InDel markers for identification of QTL underlying flowering time in soybean,” *The Crop Journal*, vol. 6, no. 2, pp. 126–135, 2018.

[50] F. Kong, H. Nan, D. Cao, Y. Li, F. Wu, J. Wang, S. Lu, X. Yuan, E. R. Cober, J. Abe, and B. Liu, “A New Dominant Gene E9 Conditions Early Flowering and Maturity in Soybean,” *Crop Science*, vol. 54, no. 6, pp. 2529–2535, 2014.

[51] Y.-C. Lee, R. L. Hamawaki, V. Colantonio, M. J. Iqbal, and D. A. Lightfoot, “The use of marker-assisted selection in developing improved varieties of soybean,” *Achieving sustainable cultivation of grain legumes Volume 2 Burleigh Dodds Series in Agricultural Science*, pp. 83–104, 2018.

[52] S. Teama, “DNA Polymorphisms: DNA-Based Molecular Markers and Their Application in Medicine,” *Genetic Diversity and Disease Susceptibility*, 2018.

[53] B. C. Y. Collard, M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang, “An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts,” *Euphytica*, vol. 142, no. 1-2, pp. 169–196, 2005.

[54] G.-L. Jiang, “Molecular Markers and Marker-Assisted Breeding in Plants,” *Plant Breeding from Laboratories to Fields*, 2013.

[55] A. K. Singh, “Discovery and Role of Molecular Markers Involved in Gene Mapping, Molecular Breeding, and Genetic Diversity,” *Plant Bioinformatics*, pp. 303–328, 2017.

[56] M. A. Nadeem, M. A. Nawaz, M. Q. Shahid, Y. Doğan, G. Comertpay, M. Yıldız, R. Hatipoğlu, F. Ahmad, A. Alsaleh, N. Labhane, H. Özkan, G. Chung, and F. S. Baloch, “DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing,” *Biotechnology & Biotechnological Equipment*, vol. 32, no. 2, pp. 261–285, 2017.

[57] B. P. Sheth and V. S. Thaker, “Plant systems biology: insights, advances and

challenges," *Planta*, vol. 240, no. 1, pp. 33–54, 2014.

[58] E. M. Bunnik and K. G. L. Roch, "An Introduction to Functional Genomics and Systems Biology," *Advances in Wound Care*, vol. 2, no. 9, pp. 490–498, 2013.

[59] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne, and A. Golshani, "Computational Methods For Predicting Protein–Protein Interactions," *Protein – Protein Interaction Advances in Biochemical Engineering/Biotechnology*, pp. 247–267, 2008.

[60] S. Pitre, F. Dehne, A. Chan, J. Cheetham, A. Duong, A. Emili, M. Gebbia, J. Greenblatt, M. Jessulat, N. Krogan, X. Luo, and A. Golshani, *BMC Bioinformatics*, vol. 7, no. 1, p. 365, 2006.

[61] C. V. Mering, "STRING: known and predicted protein-protein associations, integrated and transferred across organisms," *Nucleic Acids Research*, vol. 33, no. Database issue, 2004.

[62] A. Gioutlakis, M. I. Klapa, and N. K. Moschonas, "PICKLE 2.0: A human protein-protein interaction meta-database employing data integration via genetic information ontology," *Plos One*, vol. 12, no. 10, 2017.

[63] N. Jones, H. Ougham, H. Thomas, and I. Pašakinskienė, "Markers and mapping revisited: finding your gene," *New Phytologist*, vol. 183, no. 4, pp. 935–966, 2009.

[64] K. Chandra and A. Pandey, "QTL Mapping in Crop Improvement: A Basic Concept," *International Journal of Current Microbiology and Applied Sciences*, vol. 6, no. 12, pp. 835–842, 2017.

[65] H. Verdeprado, T. Kretzschmar, H. Begum, C. Raghavan, P. Joyce, P. Lakshmanan, J. N. Cobb, and B. C. Collard, "Association mapping in rice: basic concepts and perspectives for molecular breeding," *Plant Production Science*, vol. 21, no. 3, pp. 159–176, 2018.

[66] A. E. Lipka, C. B. Kandianis, M. E. Hudson, J. Yu, J. Drnevich, P. J. Bradbury, and M. A. Gore, "From association to prediction: statistical methods for the dissection and selection of complex traits in plants," *Current Opinion in Plant Biology*, vol. 24, pp. 110–118, 2015.

[67] S. Challa and N. R. Neelapu, "Genome-Wide Association Studies (GWAS) for Abiotic Stress Tolerance in Plants," *Biochemical, Physiological and Molecular Avenues for Combating Abiotic Stress Tolerance in Plants*, pp. 135–150, 2018.

[68] K. Dick, B. Samanfar, B. Barnes, E. R. Cober, B. Mimee, L. H. Tan, S. J. Molnar, K. K. Biggar, A. Golshani, F. Dehne, and J. R. Green, "PIPE4: Fast PPI Predictor for Comprehensive Inter- and Cross-Species Interactomes," *Scientific Reports*, vol. 10, no. 1, 2020.

[69] D. Torkamaneh, J. Laroche, B. Valliyodan, L. O'Donoughue, E. Cober, I. Rajcan, R. V.

Abdelnoor, A. Sreedasyam, J. Schmutz, H. T. Nguyen, and F. Belzile, “Soybean Haplotype Map (GmHapMap): A Universal Resource for Soybean Translational and Functional Genomics,” 2019.

[70] D. Torkamaneh, J. Laroche, A. Tardivel, L. Odonoughue, E. Cober, I. Rajcan, and F. Belzile, “Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean,” *Plant Biotechnology Journal*, vol. 16, no. 3, pp. 749–759, 2017.

[71] A. J. Severin, J. L. Woody, Y.-T. Bolon, B. Joseph, B. W. Diers, A. D. Farmer, G. J. Muehlbauer, R. T. Nelson, D. Grant, J. E. Specht, M. A. Graham, S. B. Cannon, G. D. May, C. P. Vance, and R. C. Shoemaker, “RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome,” *BMC Plant Biology*, vol. 10, no. 1, p. 160, 2010.

[72] A. Untergasser, I. Cutcutache, T. Koressaar, J. Ye, B. C. Faircloth, M. Remm, and S. G. Rozen, “Primer3—new capabilities and interfaces,” *Nucleic Acids Research*, vol. 40, no. 15, 2012.

[73] P. Kerpedjiev, S. Hammer, and I. L. Hofacker, “Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams,” *Bioinformatics*, vol. 31, no. 20, pp. 3377–3379, 2015.

[74] Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, “Predicting the Functional Effect of Amino Acid Substitutions and Indels,” *PLoS ONE*, vol. 7, no. 10, 2012.

[75] Q. Wan, S. Chen, Z. Shan, Z. Yang, L. Chen, C. Zhang, S. Yuan, Q. Hao, X. Zhang, D. Qiu, H. Chen, and X. Zhou, “Stability evaluation of reference genes for gene expression analysis by RT-qPCR in soybean under different conditions,” *Plos One*, vol. 12, no. 12, 2017.

[76] M. Xu, Z. Xu, B. Liu, F. Kong, Y. Tsubokura, S. Watanabe, Z. Xia, K. Harada, A. Kanazawa, T. Yamada, and J. Abe, “Genetic variation in four maturity genes affects photoperiod insensitivity and PHYA-regulated post-flowering responses of soybean,” *BMC Plant Biology*, vol. 13, no. 1, p. 91, 2013.

[77] M. W. Pfaffl, “A new mathematical model for relative quantification in real-time RT-PCR,” *Nucleic Acids Research*, vol. 29, no. 9, 2001.

[78] H. R. Shehata, J. Li, S. Chen, H. Redda, S. Cheng, N. Tabujara, H. Li, K. Warriner, and R. Hanner, “Droplet digital polymerase chain reaction (ddPCR) assays integrated with an internal control for quantification of bovine, porcine, chicken and turkey species in food and feed,” *Plos One*, vol. 12, no. 8, 2017.

[79] S. C. Taylor, G. Laperriere, and H. Germain, “Droplet Digital PCR versus qPCR for gene expression analysis with low abundant targets: from variable nonsense to publication quality data,” *Scientific Reports*, vol. 7, no. 1, 2017.

[80] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research*, vol. 47, no. D1, 2018.

[81] L. Ma and G. Li, "FAR1-RELATED SEQUENCE (FRS) and FRS-RELATED FACTOR (FRF) Family Proteins in Arabidopsis Growth and Development," *Frontiers in Plant Science*, vol. 9, 2018.

[82] W. Tang, W. Wang, D. Chen, Q. Ji, Y. Jing, H. Wang, and R. Lin, "Transposase-Derived Proteins FHY3/FAR1 Interact with PHYTOCHROME-INTERACTING FACTOR1 to Regulate Chlorophyll Biosynthesis by Modulating HEMB1 during Deetiolation in Arabidopsis," *The Plant Cell*, vol. 24, no. 5, pp. 1984–2000, 2012.

[83] X. Ouyang, J. Li, G. Li, B. Li, B. Chen, H. Shen, X. Huang, X. Mo, X. Wan, R. Lin, S. Li, H. Wang, and X. W. Deng, "Genome-Wide Binding Site Analysis of FAR-RED ELONGATED HYPOCOTYL3 Reveals Its Novel Function in Arabidopsis Development," *The Plant Cell*, vol. 23, no. 7, pp. 2514–2535, 2011.

[84] X. Yu, H. Liu, J. Klejnot, and C. Lin, "The Cryptochrome Blue Light Receptors," *The Arabidopsis Book*, vol. 8, 2010.

[85] Y. Meng, H. Li, Q. Wang, B. Liu, and C. Lin, "Blue Light-Dependent Interaction between Cryptochrome2 and CIB1 Regulates Transcription and Leaf Senescence in Soybean," *The Plant Cell*, vol. 25, no. 11, pp. 4405–4420, 2013.

[86] Q. Zhang, H. Li, R. Li, R. Hu, C. Fan, F. Chen, Z. Wang, X. Liu, Y. Fu, and C. Lin, "Association of the circadian rhythmic expression of GmCRY1a with a latitudinal cline in photoperiodic flowering of soybean," *Proceedings of the National Academy of Sciences*, vol. 105, no. 52, pp. 21028–21033, 2008.

[87] P.-L. Liu, L. Du, Y. Huang, S.-M. Gao, and M. Yu, "Origin and diversification of leucine-rich repeat receptor-like protein kinase (LRR-RLK) genes in plants," *BMC Evolutionary Biology*, vol. 17, no. 1, 2017.

[88] F. Zhou, Y. Guo, and L.-J. Qiu, "Genome-wide identification and evolutionary analysis of leucine-rich repeat receptor-like protein kinase genes in soybean," *BMC Plant Biology*, vol. 16, no. 1, 2016.

[89] K. J. Lee, D. S. Kim, J.-B. Kim, S.-H. Jo, S.-Y. Kang, H.-I. Choi, and B.-K. Ha, "Identification of candidate genes for an early-maturing soybean mutant by genome resequencing analysis," *Molecular Genetics and Genomics*, vol. 291, no. 4, pp. 1561–1571, 2016.

[90] J.-E. Choe, B. Kim, E. K. Yoon, S. Jang, G. Kim, S. Dhar, S. A. Lee, and J. Lim,

"Characterization of the GRAS transcription factor SCARECROW-LIKE 28's role in Arabidopsis root growth," *Journal of Plant Biology*, vol. 60, no. 5, pp. 462–471, 2017.

[91] S. W. Burgdorf, C. L. Clark, J. R. Burgdorf, and D. H. Spector, "Mutation of Glutamine to Arginine at Position 548 of IE2 86 in Human Cytomegalovirus Leads to Decreased Expression of IE2 40, IE2 60, UL83, and UL84 and Increased Transcription of US8-9 and US29-32," *Journal of Virology*, vol. 85, no. 21, pp. 11098–11110, 2011.

[92] P.-C. Li, S.-W. Yu, K. Li, J.-G. Huang, X.-J. Wang, and C.-C. Zheng, "The Mutation of Glu at Amino Acid 3838 of AtMDN1 Provokes Pleiotropic Developmental Phenotypes in Arabidopsis," *Scientific Reports*, vol. 6, no. 1, 2016.

[93] R. W. Bruce, D. Torkamaneh, C. M. Grainger, F. Belzile, M. Eskandari, and I. Rajcan, "Haplotype diversity underlying quantitative traits in Canadian soybean breeding germplasm," *Theoretical and Applied Genetics*, vol. 133, no. 6, pp. 1967–1976, 2020.

[94] Y. Liu, H. Wei, M. Ma, Q. Li, D. Kong, J. Sun, X. Ma, B. Wang, C. Chen, Y. Xie, and H. Wang, "Arabidopsis FHY3 and FAR1 Regulate the Balance between Growth and Defense Responses under Shade Conditions," *The Plant Cell*, vol. 31, no. 9, pp. 2089–2106, 2019.

[95] D. Ma, X. Li, Y. Guo, J. Chu, S. Fang, C. Yan, J. P. Noel, and H. Liu, "Cryptochrome 1 interacts with PIF4 to regulate high temperature-mediated hypocotyl elongation in response to blue light," *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 224–229, 2015.

Appendix

Appendix 1: The following figures indicate the location of amino acid sequence variation among the wild type and mutant genotypes for each candidate gene. Pairwise alignment and highlights done using ExPASy and BoxShade server (https://embnet.vital-it.ch/software/BOX_form.html).

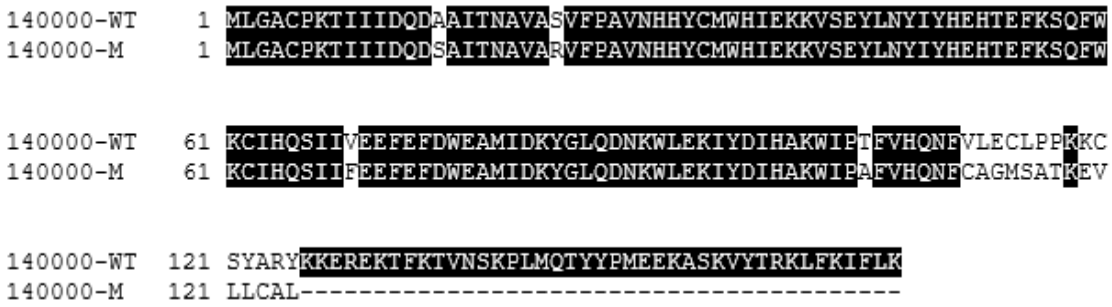
Glyma.04G111200

Glyma.04G101500

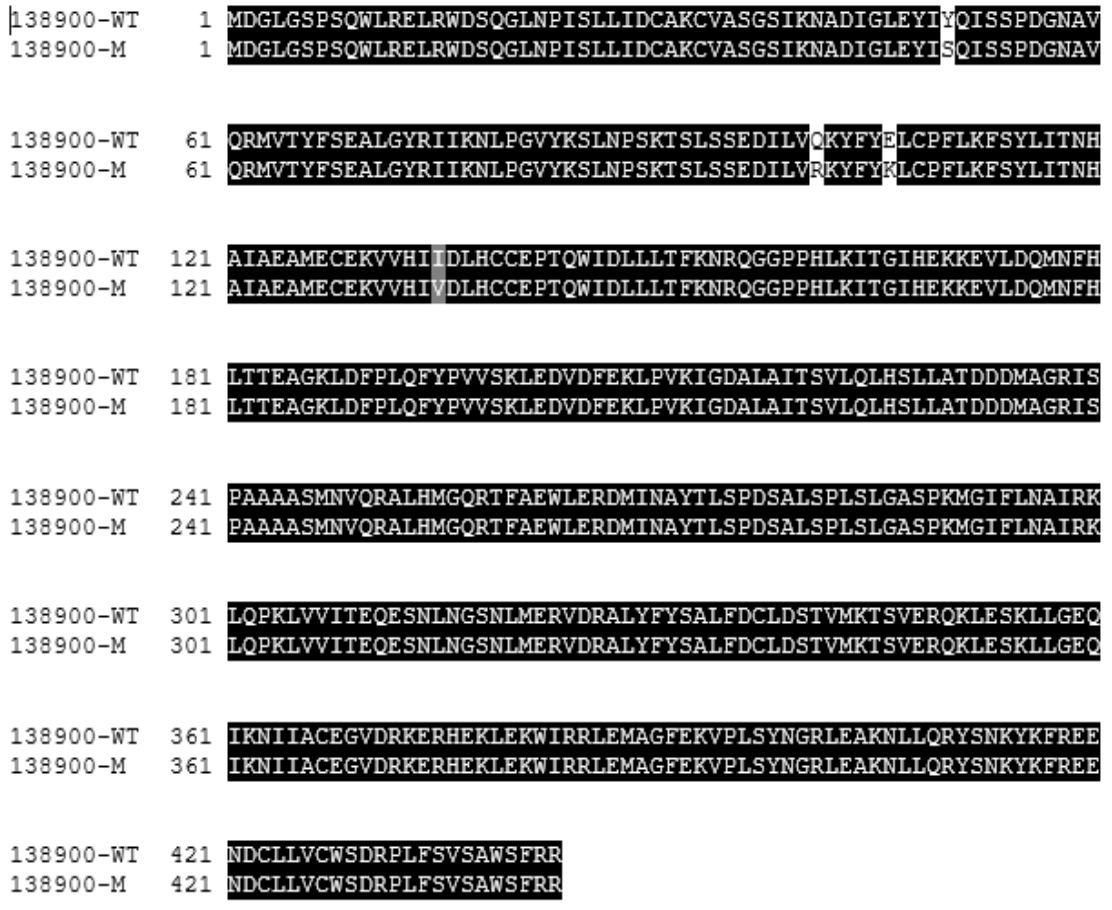
Glyma.04G124300

```
124300-WT     1 MDIDLRLPSGEHDKEDEETTTIDNMLDSEEKLHNGGIDGRNIVDTGIEVHALNGGDLNSP
124300-M      1 MDIDLRLPSGEHDKEDEETTTIDNMLDSEEKLHNGGIDGRNIVDTGIEVHALNGGDLNSP

124300-WT    61 IVDIVMFKEDTNLEPLSGMEFESHGEAYSFYQEYARSMGFNTAIQNSRRSKTSREFIDAK
124300-M     61 IVDIVMFKEDTNLEPLSGMEFESHGEAYSFYQEYARSMGFNTAIQNSRRSKTSREFIDAK

124300-WT   121 FACSRYGTKREYDKSFNRPRARQNKQDSENSTGRRSCSKTDCKASMHVKRRSDGKWVIHS
124300-M    121 FACSRYGTKREYDKSFNRPRARQNKQDSENSTGRRSCSKTDCKASMHVKRRSDGKWVIHS

124300-WT   181 FVKEHNHELLPAQAVSEQTRRMYAAMARQFAEYKTVVGLKNEKNPFDKGRNLGLESGEAR
124300-M    181 FVKEHNHELLPAQAVSEQTRRMYAAMARQFAEYKTVVGLKNEKNPFDKGRNLGLESGEAR

124300-WT   241 LMLDFFIQMQNMNSNFFYAVDLGEDQRLKNLLWIDAKSRNDYINFCDVVSFDTAYVRNKY
124300-M    241 LMLDFFIQMQNMNSNFFYAVDLGEDQRLKNLLWIDAKSRNDYINFCDVVSFDTAYVRNKY

124300-WT   301 KMPLALFVGVNQHYQFTLLGCALISDESAATFSWLFRTWLKGVGGQVPKVIITDHDKTLK
124300-M    301 KMPLALFVGVNQHYQFTLLGCALISDESAATFSWLFRTWLKGVGGQIPKVIITDHDKTLK

124300-WT   361 SVISDIFPNSSHCVCLWHILGKVSENLSPVIKKHENFMAKFEKCIYRSLTSDDFEKRWWK
124300-M    361 SVISDIFPNSSHCVCLWHILGKVSENLSPVIKKHENFMAKFEKCIYRSLTSDDFEKRWWK

124300-WT   421 IVDKFELREDECMQSLYEDRKLWAPTFMKDVFLGGMSTVQRSESVNSFFDKYVHKKTSVQ
124300-M    421 IVDKFELREDECMQSLYEDRKLWAPTFMKDVFLGGMSTVQRSESVNSFFDKYVHKKTSVQ

124300-WT   481 DFVKQYEAILQDRYEEEAKADSDTWNKVATLKTPSPLEKSVAGIFSHAVFKKIQTEVVGA
124300-M    481 DFVKQYEAILQDRYEEEAKADSDTWNKVATLKTPSPLEKSVAGIFSHAVFKKIQTEVVGA

124300-WT   541 VACHPKADRQDDTTIVHRVHDMETNKDFFVVVNQVKSELSCICRLFEYRGYLCRHALFVL
124300-M    541 VACHPKADRQDDTTIVHRVHDMETNKDFFVVVNQVKSELSCICRLFEYRGYLCRHALFVL

124300-WT   601 QYSGQSVFPSQYILKRWTKDAKVRNIMGEESEHMLTRVQRYNDLCQRALKLSEEGSLSQE
124300-M    601 QYSGQSVFPSQYILKRWTKDAKVRNIMGEESEHMLTRVQRYNDLCQRALKLSEEGSLSQE

124300-WT   661 SYGIAFHALHEAHKSCVSVNNSSKSSPTEAGTPGAHGQLSTEEDTQSRNMGKSNKKKHPT
124300-M    661 SYGIAFHALHEAHKSCVSVNNSSKSSPTEAGTPGAHGQLSTEEDTQSRNMGKSNKKKHPT

124300-WT   721 KKKKVNSEAEVITVGALDNLQQMDKFSTRAVTLEGYYGTQQSVQGMLNLMGPTRDDYYGN
124300-M    721 KKKKVNSEAEVITVGALDNLQQMDKFSTRAVTLEGYYGTQQSVQGMLNLMGPTRDDYYGN

124300-WT   781 QQTLQGLGPISSIPTSHDGYYGTHQGMPGLAQLDFLRTGFTYGIRDDPNVRATQLHEDPS
124300-M    781 QQTLQGLGPISSIPTSHDGYYGTHQGMPGLAQLDFLRTGFTYGIRDDPNVRATQLHEDPS

124300-WT   841 RHA
124300-M    841 RHA
```

Glyma.04G140000



Glyma.04G138900



Glyma.04G126000

```
126000-WT     1 MTTEAETERVVVIQDASRDVNSNAILGALEWFSVKAGDQLIIVAILDWMSSPMGYMVRVD
126000-M      1 MTTEAETERVVVIQDASRDVNSNAILGALEWFSVKAGDQLIIVAILDWMSSPMGYMVRVD

126000-WT    61 SSSMISTNKKIIEKRLTKKKEEYLMNQNIQEISNYCKLNEIGFQLEVLVGSTAEVASNAA
126000-M     61 SSSMISTNKKIIEKRLTKKKEEYLMNQNIQEISNYCKLNEIGFQLEVLVGSTAEVASNAA

126000-WT   121 KEFQATRLILVRQIHKDMKHFVRNLPCGMYRITSDNSIERLKDPKSAVSTKTFALRQENV
126000-M    121 KEFQATRLILVRQIHKDMKHFVRNLPCGMYRITSDNSIERLKDPKSAVSTKTFALRQENV

126000-WT   181 SYKEMFPGSEEEERSLLMSRSSSSDLLTSTGISSQWSTEVSTSSFGSLRYGCQYQEGKFY
126000-M    181 SYKEMFPGSEEEERSLLMSRSSSSDLLTSTGISSQWSTEVSTSSFGSLRYGCQYQEGKFY

126000-WT   241 SNKEQETTGNQSLFHISENEETSQLQVNKKEQHSRNNETSHMEEEFTNPLCSVCKNRRPN
126000-M    241 SNKEQETTGNQSLFHISENEETSQLQVNKKEQHSRNNETSHMEEEFTNPLCLVCKNRRPN

126000-WT   301 IGLKRDFSYAELHTATQGFSPKNFLSEGGFGSVYKGLLNGMKIAVKQHKYASFQGEKEFK
126000-M    301 IGLKRDFSYAELHTATQGFSPKNFLSEGGFGSVYKGLLNGMKIAVKQHKYASFQGEKEFK

126000-WT   361 SEVNVLSKARHENVVVLLGSCSEKNNRLLVYEYVCNGSLDQHLSEHSRSPLSWEDRINVA
126000-M    361 SEVNVLSKARHENVVVLLGSCSEKNNRLLVYEYVCNGSLDQHLSEHSRSPLSWEDRINVA

126000-WT   421 IGAAKGLLYLHKNNMIHRDVRPNNILITHDYHPLLGDFGLARNQNQDSIHSTEVVGTLGY
126000-M    421 IGAAKGLLYLHKNNMIHRDVRPNNILITHDYHPLLGDFGLARNQNQDSIHSTEVVGTLGY

126000-WT   481 LAPEYAELGKVSTKTDVYSFGVVLLQLITGMRTTDKRLGGRSLVGWARPLLRERNYPDLI
126000-M    481 LAPEYAELGKVSTKTDVYSFGVVLLQLITGMRTTDKRLGGRSLVGWARPLLRERNYPDLI

126000-WT   541 DERIINSHDVHQLFWMVRIAEKCLSREPQRRLNMIQVVDALTDIVEGRTCDIILRDYSPA
126000-M    541 DERIINSHDVHQLFWMVRIAEKCLSREPQRRLNMIQVVDALTDIVEGRTCDIILRDYSPA

126000-WT   601 RSDSTYSASDSDESEDEMQEPLRFESELLSHSSESIESNNISQMMHMIVRQPPSPPIQSI
126000-M    601 RSDSTYSASDSDESEDEMQEPLRFESELLSHSSESIESNNISQMMHMIVRQPPSPPIQSI

126000-WT   661 SSSSSSSYKLHYESTSDGEAHNEGEIEISNSNWGLLNS
126000-M    661 SSSSSSSYKLHYESTSDGEAHNEGEIEISNSNWGLLNS
```

Glyma.04G124600

Appendix 2: RNA sequence data compiled from Severin et al., [71]. The data below shows RNA expression across tissues; young leaf, flower, one cm pod, pod shell 10 days after flowering, pod shell 14 days after flowering, seed 10 days after flowering, seed 14 days after flowering, and seed 21 days after flowering.

Gene	*young_leaf*	*flower*	*one cm pod*	*pod shell 10DAF*	*pod shell 14DAF*	*seed 10DAF*	*seed 14DAF*
Glyma.04g124300	112	78	49	45	21	19	25
Glyma.04G101500	220	199	177	228	123	25	29
Glyma.04g119300	0	0	0	0	0	0	1
Glyma.04g146000	32	42	33	25	22	18	11
Glyma.04g143000	0	0	1	0	0	0	0
Glyma.04g143300	0	0	0	0	0	0	0
Glyma.04g156400	0	0	0	0	0	0	0
Glyma.04g141700	0	0	0	0	0	0	0
Glyma.04g126000	0	0	0	0	0	0	0
Glyma.04g093900	0	0	0	0	0	0	0
Glyma.04g138900	29	10	22	12	7	5	9
Glyma.04G110400	7	60	20	21	44	1	2
Glyma.04g111200	23	28	42	26	10	16	10
Glyma.04g140000	0	0	0	0	0	0	0
Glyma.04g124600	0	0	0	0	0	0	0
Glyma.04G147500	24	32	34	37	13	0	2
Glyma.04G111200	23	28	42	26	10	16	10
Glyma.04G110400	7	60	20	21	44	1	2
Glyma.04G099600	37	20	54	35	46	9	32

Appendix 3: Housekeeping genes and primers used for expression analysis

Gene	*Locus Name*	*Primer Sequence F (5'- 3')*	*Primer Sequence R (3'- 5')*
TUB4	Glyma.03G124400	AGCTGGTCAATGTGGAAACC	AAGCACAGCTCGAGGAACAT
ACT11	Glyma.18G290800	ATCTTGACTGAGCGTGGTTATTCC	GCTGGTCCTGGCTGTCTCC
UKN1	Glyma.12G020500	TGGTGCTGCCGCTATTTACTG	GGTGGAAGGAACTGCTAACAAT
UBQ10	Glyma.07G199900	TCCCACCAGACCAGCAGAG	CCTTGTGTTGCGTCTTCGTG
ELF1a	Glyma.19G052400	GACCTTCTTCGTTTCTCGCA	CGAACCTCTCAATCACACGC
CYP	Glyma.12G024700	CGGGACCAGTGTGCTTCTTCA	CCCCTCCACTACAAAGGCTCG
TAU5	Glyma.05G157300	AGGTCGGAAACTCCTGCTGG	AAGGTGTTGAAGGCGTCGTG
N/A	Glyma.04G124300	TCTTGGCTATTTCGGACGTG	TCAATGTCTTGTCATGGTCAGT
N/A	Glyma.04G111200	TGAATATGCCCGGCGATTAG	CCTCGGATGCTGACACAATAA
N/A	Glyma.04G124600	GGATGCAGACAAGTTGGAGTAG	CGGCAAGACCGCTTATTGT
N/A	Glyma.04G101500	CTTTAGTGGATGCTGGGATGAG	CTGCAGAACCTTCACAAAGAAAC
N/A	Glyma.04G126000	TCAGTTATGCTGAGCTCCATAC	CATTCCATTCAGCAGTCCTTTG

Appendix 4: Primers used for sequencing

Primer Name	*Sequence (5' – 3')*
111200-F1	GGGCTTTCAGGACCAATTTT
111200-F2	TGTGGCATTACACGTGACAG
111200-SF1	AAAAGCGAGTTGCACAGGTT
111200-SR1	ATCAAGCACTGGAGGGACAG
111200-SF2	CTGTCCCTCCAGTGCTTGAT
111200-SR2	CCCCAGAAACTTCCATCTCA
111200-SF3	TGAGATGGAAGTTTCTGGGG
111200-SR3	TTTTGGGCTGAAGCAGTTCT
111200-SF4	AGAACTGCTTCAGCCCAAAA
111200-SR4	AATCGGGAGATCGTAACACG
111200-SF5	CGTGTTACGATCTCCCGATT
111200-SR5	TCCCAATGTAGAACCCTTGC
111200-SF6	GCAAGGGTTCTACATTGGGA
111200-SR6	ACGCAAATTTACCTCGGATG
111200-SF7	CATCCGAGGTAAATTTGCGT
111200-SR7	TCCTCATCTCCGCTACACCT
111200-SF8	AGGTGTAGCGGAGATGAGGA
111200-SR8	AACCAGAAGGTGCAGGAAAA
111200-SF9	TTTTCCTGCACCTTCTGGTT
111200-SR9	CACCGCCTAGAGGAGTTCAG
111200-SF10	CTGAACTCCTCTAGGCGGTG
111200-SR10	CCAAGAAGCATCAGCGGTAT
111200-SF11	ATACCGCTGATGCTTCTTGG
111200-R1	TCAGCGGTCTCTTGAATCAT
111200-R2	TCAACGCATTTGTGGTTGAT
124300-F1	ACATAAGCCCATTCCGTGAG
124300-F2	GCGGGAAATGACAGGTAGAG
124300-SF1	AAAAGGGGGCATTGGTAAGT
124300-SR1	AACCTGCCCCTTCATCTCTT
124300-SF2	AAGAGATGAAGGGGCAGGTT
124300-SR2	GGGTTTGGGCTATTGGCTAT
124300-SF3	ATAGCCAATAGCCCAAACCC
124300-SR3	AGCAGCAAAGAAAGGCCATA
124300-SF4	TATGGCCTTTCTTTGCTGCT
124300-SR4	AGCATGAGGTGAGGCTAGGA
124300-SF5	TCCTAGCCTCACCTCATGCT
124300-SR5	AATTTGCGGATGAGAAATGC
124300-SF6	GCATTTCTCATCCGCAAATT
124300-SR6	TCCCCGTGAGACTCAAATTC
124300-SF7	GAATTTGAGTCTCACGGGGA
124300-SR7	TGATATCAGGGCACATCCAA

124300-SF8	TTGGATGTGCCCTGATATCA
124300-SR8	AGCTTTAGGATGGCAAGCAA
124300-SF9	TTGCTTGCCATCCTAAAGCT
124300-SR9	GCAGGTCAAAGAGGAGGTCA
124300-SF1	TGACCTCCTCTTTGACCTGC
124300-SR1	TGTGAAACCAGTTCGCAAAA
124300-SF1	TTTTGCGAACTGGTTTCACA
124300-SR1	CCAAAAGAGGTTGAACGCTAA
124300-SF1	TTAGCGTTCAACCTCTTTTGG
124300-SR1	AATGCCAAGTGAAGGCAATC
124300-R1	CGGTATGGTGGTTTGTAGGG
124300-R2	GGGGGAAGGTTAGAATTGGA
124600-F1	CCCAGAACTCCAGACCTCAA
124600-F2	TCATCTGCTTCAAAGCCTCA
124600-SF1	AGCCAATGACTGGGTCGTTA
124600-SR1	GTAGCCGTCCTGGTCCATAA
124600-SF2	TTATGGACCAGGACGGCTAC
124600-SR2	TGCAATCAATTGGTCTTCCA
124600-SF3	TGGAAGACCAATTGATTGCA
124600-SR3	GCCCATTTGAGGGGTTATTT
124600-SF4	AAATAACCCCTCAAATGGGC
124600-SR4	CCTCCACACATGTACGCAAC
124600-SF5	GTTGCGTACATGTGTGGAGG
124600-SR5	TGAAGTTGTCAGCAGGTTGG
124600-SF6	CCAACCTGCTGACAACTTCA
124600-SR6	TTTGTGTGCCAGGTTCGTAG
124600-R1	GGGGACGAGGTGGAGATTAT
124600-R2	GTTGGGGTGTAGGATTTGGA
101500-F1	TGAAATGGAGAACGCATGAA
101500-F2	GGAGAGACGTGGAAGATGGA
101500-SR1	AGGTGCCCATACGAAAACTG
101500-SF2	CAGTTTTCGTATGGGCACCT
101500-SR2	GCAGTCAGAACCTCCTTTGC
101500-SF3	GCAAAGGAGGTTCTGACTGC
101500-SR3	AAGTGCATTGCTTGCCTTCT
101500-SF4	AGAAGGCAAGCAATGCACTT
101500-SR4	CATGCACATGCAATCAAATG
101500-SF5	CATTTGATTGCATGTGCATG
101500-SR5	CAGCTCTTGAAGCTGCCTCT
101500-SF6	AGAGGCAGCTTCAAGAGCTG
101500-SR6	CCTCCAATTCAGCATTTGGT
101500-SF7	ACCAAATGCTGAATTGGAGG
101500-SR8	CCTATGGTGGCTCTTTCTCC
101500-SF9	GGAGAAAGAGCCACCATAGG

101500-R1	TGCTCCGAAAGGCTTAATGT
101500-R2	TGGTATGCTTTAGGAAGGGAAG
126000-F1	CATGTTTGGGTGTCTGTTGG
126000-F2	TTTTATGCTGCACGAGGATG
126000-SR1	TGACATAGTGGAGGGCAGAA
126000-SF2	TTCTGCCCTCCACTATGTCA
126000-SR2	CTCAGCTCATCAAGGCATCA
126000-SF3	TGATGCCTTGATGAGCTGAG
126000-SR3	TCTGCAATGGCTCACTTGAC
126000-SF4	GTCAAGTGAGCCATTGCAGA
126000-SR4	GAAGCTTGCGGTATGGTTGT
126000-SF5	ACAACCATACCGCAAGCTTC
126000-SR5	CCCGGTGTTCTATCATAGCC
126000-SF6	GGCTATGATAGAACACCGGG
126000-SR6	TGCAGGAATACTGTGGGTGA
126000-SF7	TCACCCACAGTATTCCTGCA
126000-SR7	GCTTCTTGGCCATCATTTGT
126000-SF8	ACAAATGATGGCCAAGAAGC
126000-SR8	GGCTCTACTGCAGAGGTTGC
126000-SF9	GCAACCTCTGCAGTAGAGCC
126000-SR9	AGACTGAGAGGGTGGTGGTG
126000-F10	CACCACCACCCTCTCAGTCT
126000-R1	TAGGCAATGTACAGCGCAAC
126000-R2	TCAAACCCTTATACCCACCAA
138900-F1	TGCCTGTTCATTTGTGCTTT
138900-F2	TTTCACACCTACGGCTCTCA
138900-SR1	TGAGAACTTCAAAAATGGACACA
138900-SF2	TGTGTCCATTTTTGAAGTTCTCA
138900-SR2	TCCCTGCCATATCATCATCA
138900-SF3	TGATGATGATATGGCAGGGA
138900-SR3	CTCCAAGAAGCTTGCTCTCAA
138900-SF4	TTGAGAGCAAGCTTCTTGGAG
138900-R1	GGCAATGCAAGGGATTCTT
138900-R2	TGCCACCTCTTTCAGCTGTT
140000-F1	CATTGCCACCGGATAGTCTT
140000-SF1	TGCAACAAACAACCGAGAAC
140000-SR1	TTGTTCCATTCACCGGAGTT
140000-R1	TCTTGACAAGTGAGGGGAAAA
093900-F1	ATGACGAGGGCGAAGATAAA
093900-F2	CTCATACCCTTGCATGCTCA
093900-R1	TCTCCGTCCTTCTCCTCTCA
093900-R2	TTTCTGTTCCAAACCCATCA
110400-F1	TTGGCACAATTTTTCGGTTT
110400-F2	TGGGTGCAGCTTTCCTAAGT

110400-SR1	GTTTTGGGTCGGTTTGAGAA
110400-SF2	TTCTCAAACCGACCCAAAAC
110400-SR2	GATTTGATGCCATCCTTGGT
110400-SF3	ACCAAGGATGGCATCAAATC
110400-SR3	TTCTAGTGCTGATGCGGTTG
110400-SF4	CAACCGCATCAGCACTAGAA
110400-SR4	TTGGATCCGTCCCTCCTTAT
110400-SF5	ATAAGGAGGGACGGATCCAA
110400-R1	GGTAGCAACTGCAAGCATGA
110400-R2	GTAGATGTGCCGGAAGTGGT
99600-F1	ATGTGACGAGTGGCAATTCA
99600-F2	ATGGGTTTCCCTTCTTTTGC
99600-F3	CACATCCAATTTGAGCGTGA
99600-SR1	GGGTGGGCGGTTAGTAATTT
99600-SF2	AAATTACTAACCGCCCACCC
99600-SR2	CAGCAGAGAGCATGAACCAA
99600-SF3	TTGGTTCATGCTCTCTGCTG
99600-R1	AACGGCCAAACCAAATACAC
99600-R2	TGGGATGTAGCCACTAGCAA
99600-R3	GCATGTGTTGGTGTAGGTGTG
144700-F1	CAGCCACACGAAGCATTAAA
144700-SF1	ATTGGACATATGCAGCCACA
144700-SR1	CAAGCCATGGTACATTGGAA
144700-R1	CATGCAATGAATCCCCCTTA
150800-F1	ACGGTGTTGATGGGAGAAAG
150800-SF1	TTACGGTGTTTCAAAGGCATC
150800-SR1	CATTGCTTCACAACCACACC
150800-SF2	GGTGTGGTTGTGAAGCAATG
150800-SR2	GCTTGAGGACTGGACGTAGG
150800-SF3	CCTACGTCCAGTCCTCAAGC
150800-SR3	AGATCAAGCCTTGCCTCAAA
150800-SF4	TTTGAGGCAAGGCTTGATCT
150800-SR4	TCCTCTAAGTCCTTCAAGTGGAT
150800-R1	AAGAGCTTCCTTGGCTTTCC
147500-F1	GGAAAGAGTTCGGTGACTGC
147500-F2	CAGGAGCTAAAACACCAGCA
147500-SR1	TGGCAGTGCAAAAGACAAAA
147500-SF2	TTTTGTCTTTTGCACTGCCA
147500-SR2	AAGGTGGTGGCAATAAGGTG
147500-SF3	CACCTTATTGCCACCACCTT
147500-SR3	CTAGCCACACCCTCCTCTTG
147500-SF4	CAAGAGGAGGGTGTGGCTAG
147500-SR4	ATTTGCAGCGCAATTCTTTC
147500-SF5	GAAAGAATTGCGCTGCAAAT

147500-SR5	TCCTTTCCATGATTCCATCC
147500-SF6	GGATGGAATCATGGAAAGGA
147500-R1	TCAAATCTCCTCCACCACTTG
147500-R2	ATCTGATTGGGCCAACCTAC