A Functional Genomics Approach to Identifying Potential Candidates Underlying the E7 Maturity Locus in Soybean (*Glycine max*)

by

Arezo Pattang

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

Master of Science
in

Biology

Carleton University
Ottawa, Ontario

Abstract

Soybean [*Glycine max* (L.) Merr.] is a valuable crop, with benefits attributed to its high protein and oil content, in addition to its nitrogen fixating capabilities. To expand production across Canada, breeding programs will need the availability of markers linked to desired traits, including time of flowering and maturity. The E7 locus is among the 10 maturity loci identified in this pathway, whose underlying gene remains unknown. Using functional genomics resources, and a computational approach utilizing PIPE (Protein-protein Interaction Prediction Engine), this region was narrowed to a short-list of candidates, including 3 promising candidates: *Glyma.06G200400*, *Glyma.06G200800*, and Glyma*.06G220000*. From the expression analysis performed to date, *Glyma.06G199800* and *Glyma.06G233300*, were found to have significant change in expression between the E7/e7 lines in contrasting flowering stages during long-day (LD) photoperiod. However, additional expression analysis on the remaining candidate's, along with further experimentation needs to be performed to reveal the underlying gene for E7.

Acknowledgments

First, and foremost, I would like to thank my direct supervisor Dr. Bahram Samanfar for his valuable guidance, encouragement, and providing an inspirational work environment that allowed me to both grow professionally and personally. I am forever grateful for the valuable kindness and leadership he showed throughout my studies.

I am appreciative to my co-supervisor, Dr. Ashkan Golshani and members of my thesis committee, Dr. Elroy Cober and Dr. Tim Xing for their valuable support and insightful comments throughout this project.

I am grateful to the members of Samanfar lab for welcoming me to their team, and sharing their knowledge and expertise with me. I want to thank Martin Charette, and Doris Luckert, members of the Trifoliates: Julia Hooker and Nour Nissan, Siwar Haidar, and former lab member Michael Sadowski.

I am forever indebted to my dear parents, for being the greatest role models I could have asked for. Their everlasting encouragement and support have made this dream of mine a reality. To my sisters, Maryam and Nilofar, thank you for always making sure I didn't take life too seriously and enjoy the little things.

Finally, I would like to thank my better half, Jeremy Ballentine, for always believing in me. His confidence in my ability, unconditional love, and amazing sense of humor have provided me with the strength and motivation to achieve this goal of mine.

Table of Contents

Abstract ii

Acknowledgments iii

List of Tables vii

List of Illustrations viii

List of Appendices x

List of Abbreviations xi

Chapter 1: Introduction 1

1.1 Soybean production in Canada 1
1.2 Challenges for Western Canada 3
1.3 Soybean growth and development 4
1.4 Time of flowering and maturity 8
1.4.1 Molecular mechanisms in *Arabidopsis thaliana* / *Oryza sativa* 9
1.5 Soybean genomics 12
1.6 Molecular basis of flowering in soybean 15
1.6.1 E1 (*Glyma.06G207800*) 17
1.6.2 E2 (*Glyma.10G221500*) 18
1.6.3 E3 (*Glyma.19G224200*) 18
1.6.4 E4 (*Glyma.20G090000*) 19
1.6.5 E5 (*nonexistent*) 19
1.6.6 E6 and J (*Glyma.04G050200*) 19
1.6.7 E7 (*unidentified*) 20
1.6.8 E8 (*unidentified*) 20
1.6.9 E9 (*Glyma.16G150700*) 21
1.6.10 E10 (*Glyma.08G363100*) 21
1.6.11 E11 (*unidentified)* 22
1.7 Molecular markers 22
1.8 Genetic mapping 24
1.9 Computational and systems biology 26

1.10 Computational approaches in Protein-Protein Interactions (PPI) 27
1.11 Purpose and objective 30
Chapter 2: Materials and methods 32

2.1 Plant material 33
2.2 Sample collection and DNA extraction 33
2.3 Sequencing (Genome Quebec) data analysis 34
2.4 PCR and sequencing 34
2.5 SNP database 35
2.6 Computational analysis 35
2.6.1 Protein-protein Interaction Prediction Engine (PIPE) and Gene Ontology (GO) analysis 36
2.6.2 Loss of Function (LOF) analysis 37
2.6.3 RNA-Sequencing (expression) database analysis 38
2.7 Identifying conserved domains 38
2.8 2D RNA structure analysis 38
2.9 Expression analysis 39
2.10 Digital PCR (dPCR) 40
Chapter 3: Results 40

3.1 Identification of candidate genes involved in time of flowering and maturity 40
3.2 Sequencing data analysis for contrasting E7 lines 44
3.3 Computational analysis of candidate genes 46
3.5 Expression analysis with RT-qPCR and dPCR 48
3.6 Candidate gene summary 49
Chapter 4: Discussion 61

4.1 Analysis of candidate genes 61
4.2 REDUCED VERNALIZATION RESPONSE 1 (VRN1) gene - (sub-group 1) 62
4.3 SUGARS WILL EVENTUALLY BE EXPORTED TRANSPORTER (SWEET16/17) - (sub-group 1) 64
4.4 EXOCYST SUBUNIT EXO70 FAMILY PROTEIN - (sub-group 1) 64
4.5 EARLY BOLTING IN SHORT DAYS (EBS) – (sub-group 2) 65
4.6 WRKY family transcription factor family protein - (sub-group 2) 66
4.6 PHOTOPERIOD-INDEPENDENT EARLY FLOWERING1 (PIE1) - (sub-group 2) 68
Chapter 5: Conclusion and future direction 69

5.1 Conclusion 69
5.2 Future direction 71
5.2.1 Expression analysis 71
5.2.2 Vernalization experiment and photoperiod induction 71
5.2.3 Compensation analysis 73
5.2.4 Allele-specific marker development (CAP/dCAP and KASP) 73
References 75

Appendix 86

List of Tables

Table 1. Genotype of soybean lines used in this study 33

Table 2. Short-list of candidate genes selected by whole genome sequencing analysis..... 43

Table 3. Summary of sequence, and type of variations (SNP, INDEL) identified from GQ data, location within exon or intron, along with amino acid changes resulting from exon variations, and intron variations in the form of INDELs..... 45

Table 4. Conserved domains associated with each candidate gene. Variations within each functional domain is also shown along with the size of the domain in bp 46

Table 5. Gene candidates ranked on priority based sequence variations between E7/e7 genotypes, with additional evidence provided computational analysis using PIPE paired with GO 48

List of Illustrations

Figure 1. Average soybean production in Canada 2017-2019 3

Figure 2. Soybean growth and developmental stages 6

Figure 3. Maturity groups in Northern and Southern America 7

Figure 4. Conserved GI-CO-FT pathway in Arabidopsis and rice 11

Figure 5. Experimental workflow 32

Figure 6. Simplified algorithm for PIPE workflow 37

Figure 7. Process to narrowing the E7 region to shortlist of candidates 42

Figure 8. Relative (normalized) fold change for candidate genes, A)*Glyma.06G199800*, B) *Glyma.06G180300,* C) *Glyma.06G233300* and D)*Glyma.06G239100,* 49

Figure 9. Conserved domain of *Glyma.06G220000* A) intron (black spaces)/exon (beige) map. B) plant-specific B3-DNA binding conserved 51

Figure 10. 2D RNA structure prediction of *Glyma.06G220000* 51

Figure 11. Conserved domain of *Glyma.06G200400* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) Exo70 and Cytochromse b subunit conserved domains 52

Figure 12. 2D RNA structure prediction of *Glyma.06G220000* 52

Figure 13. Conserved domain of *Glyma.06G200800* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) Laminin G and PQ loop repeat conserved domains 53

Figure 14. 2D RNA structure prediction of *Glyma.06G200800* 54

Figure 15. Conserved domain of *Glyma.06G199800* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) Ribosomal protein S2 and the C-terminal helicase conserved domain 55

Figure 16. Conserved domain of *Glyma.06G233300* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) Bromo Adjacent Homology conserved domain 56

Figure 17. Conserved domain of *Glyma.06G242200* A)intron (black spaces)/exon (beige) map with UTR's (grey). B) WRKY DNA-binding conserved domain within 57

Figure 18. Conserved domain of *Glyma.06G200200* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) PQ loop repeat, nodulin MtN21 family, and laminin G conserved domain 58

Figure 19. Conserved domain of *Glyma.06G180300* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) Neprosin and Neprosin_AP conserved domain 59

Figure 20. Conserved domain of *Glyma.06G2023000* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) p450 conserved domain..... 60

Figure 21. Conserved domain of *Glyma.06G239100* A) intron (black spaces)/exon (beige) map with UTR's (grey). B) no conserved domain..... 60

List of Appendices

Appendix 1. Amino acid sequence variation between E7/e7 genotypes 86

Appendix 2. RNA sequence data provided by Severin et al. (https://soybase.org/soyseq/). 89

Appendix Table 3. Housekeeping genes and primers used for expression analysis. 90

Appendix Table 4. Primers used for sequencing. 91

Appendix Table 5. Genome Quebec sequencing analysis for the E7 lines (OT93-26 and OT89-9) and e7 lines (OT02-18 and OT98-17). 99

Appendix Table 6. Representative PIPE raw data for candidate Glyma.06G200400. 100

Appendix Figure 1. Representative GO identified for the top 200 interacting partners identified from the raw PIPE. 103

Appendix Figure 2..Human-Soybean Allergies: Elucidation of the Seed Proteome and Comprehensive Protein-Protein Interaction Prediction 104

List of Abbreviations

Abbreviation	Definition
ABA	abscisic acid
AFLP	Amplified Fragment Length Polymorphism
AP1	APETALA 1
ARF	Auxin Response Factors
AT	*Arabidopsis thaliana*
ATP	Adenosine triphosphate
BAH	Bromo-Adjacent Homology
CAPS	Cleaved Amplified Polymorphic Sequence
CDF	Cycling DOF Factors
CGH	Comparative Genome Hybridization Array
cM	Centimorgan
CO	CONSTANS
RAV1	cold-responsive transcription factor
CXMS	Chemical Cross-Linking of Proteins coupled with Mass Spectrometry
dCAPS	derived Cleaved Amplified Polymorphic Sequence
DE	Drought Escape
DOE-JGI	Department of Energy Joint Genome Institute
dPCR	Digital PCR
E1L	E1-Like
EBS	Early Bolting in Short Days
EHD1	Early Heading date 1
EUB	Urea Extraction Buffer
EXO70	EXOcyst 70
FLC	FLOWERING LOCUS C
FRI	FRIGIDA
FT	FLOWERING LOCUS T
GI	GIGANTEA
GO	Gene Ontology
GQ	Genome Quebec
GWAS	Genome Wide Association Mapping
Hd1	Heading Date 1
IDC	Iron Deficiency Chlorosis
KASP	Kompetitive Allele Specific
LamG	Laminin Globular domain
LAV	LEAFY COTYLEDON2-ABI3-VAL
LCT	Lysomial Cysteine Transporter
LD	Long Day
LHY	Late Elongated Hypocotyl
LOF	Loss Of Function
MFE	Minimum Free Energy

MG	Maturity Group
MudPIT	Multidimensional Protein Identification Technology
NGS	Next Generation Sequencing
NIL	Near Isogenic Lines
PHD	Plant homeodomain
PIE1	Photoperiod-Independently Early Flowering
PIPE	Protein-protein Interaction Prediction Engine
PPI	Protein-Protein Interactions
QTL	Quantitative trait locus
R	Reproductive
RAPD	Random Amplified Polymorphic Detection
REM	REPRODUCTIVE MERISTEM
RFLP	Restriction Fragment Length Polymorphism
RFT1	Rice Flowering Loc T 1
RP-PPI	Reciprocal Perspective for PPI prediction
SD	Short Day
SLAF-Seq	Specific-locus amplified fragment sequencing
SNPs	Single Nucleotide Polymorphisms
SOC1	SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1
SPRINT	Scoring Protein INTeractions
SSR	Simple Sequence Repeat
TAP	Tandem Affinity Purification
TE	Transposable Elements
TF	Transcription Factor
TOG	Transporter/Opsin/G Protein-Coupled Receptor
TSF	TWIN SISTER OF FT
V	Vegetative
VC	Vegetative Cotyledons
VE	Vegetative Emergence
VIN3	VERNALIZATION INSENSTIVIE 3
VRN1	Reduced Vernilization Response 1
WGD	Whole Gene Duplication
WGS	Whole Genome Sequencing
Y2H	Yeast Two Hybrid

Chapter 1: Introduction

Soybean [Glycine max (L.) Merr.] is a major crop cultivated worldwide because of its multi-use characteristics; it is used for human consumption, animal feed and industrial products. Soybean seed has protein content greater than many other food crops at (40-42%), contributing the largest source of vegetable oil and animal feed among all food crops [1]. Additionally, soybean is key to promoting sustainable agriculture management practices due to its nitrogen fixating properties. While soybean is successfully grown across the globe, Brazil is the world's leading producer as of 2019/2020, contributing 36.9% of the total world production, followed by the United States contributing 28.8% [2] [3]. Canada currently contributes 1.8%, however there is great potential to further expand production across Western and Northern regions of Canada [2] [3].

1.1 Soybean production in Canada

Canada is ranked as the 7th leading soybean producer globally, producing approximately 6-7 million metric tons annually, according to the Food and Agriculture Organization of the United Nations in 2014 [4]. It is forecasted that soybean export will rise by 29%, and planted area will increase by 5% in 2021 – 2022 [5]. Within Canada, the leading principal field crop in 2021 was wheat (durum, spring and winter), corn, canola, and barley with 21,714,800, 14,368,100, 12,781,900 and 7,141,300 metric tonnes produced, respectively. Soybean ranks 5th with 5,886,300 metric tonnes produced based on 2021 statistics from Statistics Canada [6].

Since its introduction to Canada in the mid1800s, the vast majority of soybean production has been in Southern Ontario, contributing to 61.3% of Canada's total production, followed by Manitoba and Quebec, contributing 18.6%, and 17.3%, respectively (Figure 1) [7]. While Ontario remains the leading producer, there have been successful signs of further expansion to Western

and Northern regions of Canada. This is due to the efforts put in place by breeding programs focused on developing lines with early maturity, suitable for growth across differing photoperiods, with an emphasis placed on Western and Northern regions of Canada. This was achieved by gaining a more in-depth understanding of the genetics involved in time of flowering and maturity, that has since allowed for the successful development of early and ultra-early lines. In 1976, Manitoba and Saskatchewan had contributed 309 and 183 hectares of soybean respectively in 1976, compared to 2018 with a reported 764,900 and 164,900 hectares contribution [8]. Additionally, in Canada there has been a 103% increase in production from 2009 to 2018.

While Canada has seen great advancements in soybean expansion, there is continuous efforts to further its expansion, largely due to the improvement in sustainable agriculture production and plant rotation (nitrogen fixating properties). There has been an emphasis in expanding production across Western and Northern regions of Canada that have a large availability of fertile land. To continue these expansion efforts, cultivars suited for various environmental conditions within Canada need to be developed. This requires a greater understanding of the genetics that underlie key mechanisms necessary for successful adaptation, including a major factor; time of flowering and maturity.

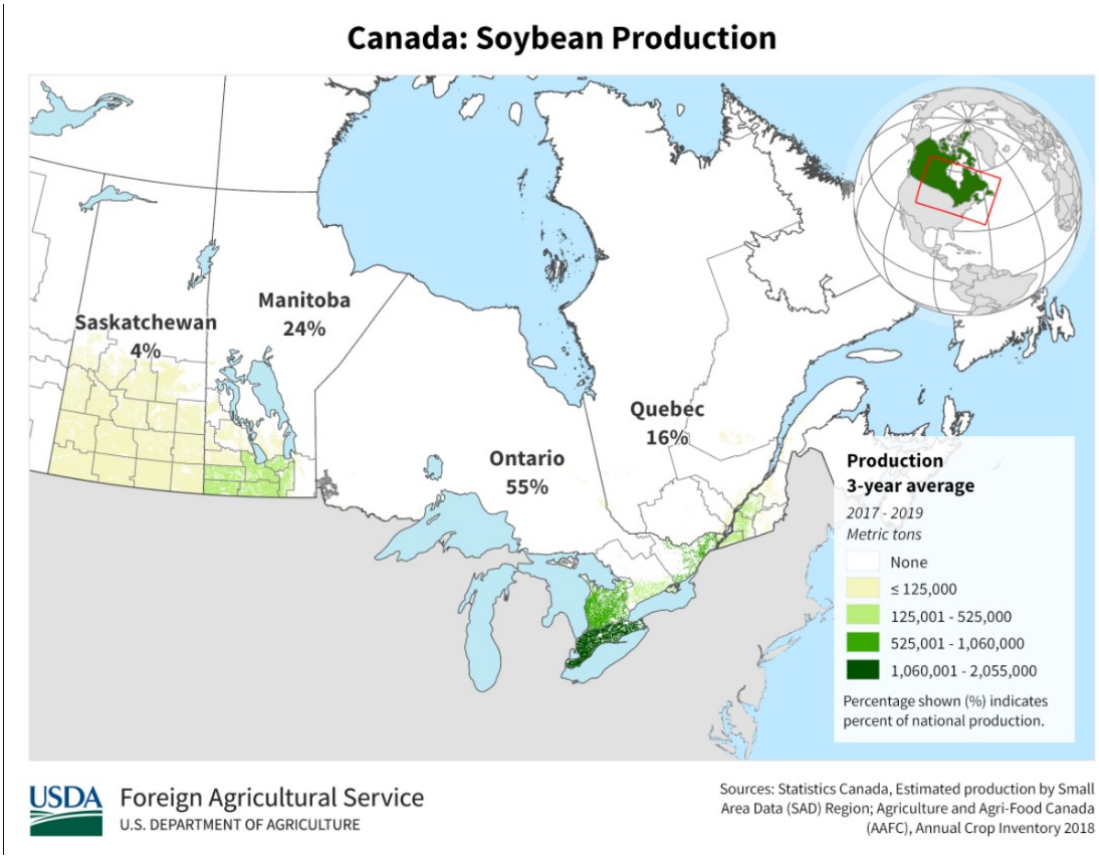


Figure 1. The 3-year average soybean crop production (metric tons) map of Canada from 2017-2019 [9].

1.2 Challenges for Western Canada

Soybean expansion across Western and Northern regions of Canada has posed a challenge, as soybean needs to overcome the external limitations of the region while also maintaining reasonably high yield and seed quality. In addition to the longer photoperiod that is a limitation to short-day (SD) crops, such as soybean, the crop needs to withstand abiotic stresses including, drought, flooding, iron deficiency chlorosis (IDC) and salinity. Soybean also needs to overcome challenges brought by biotic stresses including disease, pathogens (i.e. soybean cyst nematode (SCN)), and weeds. When a SD crop is grown in regions with LDs, this delays the time of flowering and maturity of the crop. This is especially concerning when growing soybean in Western and Northern regions of Canada where there are LDs, and shorter growing seasons. Thus,

posing a risk that the crop will not reach full maturity before the first frost of the season occurs. To be able to expand soybean production across those regions, a deeper comprehension of the underlying genetic mechanisms influencing time of flowering and maturity is essential. Only then will it be possible to develop soybean varieties adapted to diverse environmental conditions, or more specifically, developing early and/or ultra-early maturing varieties adapted for Western and Northern regions of Canada.

1.3 Soybean growth and development

Through human intervention, soybean has successfully migrated across the globe, adapting to diverse environmental conditions. To continue the successful migration of soybean to diverse ecological conditions, there needs to be a greater understanding of the genetic mechanisms involved in the growth and development of soybean, including influential genetic and environmental factors. Soybean plant development (both vegetative and reproductive) is dependent on many factors including day-length, temperature, maturity group (MG), etc. [10]. Soybean is referred to as a SD crop (i.e. long nights) as they initiate floral induction more rapidly under SD conditions. Soybean growth stages can be separated into three phases; vegetative (V), reproductive (R) and senescence (maturity). However, depending on the variety there are three different growth habits that exist; (a) indeterminate - the plant undergoes vegetative and reproductive growth simultaneously until the V5 growth stage, (b) determinate - the plant undergoes most of its vegetative growth before initiating reproduction and (c) semi-determinate – the plant continues vegetative growth even after flowering.

The vegetative growth stages (Figure 2, A) is descriptive of the plant from emergence to right before the first flower. These growth stages are determined based on the number of nodes on the main stem, beginning with the unifoliate nodes with a fully developed leaf, and go from

descending order “VE, “VC”, “V1” to V(n)” [11]. The emergence of the cotyledons from the soil surface is referred to as the vegetative emergence (VE) stage. This is followed by the vegetative cotyledons (VC) stage, observed by the elongation of the hypocotyl, and the production of unrolled unifoliate leaves. As the main stem continues to grow, and the leaves of the unifoliate nodes develop is denoted as V1. All vegetative stages following VC are designated as V (n), where “n” represent the number of nodes on the main stem with fully developed leaves [11]. The reproductive stages follow, and are classified based on flowering, pod and seed development, and plant maturation. The reproductive stages are categorized by “R1” to “R8”, using the main stem to determine the respective stage (Figure 2, B) [12]. The R1 growth stage is during floral initiation when the first flower on the main stem is open. R2 is when flowers are present within a fully developed trifoliate leaf. Depending on the variety, it is possible to have the R1 and R2 growth stages occur simultaneously, seen in determinate varieties. The beginning of pod formation and full pod formation define the R3 and R4 growth stages, respectively. While the beginning of seed formation, and full seed formation are classified as the R5 and R6 growth stages, respectively. The crop begins maturity when at least one pod on the main stem has reached its maturity color, often indicated by a brown color, this is classified as the R7 growth stage. Lastly, when 95% of the pods have reached maturity, this indicates that the seed growth has completed and the plant is considered fully mature, classified as the R8 growth stage [11].

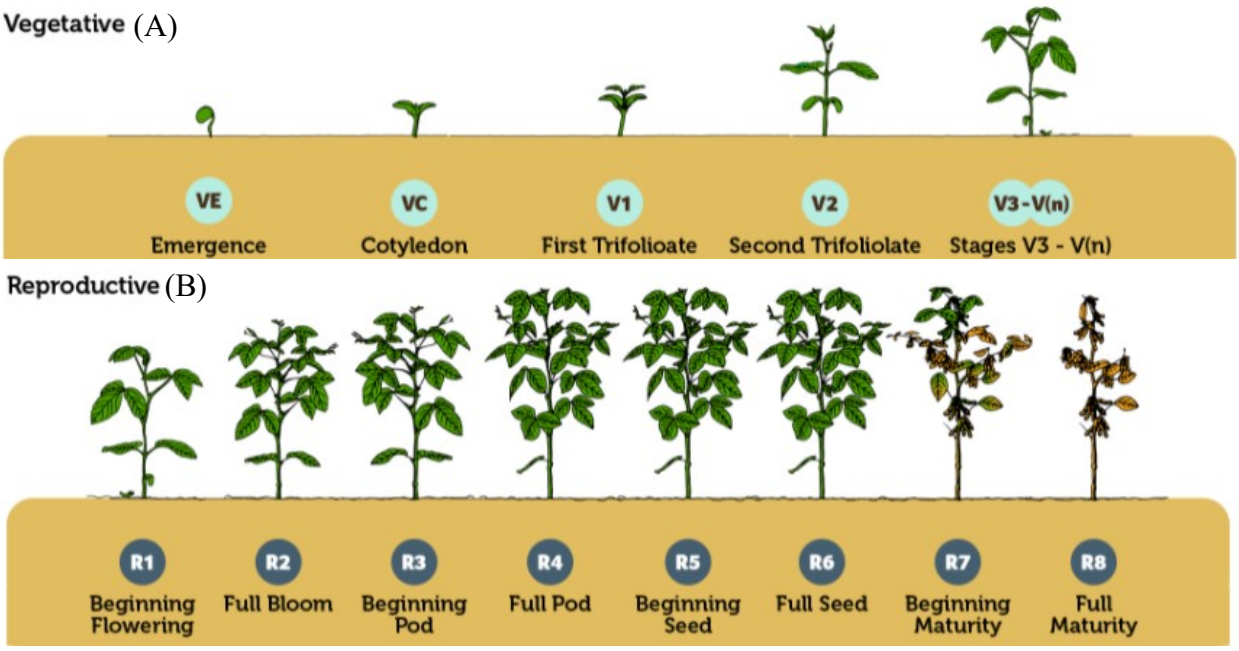


Figure 2. Soybean growth during both (A) vegetative stages that range from the emergence from the soil (VE) to the development of nodes on the main stem with fully developed leaves (V1 – V(n)). (B) the reproductive stages that range from floral initiation (R1) to pod development (R3 and R4), seed formation (R5 and R6) and plant maturation (R7 and R8) [13].

The successful adaptation of soybean across continents has been made possible in part due to the development of genetically diverse cultivars that vary in time to flowering and maturity. This divergence in time to flowering, along with allelic variations at the soybean maturity loci are used to classify soybean in maturity groups. Soybean cultivars are categorized with respect to their response to photoperiod in select geographic regions, and the output in which the highest yield is obtained. The Regional Soybean Laboratory of the USDA has taken into consideration the soybean response to varying latitudes, creating a MG classification system. Cultivars have been categorized into 13 MGs, ranging from MG 000 found in Southern Canada that are comprised of very early maturing varieties that are able to thrive in higher geographic latitudes with longer photoperiods (Figure 3). The MG X that are late maturing varieties grown in lower latitudes and shorter photoperiods that can be found in Mexico and the Caribbean Islands [14] [15]. In Canada, the MGs range from MG 000 to MG III, with MG III primarily grown in Southern regions of Canada,

MG 0 to MG 00 dominant in Ontario, and MG 000 dominant in Northern regions including Manitoba and Saskatchewan [14].

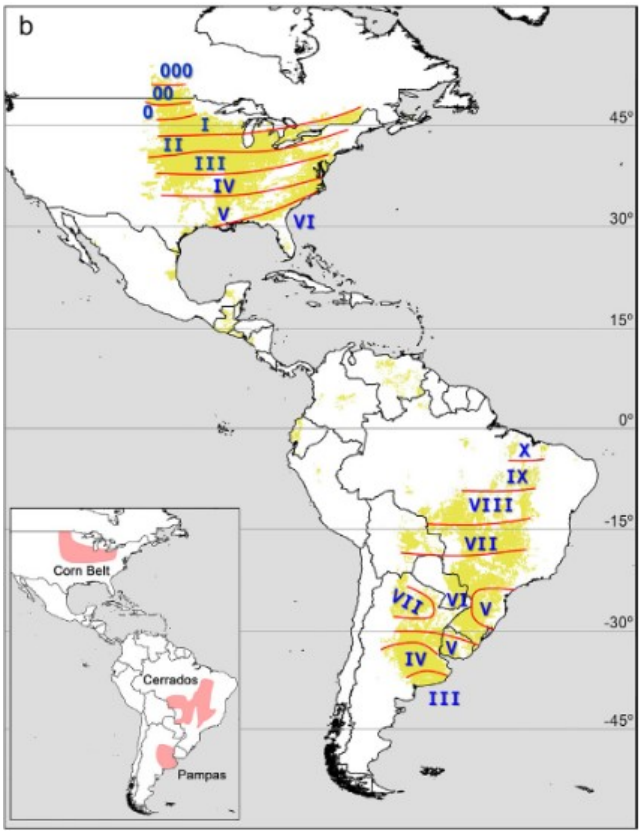


Figure 3. Distribution of soybean cultivars categorized into maturity groups, showing dominant maturity group respective to geographic latitude [16].

The adaptability of flowering plants is dependent on their ability to recognize environmental stimuli and use this to maximize their reproductive development [17]. The full scope of the molecular mechanisms involved in processes related to how plants utilize photoperiodic information and relate this to flowering is yet unknown. However, it is necessary to gain a greater understanding of these mechanisms, as time of flowering and maturity are important components necessary for the successful adaptation of soybean plants to new geographic regions.

1.4 Time of flowering and maturity

The transition from vegetative phase to the reproductive phase is strongly dependent on the plants ability to use external and internal (physiological) information in a meaningful way towards floral initiation, including day length, temperature, and developmental age [17]. Through a comprehensive understanding of the major genes underlying the mechanisms that determine time of flowering and maturity, it can be possible to gain a more in-depth understanding of all the mechanisms involved in the regulation and coordination of a flowering response [18]. While the full scope of mechanisms involved in the photoperiod pathway are not fully understood, it is known that photoperiod plays a key role in the flowering response [19]. Plants are classified into three groups based on their flowering response, (a) plants that flower as the day length increases (LD), (b) plants that flower as the day length decreases (SD), and (c) plants that flower independent to photoperiod (day-neutral plants) [17].

The full scope of mechanisms and pathways involved in time of flowering and maturity in legumes, such as soybean still remains unclear, compared to well-studied model organisms including *Arabidopsis thaliana* (LD) and *Oryza sativa* (SD). Through comparative mapping analysis, it has been determined that the flowering pathways have been conserved in diverse plant species, including but not limited to Arabidopsis, brassica and barley [20]. These species can therefore be used as model organisms to better understand the mechanisms involved in time of flowering and maturity in soybean. In addition, due to the polyploidization of the soybean genome, it is considered more complex, with many redundant genes, low transformation efficiency, and long transgenic regeneration process. Therefore, to confirm the involvement of candidate genes in these pathways, one can use the well-studied model plants as an intermediate tool. There is high

confidence in using this approach, since it has also been used in the past to successfully confirm the involvement of other maturity genes in time of flowering and maturity [21].

1.4.1 Molecular mechanisms in *Arabidopsis thaliana* / *Oryza sativa*

The model plant, *Arabidopsis thaliana* has been extensively studied and has greatly contributed to the understanding of the molecular mechanisms and pathways involved in (but not limited to) time of flowering and maturity, including both environmental and endogenous factors. Despite the estimated 200 million years monocot-dicot divergence, there are still key genes conserved between *Arabidopsis thaliana* and rice (*Oryza sativa*) [22].

A key contrast between Arabidopsis and rice is their floral initiation in response to photoperiodic cues. Arabidopsis is a facultative LD species, with flowering initiated under LD conditions. In contrast, rice is a SD species with flowering initiated under SD conditions. The circadian clock provides plants the ability to anticipate changes in day length, distinguishing between LD and SD. This endogenous mechanism provides plants the ability to modulate their developmental programs to run in conjunction with environmental factors, including day length and temperature [23] [24]. Thus, the circadian clock is key to the success in important developmental transitions, including time to flowering. In Arabidopsis for example, the expression of several hundred genes have been identified to be regulated by this endogenous circadian oscillation, with expression oscillating between the day and night cycle. The expression of a key component to the photoperiodic pathway, CONSTANS (CO) is also regulated by the circadian clock [23].

The underlying mechanism behind the photoperiod pathway can be separated into three key parts, including the input of light information, circadian clock and output of information that induces the expression of select genes. The input of light information is integrated into the

photoperiodic mechanism that is influenced by the circadian clock to induce the expression of select genes that regulate flowering [24]. In Arabidopsis, genes involved in the photoperiod pathway are regulated by endogenous factors (i.e. circadian clock), and environmental factors (i.e. photoperiod) to regulate time of flowering and maximize development success. A key pathway identified in controlling the time of flowering in response to photoperiod cues include *CONSTANS* (*CO*), a zinc finger transcription factor (TF), GIGANTEA (GI), and *FLOWERING LOCUS T (FT),* a florigenic protein that promotes flowering. *CO* is key in optimal floral initiation, its expression is dependent on the circadian clock, as it is only able to accumulate at the end of a LD. Only then is it able to accumulate and activate the expression of FT in the vascular tissue of leaves and induce flowering (Figure 4, A). The expression of the CO-FT pathway, however, is strongly dependent on the activity of the protein complex that is formed between GI and FLAVIN BINDING, KELCH REPEAT, F-BOX 1 (FKF1). This complex is also dependent on the circadian clock, as the expression of the two genes need to coincide under LD conditions in order to form a stable complex. This GI-FKF1 complex is only then able to target the degradation of CYCLING DOF FACTORS (CDF) genes. These CDF proteins prevent the expression of CO and FT, by binding to their promoter sites during SD conditions. However, under LD conditions when the GI-FKF1 complex is stable, it is able to degrade the CDFs, and as a result allow for the non-repressed CO and FT to form a complex and allow for flowering to occur [25].

A similar photoperiodic pathway that mediates day length response in Arabidopsis can be found in rice. Homologues for the Arabidopsis *GI* and *CO* were identified in rice, as *OsGI* and *Heading Date 1 (Hd1)*, respectively (Figure 4, B). In addition, several FT homologues are identified in the rice genome, including *Hd3a*, *RICE FLOWERING LOC T 1 (RFT1)* and *FT-like 1 (FTL1)* that are found to promote flowering when overexpressed. Similar to Arabidopsis, the

expression of *Hd1* is key in transition to developmental stage. However in contrast to *CO*, whereas its expression and consequently flowering are activated under LD conditions, *Hd1* was identified as bi-functional, promoting flowering under SD conditions by activating the expression of Hd3a and repressing flowering under LD conditions conversely repressing the expression of Hd3a [26]. Unlike soybean the OsGI-Hd1-Hd3a pathway also contains two proteins that promote flowering under both SD and LD conditions via regulation of the *Hd3a*. Including *Early Heading date 1 (Ehd1)* and *Grain number, plant height and heading date 7 (Ghd7)* [25].

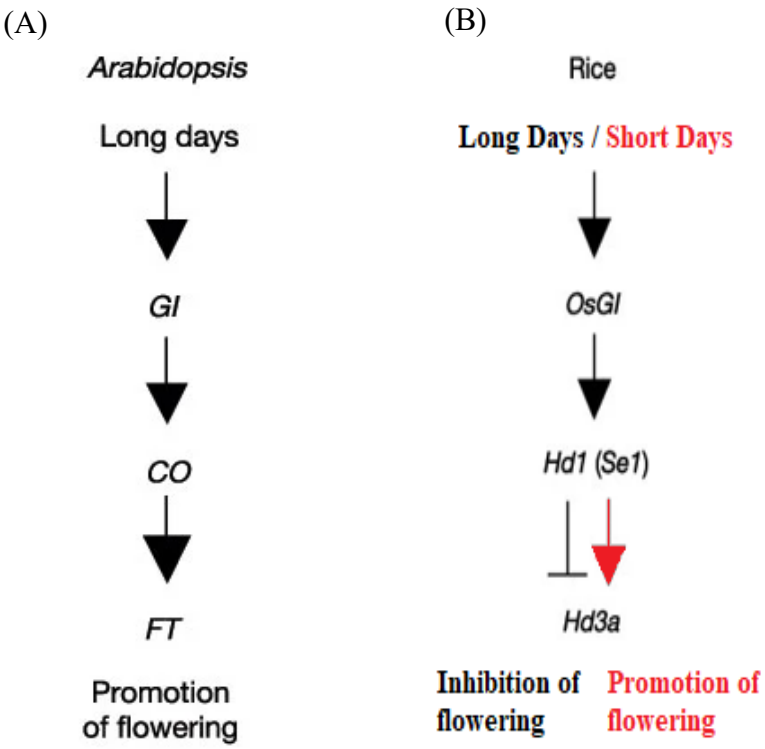


Figure 4. Conserved and distinct mechanisms for the regulation of flowering in rice (SD plant) and Arabidopsis (LD plant). (A) the expression of the floral initiator, FT is dependent on CO under LDs, (B) the expression of the FT homolog in rice Hd3a is dependent on the CO homologue, Hd1 that has a bi-functional role, promoting floral initiation under short-day, and inhibiting under long day conditions [26].

The mechanisms and pathways involved in time of flowering and maturity in soybean is not as well understood in comparison to model plants such as Arabidopsis and rice. To identify the best model plant for this project, a phylogenetic analysis of genes known to be associated with

the photoperiod flowering pathway across Arabidopsis, rice and soybean had been conducted. The findings had confirmed that commonly known photoperiod flowering genes, among Arabidopsis and soybean had a closer sequence homology than that of rice and soybean [26] .While both rice and soybean are SD crops, it is hypothesized that the greater similarity between Arabidopsis and soybean has to do with their physiology as both are dicots, while rice is a monocot. In addition to the greater understanding of the time of flowering and maturity mechanism in Arabidopsis, it also has a closer sequence homology to soybean. Therefore, Arabidopsis has previously been used as a model organism for soybean, due to its small genome, short life cycle and its successful use in previous studies to elucidate the function of novel soybean genes. Having said that, Arabidopsis seems to be an excellent model plant to further investigate genomic approaches in soybean.

1.5 Soybean genomics

Soybean, *Glycine max* (L.) Merr., belongs to the Leguminosae family; the domestication of the modern cultivated soybean is believed to have occurred approximately 6,000 to 9,000 years ago in East Asia, from wild soybean (*Glycine soja* Sieb. & Zucc) [16]. The domestication is believed to have been a gradual process, with recent analyses of the genetic differentiation and gene flow among soybean accessions including (*Glycine max, Glycine gracilis* and *Glycine soja)* concluding that *G. gracilis* is a transitional species, and that *G. soja* is the progenitor of both *G. gracilis* and *G. max*. This is due to significant gene flow observed from *G.soja* to *G. gracilis* and from *G. gracilis* to *G. max*, in addition to moderate gene flow from *G. soja to G. ma*x [27].

The soybean genome is considered moderately large at ~1.1 GB (20 chromosomes), it is the product of a diploid ancestor (n=11) that had undergone aneuploid loss (n=10), polyploidization (2n=20) and diploidization (n=20). The genome had undergone two rounds of whole genome duplication (WGD) and has a predicted 46,430 high confidence protein coding loci

and another ~ 20,000 predicted loci with low confidence and 88,647 transcripts [2] [3] [4]. The soybean genome was first sequenced using whole genome shot-gun approach by the soybean research community in collaboration with the Department of Energy Joint Genome Institute (DOE-JGI) in 2010. Using Sanger sequencing, ~85% of the reference cultivar, Williams 82 genome was assembled, including 1.9% gaps. Approximately 57% of the genome was determined to be repeat rich, with the majority classified as transposable elements (TE)s. Further analysis of these repeat rich regions, specifically the repeat rich low recombination heterochromatic regions, located them in centromeres. Among the repetitive elements found in the soybean genome, the long terminal repeat retrotransposons account for 42%, consisting of 510 families and 14,106 intact elements [28].

The closest relative of the cultivated soybean (*G. max*) is the undomesticated wild soybean (*G. soja*), both of which have been reported to have (2n = 40) number of chromosomes. While morphologically they differ, with *G. soja* characterized as having specific traits that were of interest during domestication, including: seed hardness, stem growth, pod color, time of flowering and pod shattering [29]. However, the genetic diversity that was prevalent when soybean was first discovered is no longer the case. Through farmer selection, *G. soja* was domesticated to Asian landraces, with only phenotypic and genetic traits of interest highlighted. This genetic bottleneck was intensified upon the introduction of landraces to North America, with only traits ideal for growth in the new environment considered of interest.

To evaluate the genetic diversity retained from the domestication of *G. soja*, many studies were performed, comparing the two genotypes. Findings support the hypothesis that domestication resulted in a genetic bottleneck, with the mean proportion of diversity retained from the 5 independent studies calculated to being 0.65 [30]. A genomic comparison of *G. max* and. *G. soja*

was also made by [31], concluding that the *G. soja* genome sequence covered 91.65% of the *G. max* genome sequence, with the two strains varying by 35.3 Mb. In the regions where the two sequences aligned, *G. soja* was found to consist of 0.267% SNPs, 0.043% INDELs, and 3.45% large deleted sequences. As a result of continuous selective breeding, the current soybean genome has been reported to have lost 50% of its genetic diversity, 81% of its rare alleles and 60% of the genes have shown a change in allele frequency [31]. Advancement in understanding the mechanisms involved in time of flowering and maturity are strongly dependent on understanding how molecular components interact and give rise to biological systems and their corresponding phenotypes that would be of agronomic significance. While genomic analysis of *Glycine max* accessions has provided a basis to identify novel genes that would otherwise not be present in the reference genome. There is still significant amount of research that needs to be conducted, often determining the function of a gene through experimental investigation. An approach that has helped bridge the gap between genotype and phenotype is functional genomics. This approach leverages experimental evidence obtained from various –omics, including analyses of gene expression at the mRNA (transcriptomics) and protein (proteomics) level, as well as the accumulation of metabolites (metabolomics), and genomics [32].

A key facilitator in integrating the various –omics was the completion of the soybean genome in 2009 [33]. Genomics is the study of an organism's entire genome, providing up-stream level insight into the adaptability of organisms to various environmental conditions based on the genetic variability seen among different populations. The soybean genome initially was used to target phenotypic traits that were controlled by single dominant genes or a major quantitative trait locus (QTL). There was a greater emphasis on mutations easily observed through phenotype, targeting candidate genes that were specifically correlated to traits including resistance, abiotic

stress tolerance, silencing, and chlorophyll content. The reference genome was also used to identify structural variants within and between soybean accessions, through the development of a comparative genome hybridization array (CGH). Another –omics approach used in functional genomics studies, transcriptomics investigates an organism's entire transcriptome, giving insight into how varying external and internal conditions give rise to the regulation of gene expression of cells. Taking into consideration both the soybean reference genome and high-throughput sequencing has allowed for the use of genome-wide mRNA sequencing (RNA-seq) analysis, and the development of SSR and SNP markers, etc. [32] [34]. This additional information has provided a more in-depth understanding regarding the gene sequence, alternative splice variants, gene expression profiles, and polymorphic sites. Proteomics is aimed at investigating the proteome of an organism including analyzing changes in protein expression, structure, function and post-translational modifications. Metabolomics is the study of molecular metabolites produced by an organism under a given set of conditions, with extensive variations in respect to both their physical and chemical properties [31] [32].

The advancement of high-throughput technologies that include next generation sequencing (NGS), microarray analysis, multidimensional protein identification technology (MudPIT), etc., have contributed to a significant increase of data sets among the biological systems. Thus, the computational and bioinformatics tools are useful when attempting to understand the function of a cell along with the interactions of its macromolecule constituents [32].

1.6 Molecular basis of flowering in soybean

As of today, 10 maturity loci (known as the E-series) have been identified to be associated with time of flowering and maturity in soybean, including, E1-E4, E6/J, E7-E11 [35]- [36]. The reduced sensitivity to LD photoperiods is attributed to the dysfunctional alleles at the E1-E4, E7,

E8, and E10 locus, or partially functional allele in the case of e1-*as*, while the recessive alleles at E6/J, E9 and E11 attributed to the delay in time to flowering and maturity. The underlying gene for E1 to E4, E6/J, E9, and E10 have been characterized at a molecular level using fine mapping, and molecular biology approaches [37]. It should also be noted that there are many other genes known to be involved in time of flowering and maturity but not categorized under the E-nomenclature, including Dt1 (*Gm*TFL1b) and Dt2 (*Gm*TFL1b) [38].

The GI-CO-FT module in Arabidopsis involved in the photoperiod flowering pathway is highly conserved in species including rice and soybean. In Arabidopsis the FT gene family plays a vital role in time of flowering, encoding for a florigen that induces the development of floral meristems. In the soybean genome, 11-FT like genes, including (*GmFT1a*, *GmFT1b*, *GmFT2a* (E9), *GmFT2b*, *GmFT2c*, *GmFT3a*, *GmFT3b*, *GmFT4 (E10)*, *GmFT5a*, *GmFT5b*, *and GmFT6)* have been identified [39]. Among these, (*GmFT2a*, *GmFT2b*, *GmFT3a*, *GmFT3b*, *GmFT5a*, and *GmFT5b*) were found to promote flowering, *GmFT6* does not seem to affect flowering, and GmFT4 represses flowering [39]. However, a greater emphasis has been placed on understanding the mechanism behind *GmFT2a* and *GmFT5a*, as their overexpression was found to play a major role in photoperiod flowering response. The Arabidopsis homologues, PHYA (E3 and E4) and GI (E2) were also found to regulate flowering, as their recessive alleles increased the expression of GmFT2a under LDs, promoting flowering.

The soybean genome has 23 CONSTANS-like (COL) genes, including two pairs of homologous genes, GmCOL1a and GmCOL1b and GmCOL2a and GmCOL2b that have the highest sequence similarity to the Arabidopsis CO [39]. While both pairs functioned similarly to CO in flowering in Arabidopsis, it was concluded that the role of legume COL genes in the control of flowering varies within the legume species.

1.6.1 E1 (*Glyma.06G207800*)

The E1 locus was discovered as having the greatest influence on time of flowering and maturity, resulting in ~31% of observed variations seen among varieties [35] [36] [40]. The E1 locus was mapped to chromosome 6, with the underlying gene being *Glyma.06G207800,* consisting of one exon within a 864-bp region. The E1 locus was identified as a flowering repressor, encoding for a TF that down regulates *Gm*FT2a and *Gm*FT5a, that are orthologs of the Arabidopsis FT, that functions as a floral inducer [41]. Positional cloning determined that E1 is related to the RAV subfamily of B3 domain proteins, among which includes the Arabidopsis TEMPRANILLO genes that have been previously identified as transcriptional repressors of FT [42]. At least five recessive alleles have been identified, including e1-*as* (amino acid substitution in the putative nuclear localization signal), e1-*fs* (a 1-bp deletion in codon 17, resulting in a frame shift) and e1-*n1* (a 130 kb deletion that includes the entire E1 gene), *e1-b3a* (5 bp mutation in the middle of the B3-like domain), and the mutations in *e1-re, e1-p* occur only at the 5'UTR region of the E1 gene, resulting in a (retrotransposon insertion), and (allele from cultivar 'Peking') respectively, with all recessive alleles resulting in an earlier time to flowering [43] [44].

Besides E1 on chromosome 6, two E1 homologues have been identified on chromosome 4. These two E1-like (*E1L*) homologs are found to be 96% identical to each other, and 91% identical to E1. In contrast to E1, these homologs only have minor effect on time of flowering and maturity. The two E1L genes, *Glyma.04G156400* (E1La) and *Glyma.04G143300* (E1Lb), are located in the pericentromeric region of chromosome 4, which is homologues to the pericentromeric region of chromosome 6, where E1 has been located [45]. A recent study had also identified that E1Lb delayed time of flowering independently of E1 [46]. As the e1lb (loss-of function allele) was found to upregulate the expression of FT2a and FT5a, and flowered earlier

than E1Lb, under LD conditions. It appears that the E1L genes may have a weaker effect on time of flowering compared to E1, as it was found that the complete loss of the E1 function (e1-*nl*), or partial lack of function (e1-*as*) was not compensated by the functional E1L genes [45].

1.6.2 E2 (*Glyma.10G221500*)

The E2 locus was discovered by [35], located on chromosome 10. The underlying gene *Glyma.10g221500* consists of 14 exons and 14 introns in a 21.39 Kb region. E2 was identified as an ortholog of the *Arabidopsis* flowering gene GI, found to delay flowering under LD conditions through regulating the expression of *Gm*FT2a [47]. The alleles identified for this gene include E2-*in*, E2-*dl* and e2-*ns*. The E2-*in* and E2-*dl* alleles have been distinguished by either an insertion or deletion of 36 bp in the 8th intron, while e2-*ns* consists of a single base substitution in the 10th exon, resulting in a premature stop codon [48] [47].

1.6.3 E3 (*Glyma.19G224200*)

The E3 locus was identified by [49], located on chromosome 19. The underlying gene *Glyma.19g224200*, consists of four exons and four introns within an 89.00 Kb region [48]. E3 encodes a PHYTOCHROME A gene, *Gm*PHYA3, which delays time to flowering through suppressing the expression of *Gm*FT2a and *Gm*FT5a [50]. Five alleles for the gene has been identified, including two functional alleles; E3-*Ha* (allele from cultivar 'Harosoy' that has a 2.6 kb insertion after the third exon) compared to E3-*Mi* (allele from the cultivar ' Misuzudaizu'), both alleles contribute to the delay in flowering under LD conditions [51]; and three non-functional alleles that contribute to early time of flowering, including; e3-*tr* (large deletion after the third exon, resulting in a 13.3 kb deletion that includes exon 4), e3-*ns* (nonsense mutation), e3-*fs* (frameshift mutation) [51] [52].

1.6.4 E4 (*Glyma.20G090000*)

The E4 locus was identified by [49], located on chromosome 20. The underlying gene *Glyma.20G090000,* consists of four exons and four introns within a 56.74 Kb region. E4 encodes a PHYTOCHROME A gene, *Gm*PHYA2, identified to delay time to flowering through suppressing the expression of *Gm*FT2a and *Gm*FT5a [50]. Five alleles have been identified, including a functional E4 allele that is photoperiod sensitive and contributes to the delay in time to flowering under LD conditions, and five non-functional alleles that consist of a single-base deletion resulting in truncated proteins of different lengths, including: e4-SORE-1 (insertion of a T*y1/copia*-like retrotransposon in exon 1), and less common; e4-*oto* (allele from cultivar 'Otomewase'), e4-*tsu* (allele from cultivar "Tsukue-4'), e4-*kam* (allele from cultivar 'Kamaishi-17'), and e4-*kes* (allele from cultivar 'Keshuang) all have single base deletions at different positions in exon 1 and 2 [51, 52].

1.6.5 E5 (*nonexistent*)

The E5 locus has been identified by [53], using genotypes from MGs II and IV, specifically detected in the cross of Harosoy (e5) and PI 80837 (E5). This locus was described as having a similar effect on time to flowering and maturity as E2 [54]. However, a later study by [54] failed to identify a QTL for E5 when assessing the same cross, therefore concluding that the lateness in time to flowering and maturity originally believed to be E5 does not exist but is rather a result of the E2-*dl* allele [54].

1.6.6 E6 and J (*Glyma.04G050200*)

Recently, E6 and *J* were identified as the same locus [55]. The delay in time to flowering under SD conditions has been defined as a long-juvenile trait. It was first identified in PI 159925, with the underlying gene identified as *Glyma.04G050200,* consisting of four exons and five introns

within a 53.60 kb region. The E6 locus was identified by [56], from natural mutations identified in Parana (MG VI). The E6/*J* locus is located on chromosome 4, recently identified as different alleles of the EARLY FLOWERING 3 (ELF3), that suppresses the transcription of E1. There are ten *J* alleles (*J*, *j-1* to *j-8* and j-x), all the recessive alleles have some type of loss-of function mutations, resulting in the delay in time to flowering under LD conditions [55].

1.6.7 E7 (*unidentified*)

The E7 locus was identified by [57], as a locus involved in time of flowering and maturity influenced by photoperiod response, with photoperiod experiments concluding that the E7 results in delayed time to flowering, with the recessive allele flowering ~6 days earlier in LD conditions. Using 'Harosoy'-derived NILs, it was determined that E7 is located on LG C2 and while tightly linked to E1, E7 has been previously shown to be non-allelic to E1 [58]. The most probable location for E7 was determined to be a 22.2-cM region, between Satt100 and Satt460, which also contained the diagnostic marker Satt319. Supporting the findings of [57] [59], a major QTL related to time to flowering and maturity was mapped to the same region on chromosome 6 [60]. Taking these previous findings into consideration, the E7 region has been narrowed down to the physical interval of 15,742,176 bp – 42,126,497 bp on chromosome 6, consisting of ~1350 genes.

1.6.8 E8 (*unidentified*)

The E8 locus was identified by [61], located on chromosome 4. With the use of SSR marker analysis the E8 locus was found to be linked to LG C1, between Sat404 and Sat136 [61]. In accordance with studies conducted by [61] [62] [60], the underling gene for the E8 locus has been confined to the physical interval 7,137,389 bp to 43,856,157 bp on chromosome 4, consisting of ~1000 genes. While the underlying gene for E8 has yet to be discovered, it is known that E8 results in delayed time to flowering, with the recessive e8 flowering ~5-8 days earlier in LD conditions

[61]. Certain time to flowering candidates have also been suggested to be controlled by E8 due to their proximity to the QTL, including *Gm*CRY1a and *E1La* and *E1Lb,* however additional studies need to be performed to confirm their candidacy.

1.6.9 E9 (*Glyma.16G150700*)

The E9 locus was identified by [63], the underlying gene (*Glyma.16G150700*) was found to consist of four exons and three introns within a 245 kb region. The dominant *Gm*FT2a promotes early flowering under LD conditions. The recessive allele for *Gm*FT2a consists of a *Ty1/copia*-like retrotransposon, SORE-1, inserted in the first intron that contributes to its allele-specific transcriptional repression and thus weakening the *Gm*FT2a expression, and therefore delays time to flowering [64].

1.6.10 E10 (*Glyma.08G363100*)

The E10 locus was identified by [65], located on chromosome 8, and contains four exons and three introns within a 16.88 kb region. The recessive allele consists of three SNPs located in the 5' UTR, 3' UTR and the fourth exon, resulting in 5-10 earlier maturity under SD conditions [65]. This new gene was first discovered using presumed E10E10 and e10e10 genotypes, OT98-17 and OT02-15 respectively. While both were identified as having the same genotype for the known maturity loci (e1 e2 e3 e4 e7 e7), OT98-17 still matured 6 days earlier. The QTL for the E10 locus was identified through investigating SSR and SNP haplotypes between contrasting genotypes, and their association to time of flowering and maturity. A functional genomics approach was used to investigate the 75 genes identified in the genomic region of this novel QTL, including; (a) Protein-protein Interaction Prediction Engine (PIPE) to identify top candidates found to interact with known maturity and flowering genes, (b) comparing the allelic sequence variation of the mRNA 2D structure, and (c) comparing SNPs between the two candidates. Through PPI

analysis using PIPE, two FT4 candidate genes were identified as most likely underlying the E10 locus, while sequence analysis identified three SNPs resulting in differential mRNA structure. Therefore, it was suggested that E10 encodes *Gm*FT4, *Glyma.08G363100,* an orthologue of the Arabidopsis *FT* that acts downstream of E1 to repress flowering under LD conditions.

1.6.11 E11 (*unidentified)*

The E11 locus was identified by [36], located on chromosome 7. The dominant allele was found to promote time to flowering and maturity under LD conditions. With the use of specific-locus amplified fragment sequencing (SLAF-Seq) technology, 3 stable QTLs related to time to flowering were discovered on chromosome 5, 6, and 7 that were qTOF5, qTOF13, and qTOF7, respectively. However, in accordance to previous literature, the QTL responsible for E11 was constricted to qTOF7 located on chromosome 7. Using RHLs, the E11 gene was fine mapped to a 138 kb interval, with three candidate genes identified through amino acid sequence analysis. However, in accordance to previous studies, it was suggested that *Glyma.07G48500* is the most likely candidate gene for E11, and it is a homolog of LATE ELONGATED HYPOCOTYL (LHY). The expression of LHY has been found to influence time of flowering independent of photoperiod, as the loss of function allele, *lhy-1* resulted in earlier flowering under both SD and LD conditions [36]. E11 allele contributes to early flowering and maturity under LD conditions, and e11 allele contributes to delayed flowering and maturity.

1.7 Molecular markers

Approximately 75% of the genes currently present in North American soybean cultivars can be traced back to 17 ancestors [66]. Conventional breeding methods have been used in the past by artificially mating or cross-pollinating superior plants in order to develop plants with advantageous traits. However, these methods were often found to be time consuming, labor

intensive, limited to the same or closely related species and visibly observable traits. Modern methods, however, are dependent on genetic engineering that is based on selecting specific gene(s) that have been discovered to be responsible for desired traits, and not limited to phenotypic traits. This advancement has allowed for breeders to obtain optimal crops in a shorter period of time. In the past, molecular markers have been used to identify alleles of interest in many plants including soybean. This includes alleles responsible for traits such as resistance to SCN, brown stem rot, among many others in soybean cultivars [67] [68].

Molecular markers can be classified as either a) classical markers, which include morphological and biochemical markers, or b) DNA markers. Morphological markers are based on identifying phenotypic traits that are closely linked to genetic traits of interest. Morphological markers can be considered the most direct method for measuring phenotype, however, these markers have their limitations. For example, given that environmental variability may influence phenotypic outcome of the crop, the usefulness of such markers is limited, and minimum association with traits of economic value, including yield and seed weight have been identified [69]. Biochemical markers are based on protein variations, with two known classes; a) isozymes which are described as allelic variants of the same enzyme, but often encoded by different loci and b) allozymes which are different proteins that are encoded by different genes, but have the same function. The variation among such proteins can be determined according to their net charge and size [69]. While biochemical markers are co-dominant, and cost effective they are unable to detect many of the possible polymorphisms, with the data influenced by the plants growth stage at the time of harvest and is also affected by the extraction methodology [69].

DNA markers, also referred to as genetic markers, play a significant role in molecular breeding. These markers often have a known location, either in the form of a gene or DNA sequence that is associated with a specific trait of interest. DNA markers are beneficial when wanting to identify polymorphisms in the form of mutations among alleles in a population or gene pool [69]. The ideal DNA marker is said to be within the target gene [70]. Commonly used molecular techniques that have assisted in finding genetic markers are in the forms of restriction fragment length polymorphism (RFLP), random amplified polymorphic detection (RAPD), amplified fragment length polymorphism (AFLP), simple sequence repeat (SSR), and single nucleotide polymorphisms (SNPs) to name a few. Aside from these traditional methods, there has been the development of additional techniques, including cleaved amplified polymorphic sequence (CAPS), also referred to as RFLP-PCR, and a variation of CAPS referred to as derived cleaved amplified polymorphic sequence (dCAPS). Both CAPS and dCAPS are able to detect polymorphisms in the form of SNPs and INDELs within a population, based on the presence or absence of restriction sites present following PCR amplification and gel electrophoresis [71]. Kompetitive allele specific PCR (KASP) assay is another genotyping approach used to detect SNPs and short INDELs by using allele-specific PCR with fluorescence-based reporting system [72].

1.8 Genetic mapping

Genetic maps are used to show the position of distinct features within a genome, including the position of markers, QTLs and genes along a chromosome (linkage groups). Genetic maps are a useful tool for navigating across the genome, providing information on the proximity of genes or regions of interest to genetic markers.

Crops can display a variety of different traits that are controlled by a single or a small number of genes. These are referred to as qualitative traits, and often follow the Mendelian fashion of inheritance. However, polygenic traits that are more often of an agronomic importance, including yield, and disease resistance, are often controlled by a complex group of genes. These are referred to as quantitative traits, with the regions containing these gene(s) referred to as QTLs. For QTL mapping, the parental lines used to generate the mapping populations should ideally differ for traits of interest, with the resulting population segregating for these traits [73]. Commonly used mapping populations include near-isogenic lines, backcrosses, recombinant inbred lines and double haploid [73]. The typical number of individuals for a mapping population is from 50-250 individuals, when generating maps for minor traits, larger populations (>500) are used [74]. Commonly used DNA markers for QTL mapping include RFLPs, RAPDs, SSRs, AFLPs and SNP [69]. This method of identifying QTLs however has limitations, including: a) environmental factors, b) experimental errors in phenotyping and c) size of population. The expression of quantitative traits can be influenced by environmental factors, including temperature, region and photoperiod. However, having consistency in external factors can be challenging and may influence the expression of traits of interest from one growth season to the next. There can be a great level of variation when describing traits of interest, thus resulting in variability in phenotypic data. QTL mapping is also limited to the allelic diversity present within the population being assayed, and the number of recombination events that limits the mapping population resolution [69] [75]. An alternative approach to QTL mapping is association mapping that is dependent on the structure of linkage disequilibrium within the genome. The efficiency of association mapping is strongly dependent on the sample size and the genetic diversity present within the population [76]. Association mapping can be categorized into two methods; a) candidate

gene association mapping and b) genome-wide association mapping (GWAS). Candidate gene association mapping is arguably the more expensive strategy, but emphasizes identifying important alleles, as well as rare alleles that would otherwise be missed. This approach analyzes the presence of polymorphisms within a gene of interest starting from a limited germplasm collection (typically 24 to 48), polymorphisms identified would then be screened across a larger germplasm collection (100 to 1000) to confirm the SNP-influencing phenotype association. GWAS is used to identify SNP-influencing phenotype, through a process of screening large populations and comparing their respective phenotypes [76]. There are two ways in assessing SNP-influencing phenotypes, (a) direct association, when the SNP is directly genotyped and statistically associated to the phenotype, and (b) indirect association, when the SNP is not directly genotyped, but statistically associated to the phenotype [77].

1.9 Computational and systems biology

Organisms are able to execute diverse biological functions through complex biological systems. While traditional wet-lab approaches used to investigate biological systems involve studying each part individually, this is not ideal as these approaches are costly and have low-throughput. Traditional approaches are also limited in providing an understanding of how living systems function as a whole, as these functions do not occur in response to an individual component, but in response to interactions that occur between several components all as part of one system, giving rise to new properties and function.

In recent years, there has been significant advancement in computational and systems biology. This is in part due to the increased availability of biological data in the form of DNA, RNA and protein sequence, and in part due to the continuous increase in computational power. As a result, the scientific community has made major advancements in understanding how biological

systems function at a cellular and molecular level, giving rise to the discovery of novel protein interactions and complexes, new metabolic pathways, and much more. Advancements in computer power has provided a useful tool in identifying PPIs, and is currently used to predict novel interactions based on previously provided data. This does however result in limitations, as the computational predictions made are dependent on the data provided. Machine learning and artificial intelligence in computational and systems biology have provided the opportunity in providing models for predicting protein function, and interacting partners based on the primary amino acid sequence [78]

There are currently three different approaches used to better understand biological systems, including a bottom-up approach in which information gathered at a molecular level (cells and tissues) is used to generate insight on the dynamic behavior and function of cells, organs and organisms [79]. The top-down approach in contrast uses bioinformatics, specifically investigating the -omics related data, as described below. And lastly, there is the middle-out approach that takes into consideration both bottom-up and top-down approaches. For the purposes of this study, the "top-down' approach has been selected to identify the underlying gene for the E7 maturity locus.

1.10 Computational approaches in Protein-Protein Interactions (PPI)

The majority of biological processes, including gene expression, cell growth, and proliferation are facilitated by a network of proteins through Protein-protein interactions (PPI). Investigating novel PPIs, especially in the proper biological context has been found to be difficult, due to the dynamic processes involved in protein expression, as well as the influences by various stimuli. In theory, if an unknown protein interacts with other proteins that are known to be involved in a specific pathway or function, then it is most likely that the unknown protein is also involved

in that pathway or function, a concept known as *guilt by association*. Therefore, when investigating an unknown protein, it is beneficial to investigate it in accordance to its interacting partners.

Experimental methods used to investigate PPIs, include yeast two hybrid (Y2H) that uses the Bait and Prey model to identify the interaction between protein pairs, however, this method has a high rate of false positives [80]. There are also pull-down assays such as tandem affinity purification (TAP), that is able to identify protein complexes that exist *in vivo*. However, the introduction of a tag to the protein may alter the protein's properties. Other methods involved in characterizing PPIs include using chemical cross-linking of proteins coupled with mass spectrometry (CXMS), and thermal stability shift analysis that could be used under certain circumstances to monitor PPI [81]. While these experimental methods have provided an insight into PPIs they have associated technical limitations that include being (a) time consuming and labor intensive, (b) difficult to identify weak PPIs, and (c) prone to false positives and false negatives [82].

Recently, computational methods for predicting novel PPI have been used based on previously acquired experimentally derived PPI data from genomic information, evolutionary relationship, three-dimensional protein structure, conserved domains and primary protein structure. The genomic information provides insight to the proximity of two genes in different genomes and can be used to predict their interaction. The evolutionary relationship provides insight to which protein pairs with similar phylogenetic profiles in different genomes are predicted to interact. The three-dimensional protein structure analysis uses the protein structures to predict the most compatible regions for their interaction. Identifying proteins that interact and share a conserved domain can provide as a useful tool for predicting other proteins that also contain the

same domain and are predicted to interact. The primary protein structure of known interacting protein partners can also be used as tool for predicting novel PPIs [80].

A computational tool that has shown to be successful in predicting novel PPIs based on the primary structure, and has proven to be the method of choice for predicting complex PPIs is the Protein-Protein Interaction Prediction Engine (PIPE) [83] [84] [65] (Appendix Figure 2). The PIPE tool was developed by a multi-disciplinary team from the Departments of Biology, Engineering and Computer Science at Carleton University [80]. The PIPE4 has shown to be successful in the elucidation of comprehensive interactomes for organisms including *Homo sapiens*, *Arabidopsis thaliana*, *Glycine max* and others. PIPE has also been used to generate a comprehensive interactome for cross-species predictions, for example the model species *Arabidopsis thaliana* has been used as a proxy species to predict the comprehensive *Glycine max* interactome [85]. PIPE is able to overcome limitations associated with other sequence similarity based methods, by applying the Reciprocal Perspective modeling framework, termed Reciprocal Perspective for PPI prediction (RP-PPI). This model takes into consideration the putative interaction between two elements, A and B in the context of all putative interactions from the perspective of A, and all putative interactions from the perspective of B [84]. The domain-based method is dependent on the availability of previously identified domains, without the potential in identifying novel interaction sites that are beyond the known interacting domain. In contrast sequence similarity-based methods are able to identify sequences that facilitate known PPIs based on key properties of the amino acid residues from previously defined experimental data. However, the sequence based method cannot be used for high-throughput protein network wide analysis, as protein interactions are typically mediated by only small protein segments and the prediction accuracy is limited. When compared to other PPI methods, PIPE was found to outperform the other available methods in recall-

precision, specificity, sensitivity as well as its improves processing speed. PIPE was found to have an especially high specificity (up to 99.95%, with less than 0.05% false positives), and will be discussed in more detail in the methods section.

1.11 Purpose and objective

The objective of this research is to identify the underlying gene for the E7 maturity locus in soybean. The soybean is a very important crop to Canadian agriculture due to its economic and sustainability characteristics. The majority of soybean production is currently in Southern Ontario, however there has been interest to further expand its production across Western and Northern regions of Canada. While there has been substantial increase in soybean production across Western provinces of Canada, there are still environmental limitations, including photoperiod that are restricting further expansion. Since soybean is a SD plant, when grown in LD conditions present in Western and Northern regions of Canada, flowering and maturity is delayed to an extent in which the crop is unable to mature before the first frost of the season. To overcome these limitations, soybean breeding programs need to develop early and/or ultra-early cultivars that would be suitable for these regions with an acceptable yield and grain quality.

As the world demand for soybean continues to grow, Canadian soybean needs to stay competitive in the global market. It is projected by Soy Canada (https://soycanada.ca) that Canada will double soybean production to 13 million tonnes by 2027, through integrated pest management, improved agronomic practices, and new varieties with improved genetics [86]. To meet Soy Canada's projections, there needs to be a greater understanding of the mechanisms behind key factors, including time of flowering and maturity. While these pathways have been extensively investigated over the past several decades, there is still a gap in our understanding of the key components involved. Continuous efforts to identify and characterize novel genes will advance the

understanding of the different components that work together to regulate these pathways, including time of flowering and maturity. To date, 10 maturity loci (known as the E-series) have been identified to be associated with time of flowering and maturity in soybean, with the underlying genes for E1 to E4, E6/J, E9, and E10 characterized at a molecular level using fine mapping, and molecular biology approaches.

The E7 locus has previously been characterized by three independent studies on chromosome 6, within a region consisting of ~1350 genes. Considering the sheer magnitude of genes, the region was first analyzed based on the assumption that a mutation between the wild and mutant lines is contributing to the differing phenotypic trait in time to flowering and maturity. This was achieved by first employing a Illumina sequence analysis on the wild and mutant lines, to identify consistently different SNPs and INDELs between the two types. The candidate genes that were found to consist of polymorphisms consistently different between E7 and e7 were placed in a short-list and further analyzed using a bioinformatics approach, PIPE paired with GO. To predict if any of the genes within the short-list have an association with genes known to be involved in time of flowering and maturity. In addition, other resources were also employed including available LOF mutants, and RNA sequence data. Using these approaches, the short-list of candidates was prioritized based on evidence supporting their involvement in time of flowering and maturity. To determine the potential association of these candidates with the E7 maturity locus in soybean, they were further analyzed, by applying molecular biology related practices (wet-lab), including but not limited to expression analysis via qPCR.

While E7 is only a component to this highly complex process, identifying and elucidating the function of its underlying gene will provide additional evidence to support how different components work together to regulate time of flowering and maturity. Identifying the underlying

gene for E7 maturity locus, and developing user-friendly and accurate allele-specific markers will also assist soybean breeding programs in developing ultra-early maturity soybean cultivars suitable for Western and Northern regions of Canada.

Chapter 2: Materials and methods

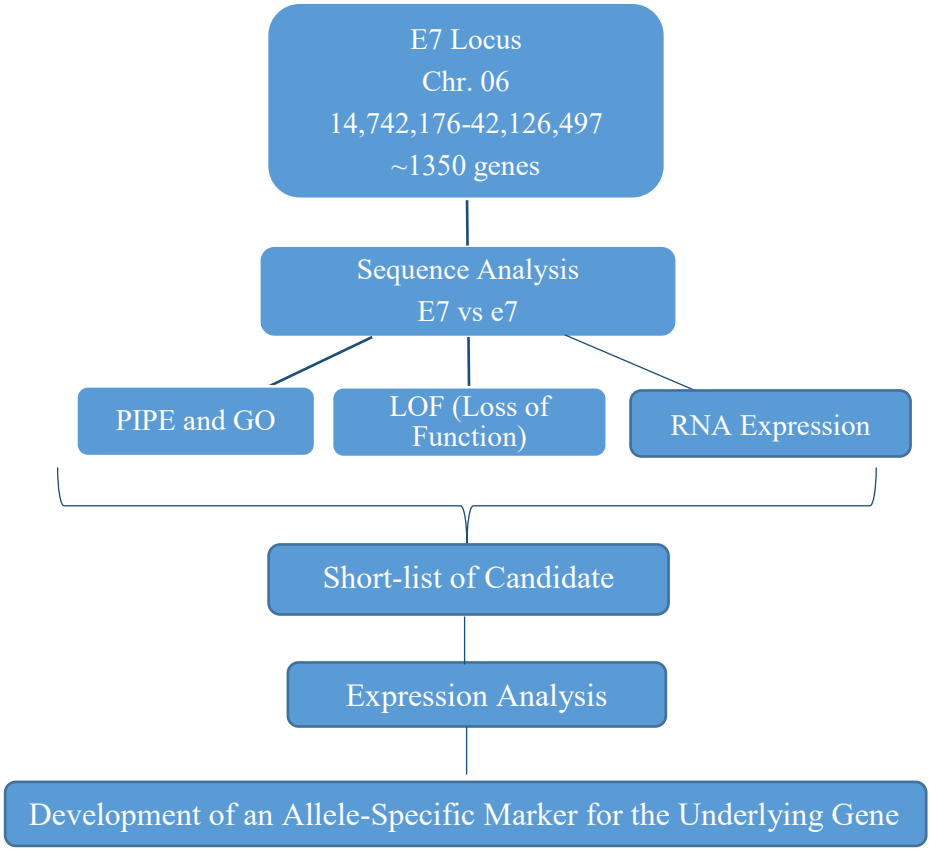


Figure 5. Experimental workflow to identify the underlying gene for E7 maturity locus.

2.1 Plant material

AAFC-Ottawa RDC soybean breeder Dr. Elroy Cober provided soybean lines with contrasting E7/e7 genotypes.

Table 1. Genotype of soybean lines used in this study.

Line	Genotype	Pedigree	Reference
OT93-26	T Dt1 E1 e3 e4 E7	OT89-5/L71-802	[59]
OT89-5	t Dt1 e1 e3 e4 E7	PI 438477/2*'Evans'//7*L62-667	[59]
OT94-41	t Dt1 e1 E3 e4 E7	OT89-5/L67-153	[59]
OT94-51	t dt1 E1 e3 e4 E7	L71-802/OT89-5//OT89-6	[59]
OT94-47	t Dt1 e1 e3 e4 e7	OT89-5//PI 196529/6*L62-667	[59]
OT02-18	e1 e3 e4 e7	X824A-ve/3 * Maple Presto/2/3 * OT89-5/3/3 * OT94-47	[61]
OT98-17	e1 e3 e4 e7	X824A-ve/7 * Maple Presto	[61]
OT89-9	T dt1 e1 e3 e4 e7	L67-153/7*'Maple Presto'	[59]

2.2 Sample collection and DNA extraction

The soybean plants were propagated in the Greenhouse at Ottawa, Canada 45°23′33″N 75°43′00″W, using soil containing; 75% Promix BX, 24% Black Earth, and 2% Lime. The plants were grown in 13-hour photoperiod with day temperature 25°C and night temperature of 20°C. Leaf tissue from each soybean line was collected during the V4-V5 growth stage from the young trifoliate leaves and immediately frozen in liquid nitrogen. Genomic DNA was extracted from the tissue samples using a urea extraction buffer (EUB) method [59]. Using ~200 mg of ground tissue (frozen) and 500 µl of EUB (8M urea, 700 mM NaCl, 100 mM tris (pH 8), 20 mM EDTA, 3% sarcosyl, and 3% SDS at 55°C), the sample was mixed by inversion and incubated at room temperature for ~ 5-10 min with occasional mixing. This was followed by the addition of 500 µl phenol/chloroform/isoamyl (25/24/1), mixed thoroughly then centrifuged at 15,000 for 10 min at 4°C. The upper aqueous phase was transferred to another tube, and 500 µl chloroform/isoamyl (24/1) was added, mixed by inversion and centrifuged at 15,000 for 10 min at 4°C. The aqueous phase was transferred, and 200 µl of 5M NaCl was added and mixed by inversion. This was

followed by the addition of 500 µl of isopropanol that was mixed by inversion and incubated at -20°C for 10 min, and centrifuged at 15,000 for 5 min at 4°C. The sample was washed using, 1 ml of 75% ethanol with 10 mM ammonium acetate, incubated for 10 min at 4°C then centrifuged for 2 min at 12,000 at 4°C. The supernatant was removed, and the process repeated an additional 2 times. The liquid was removed, and sample dried using the SpeedVac. The DNA pellet was then suspended in 100 µl of TE buffer (10mM Tris, 1mM EDTA (pH8.0)) and incubated for 30 minutes in a 37°C heating block. Both the purity and concentration of the extracted DNA was determined using a NanoDrop 2000 spectrophotometer.

2.3 Sequencing (Genome Quebec) data analysis

Nucleotide variations (NVs) between the wild and mutant type E7 alleles were identified by investigating four accessions including two E7 lines (OT93-26 and OT89-5) and two e7 lines (OT02-18 and OT98-17) that were outsourced for sequencing through Genome Quebec (GQ), using Illumina NovaSeq 6000 S4 PE150. NVs, in the form of INDELs, and SNPs that were consistently different between the E7 and e7 genotypes and in close proximity to a gene were further investigated, i.e. being that it was upstream, or downstream a gene (Appendix Table 1). The candidates identified as homozygous reference (0/0), or alternative (1/1) were of interest, in addition to candidates that were identified as inconclusive (./.).

2.4 PCR and sequencing

In addition to the results provided by GQ, a short-list of candidates was re-sequenced at an in-house sequencing facility at Agriculture and Agri-Food Canada, Ottawa Research and Development Centre (AAFC-Ottawa RDC) to reconfirm variations among more lines. The genomic sequence for each candidate gene was extracted from www.soybase.org, with primers designed using PRIMER3 web software (https://primer3.ut.ee/) (Appendix Table 4). To identify

the optimal annealing temperature for each primer pair, a PCR gradient from 52.0°C to 61.9°C, using TaKaRa Ex Taq™, with TaKaRa recommended reagents and cycling conditions were performed. The optimal conditions were then tested against the 8 lines as shown in Table 1, to confirm the presence of only the targeted product. The samples were then purified using the ExoSAP-IT reagent, followed by performing sequencing reaction using BigDye™ Terminator v3.1, using the respective recommended ThermoFisher protocols [87]. The samples were then sent to a sequencing facility at AAFC-ORDC, with chromatograms analyzed using the MegAlign 15 and SeqMan Pro 15 programs offered alongside the DNASTAR Lasergene 17 software.

2.5 SNP database

A collection of 530 Canadian and international soybean accessions were subject to whole genome sequencing (WGS) [88] [89]. Among these accessions were four wild-type E7 lines (OT93-26, OT89-5, OT94-41, and OT94-51) and four mutant-type e7 lines (OT04-47, OT02-18, OT98-17, and OT89-9) (Table 1). The SNPs were only recorded if they consisted of a distinct nucleotide differentiation between the two alleles. The SNPs were further refined to the E7 region and their location analyzed, whether it being upstream, or downstream of a gene, as well as within an exon or intron [58] [59] [60].

2.6 Computational analysis

The E7 region was previously identified by [58] and supported by [57] [59] to a physical interval of 15,742,176 bp – 42,126,197 bp on chromosome 6 that consisted of approximately 1350 genes. These 1350 genes were narrowed to a short-list of 10 candidates using NGS. These 10 candidates were further investigated by analyzing available genomics, transcriptomics and

proteomics data such as PIPE, GO, RNA, LOF mutants, and RNA-sequence data for the changes in expression during the different growth stages.

2.6.1 Protein-protein Interaction Prediction Engine (PIPE) and Gene Ontology (GO) analysis

PIPE is a computational tool that predicts proteome-wide PPIs based on short reoccurring amino acid sequences, and a database of known interactions [80]. The PIPE algorithm relies on a database of experimentally verified protein interactions, and while experimentally validated protein interactions can have large number of false positives (up to 40%), the PIPE database is carefully constructed to avoid false positives, only including protein interactions that have been validated through multiple experiments. A simplified algorithm for PIPE is presented in (Figure 6), within the known interaction database, the PPIs among V, W, X, Y and Z are known. To determine whether or not A and B are interacting, PIPE will investigate the similarity, similarity matrix, within a 20 amino acid long window from protein A along with all the known interacting data set shifting by one amino acid every round until the end of protein A is reached. Assume it found that a subsequence in protein A, resembles a subsequence in protein V and W from the database; PIPE then selects all known interacting partners with V and W, in this case being X, Y and Z, and investigates the similarity, similarity matrix, of a 20 amino acid long window from protein B across the interacting partners (X,Y, Z). Finally, using computational calculations that include similarity matrix indexes, PAM scores, etc., PIPE is able to predict if A and B are interacting or not (at a given specificity (99.95%) and sensitivity (23.3%)). The specificity was chosen to minimize the number of false positives, calculated as (TN/(TN+FP)) [%], while the sensitivity of PIPE represents PIPE's ability to detect correctly true interacting partners, calculated

as (TP/(TP+FN)) [%], where TP is the number of true positive, FN is the number of false negatives, TN is the number of true negatives, and FP is the number of false positives [84] [80] [90].

The PIPE data was used alongside GO analysis to further support the involvement of the candidate genes in time of flowering and maturity. Among the short-list of candidates, the top 200 interacting partners for each candidate was identified using PIPE. These top interacting partners were analyzed through the "GO Term Enrichment Tool" using www.soybase.org, with interacting partners related to biological processes known to be involved in time of flowering and maturity, including flowering, developmental processes, embryo development, etc. identified. The total number of interacting partners identified for each candidate was used to further support the involvement of the identified candidates in time of flowering and maturity.

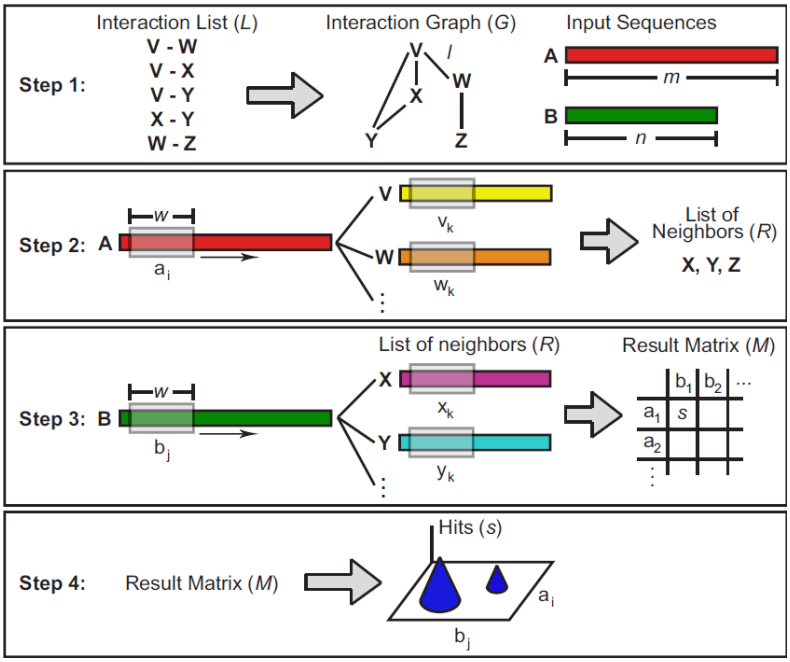


Figure 6. A simplified algorithm for PIPE workflow [91].

2.6.2 Loss of Function (LOF) analysis

The phenotypic difference seen between the wild and mutant E7 lines may be as a result of a mutation. In extreme cases mutations that result in premature stop codons or shifts in the reading

frame can be consequential to gene expression, resulting in a functional impact that can be seen as differential response to photoperiod. A database, consisting of whole-genome sequencing data for 1,007 national and international soybean accessions, including 18,031 LOF mutations in 10,662 genes was used identify LOF mutations among the candidates genes underlying the E7 region [88].

2.6.3 RNA-Sequencing (expression) database analysis

To assess the correlation between the candidate genes identified in this study to time of flowering and maturity, the expression data obtained from www.legumeinfo.org and www.soybase.org for the respective genes were analyzed. RNA expression across four tissues was assessed; flowering, leaf, pod, and roots, with an emphasis placed on the direction of change in expression, as well as the expression levels in flowering.

2.7 Identifying conserved domains

To determine the presence of conserved domains among the candidate genes, the NCBI conserved domain search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) was used. An emphasis was placed on identifying where the conserved domains lie within the candidate protein as this would provide insight into the functional impact a NV in the form of SNP or INDEL would have.

2.8 2D RNA structure analysis

Among the candidate genes, NVs between the E7 and e7 genotypes were identified. To determine if these NVs within the exon and UTR regions, resulted in significant structural variations the 2D RNA structure for the wild and mutant alleles were determined using the RNAfold WebServer from Vienna University (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi).

2.9 Expression analysis

The soybean lines with contrasting E7 genotypes (Table 1) were grown in the COVIRON growth chambers at Ottawa, Canada 45°23′33″N 75°43′00″W, using soil containing: 75% Promix BX, 24% Black Earth, and 2% Lime. The plants were grown under both SD (12-hour) and LD (20-hour) photoperiod under 25°C/20°C (day/night) temperature. Tissue samples were collected at three growth stages, including before flowering at the V2-V3 growth stages, during flowering at the R1-R2 growth stage, and after flowering, during full seed development at the R6 growth stage and placed in liquid nitrogen, then into a -80°C freezer. RNA was extracted using the TRIzol reagent (Cat. 15596026) from ThermoFisher, following the recommended procedure using 100 mg of tissue. The isolated pellet was suspended in 50 µl of RNase-free water, with the concentration and quality measured using the NanoDrop 2000 by ThermoFisher ensuring that the A260/A80 was at ~2.0. Following RNA extraction, the High-Capacity cDNA reverse transcription kit with RNase inhibitor from ThermoFisher (Cat 4374966) was used then for cDNA synthesis, using 1 µg of RNA. The cDNA samples were then amplified using traditional PCR, with the amplified fragments analyzed using gel electrophoresis, to confirm specific amplification, visible through one band.

The primers were designed following a similar outline used for sequencing, with a template size ranging from 50-200 bp, and GC content between 40 – 60% (Appendix Table 3). To quantify gene expression, the protocol recommended for the PowerTrack SYBR Green Master Mix (cat A46012) was used. Using the Micheal PfaffI method ($2^{-\Delta\Delta C}$) [92], a standard curve was generated for each candidate, along with the housekeeping gene (Tubulin), to ensure the efficiency was between the recommended 90 to 110% [5]. The mutant, e7 samples, were tested against the wild type, E7 samples, using three biological and technical replicates. The controls used included: a no

template control that contained no RNA or DNA to confirm there was no nucleic acid contamination, a no reverse transcriptase control that contained RNA instead of cDNA to identify contamination of DNA, as well as a no amplification control that contained no fluorescent dye to measure background fluorescence.

2.10 Digital PCR (dPCR)

Digital PCR (dPCR) was also used to quantify gene expression, by partitioning the reaction into many sub-reactions, each containing very few or no target sequence with the target detected using fluorescence following amplification. With the concentration of the target sequence determined using Poisson's statistics [93]. Similar to RT-qPCR, gene expression is quantified using PCR, Taq polymerase, and pre-validated primers. Similar to the RT-qPCR reaction, the same primers and cDNA samples were used but in lower concentrations. To identify the optimal annealing temperature, a gradient was run for the template cDNA using the EvaGreen Supermix from BIORAD. To monitor nucleic acid quantitation independent of reaction efficiency, measurements are made at the end-point using the Qiagen QuantStudio Absolute Q Digital PCR System.

Chapter 3: Results

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

The E7 locus was initially identified to be responsible for a +/-14-day delay in time of flowering under LDs [58]. While the underlying gene for E7 has yet to be identified, it was found to be tightly linked to E1 and T (tawny pubescence) all mapped to chromosome 6. Linkage between E1 and E7 was estimated to be 6.2 centimorgan (cM), and linkage distance between E7 and T was estimated to be 3.9 cM. The order was identified as T, E1, and E7, alternative order E1, T, and E7

has also been reported, suggesting the region is unstable with a high recombination rate, or also due to low resolution in the mapping population [59]. The most probable location for E7 was confirmed based on analysis using 'Harosoy'-derived near isogenic lines (NIL), to LG C2 within a 22.2 cM interval, specifically between SSR markers: Satt100 and Satt460, and linked to Satt319 [59]. Additionally, a major flowering and maturity QTL was mapped within the physical interval of 15,742,176 – 42-126,497, with approximately 1350 genes located within this region [60] [58] [59].

To establish a more manageable list of candidates to further investigate, NVs in the form of INDELs and SNPs identified by Illumina sequencing that were consistently different between the wild and mutant E7 lines were identified, through sequence analysis of four accessions including two E7 lines (OT93-26 and OT89-5) and two e7 lines (OT02-18 and OT98-17). Among the 1350 genes within the E7 region, only 10 candidates were identified to carry NVs that were consistently different between the E7 and e7 lines (Appendix Table 5). These candidates were then assessed using a computational tool, soybean PIPE (version 4) to identify the top 200 interacting partners for each candidate. These top 200 interacting partners were then analyzed using GO terms related to "Biological Processes", including embryo development, flowering, developmental processes involved in time of flowering and maturity, etc. (Appendix Figure 1, Table 5). The candidates were then ranked from having the most to least interacting partners known to be involved in time of flowering and maturity. These candidates were further screened using RNA-seq and loss of function (LOF) database and predicted 2D structure analysis (Table 2).

The combined wet-lab and computational tools has allowed for the successful identification of 10 candidate genes, that were prioritized based on NVs (Table 3) with 3 candidates identified

to consist of SNPs and INDELs within the exon region, 4 found to have INDELs in the intron region, and 3 candidates with SNPs in the intron region.

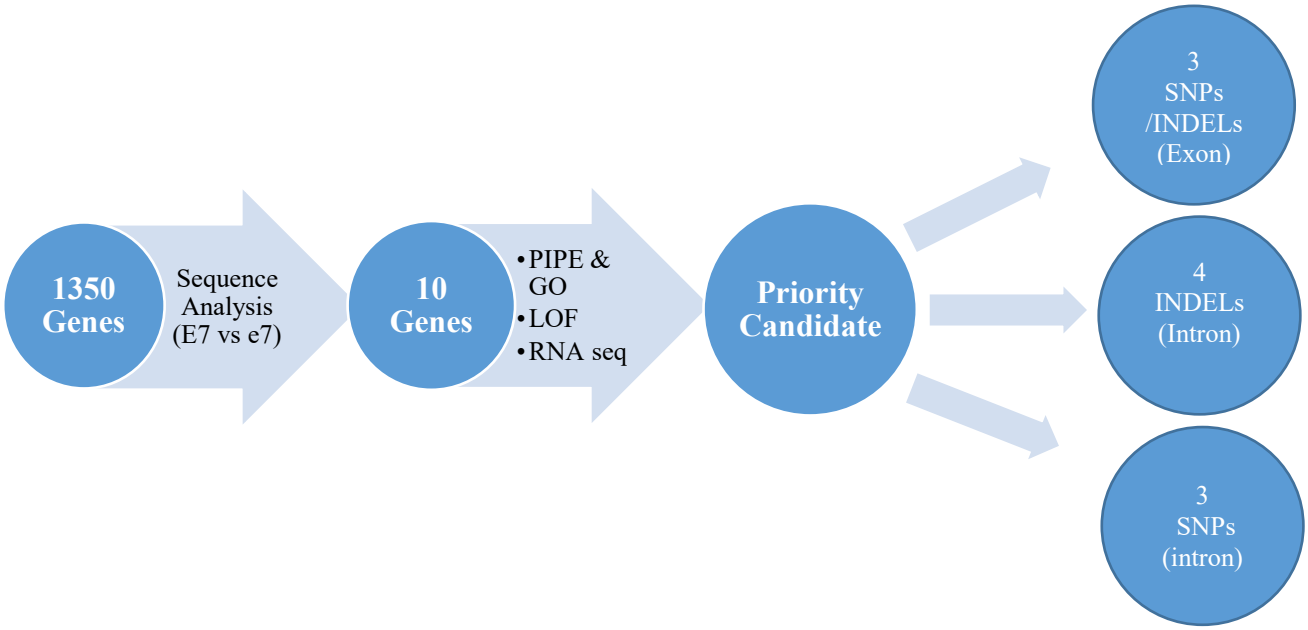


Figure 7. Process of identifying and prioritizing the 10 candidate genes in order of likeliness in being the underlying gene for the E7 maturity locus in soybean.

Table 2. Short-list of candidate genes selected by whole genome sequencing analysis, each with their annotation is Arabidopsis (AT) and soybean (Soy) and presence of LOF mutations, occurrence of SNPs and any available RNA-sequence data provided by database investigations (Appendix 2) [88] [94] [95].

Gene	**Annotation (AT)**	**Annotation (Soy)**	**LOF**	**SNP database**	**RNA *seq.* data**
Glyma.06G200400	exocyst subunit exo70 family protein B1	Exo70 exocyst complex subunit	no	no	no
Glyma.06G200800	Nodulin MtN3 family protein	Glycosyl hydrolases family 16 Sugar efflux transporter for intercellular exchange	yes	no	no
Glyma.06G220000	AP2/B3-like transcriptional factor family protein	B3 DNA binding domain	no	no	no
Glyma.06G180300	Protein of Unknown Function	Domain of unknown function	yes	no	yes
Glyma.06G199800	SNF2 domain-containing protein / helicase domain-containing protein / HNH endonuclease domain-containing protein	Helicase conserved C-terminal domain HNH endonuclease	no	no	yes
Glyma.06G200200	Nodulin MtN3 family protein	Sugar efflux transporter for intercellular exchange	yes	no	no
Glyma.06G202300	Cytochrome P450 superfamily protein	Cytochrome P450	yes	no	yes
Glyma.06G233300	PHD finger family protein / bromo-adjacent homology (BAH) domain-containing protein	PHD-finger BAH domain	no	no	yes
Glyma.06G239100	Nuclear pore complex protein	protein binding	no	yes (intron)	no
Glyma.06G242200	WRKY family transcription factor family protein	WRKY DNA -binding domain	no	no	no

3.2 Sequencing data analysis for contrasting E7 lines

The sequencing results provided by GQ allowed for an efficient manner to identify all NVs between the E7 and e7 lines on chromosome 6 (Appendix 5). The 1350 genes originally identified within the E7 region was narrowed to 10 candidate genes. Among the candidates identified, 3 consisted of a NVs within the exon region, *Glyma.06G200400*, *Glyma.06g200800* and *Glyma.06G220000* in the form of SNPs or INDELs that resulted in amino acid changes (mutations) (Table 3), (Appendix 1). While the remaining 7 candidates consisted of NVs within the intron region, including four that consisted of INDELs and SNPs that could potentially affect the splicing process, *Glyma.06G199800*, *Glyma.06G233300, Glyma.06G239100* and *Glyma.06G242200*, and three that only consisted of SNPs, *Glyma.06G180300*, *Glyma.06G202000*, *Glyma.06G202300.*

Insight into an unknown proteins' function can be made by investigating its sequence, and its similarity to other proteins with known functions. These sequence similarities among many proteins that share the same or similar function are referred to as their conserved domain. Analyzing the candidate genes for the presence of conserved domains, can potentially provide insight in the function and involvement of the gene in time to flowering and maturity. To determine if the candidates or the variations reside in important functional domains, the candidates were searched for conserved domains using NCBI's conserved domain search engine (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Based on the search, only one candidate, *Glyma.06G200400* consisted of amino acid substitutions within its respective functional domains that included Phospholipid-translocating P-type ATPase and Cytochrome b super family, respectively (Table 4).

Table 3. Summary of sequence, and type of variations (SNP, INDEL) identified from GQ data, location within exon or intron, along with amino acid changes resulting from exon variations, and intron variations in the form of INDELs.

Gene	**Intron Variation**	**Exon Variation**	**Amino Acid Change/Sequence Change (E7>e7)**
Glyma.06G180300	1 SNP	N/A	N/A
Glyma.06G199800	1 INS	N/A	N/A
Glyma.06G200200	1 SNP	N/A	N/A
Glyma.06G200400	N/A	1 SNP	THR>SER
Glyma.06G200800	N/A	1 SNP, 1INS	VAL > GLY THR ARG, ARG, GLU,THR,LEU,ARG,LEU,THR,HIS,GLY ,ALA,ARG > DELETED
Glyma.06G202300	2 SNP	N/A	N/A
Glyma.06G220000	3 SNP	2 SNP	VAL>PHE ALA > GLY
Glyma.06G233300	4 SNP, 2 INS	N/A	N/A
Glyma.06G239100	1 INS, 3 SNP	N/A	N/A
Glyma.06G242200	7 SNP, 2 INS	N/A	N/A

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

Gene	**Conserved Domain**	**Size of domain (bp)**	**Variation within domain**
Glyma.06G180300	Neprosin	4445-4678 3617-3712 25-378 3021-3302 4827-4916	No
Glyma.06G199800	Ribosomal protein S2 C-terminal helicase domain of the DEAD-like helicases	9679-9747 2475-2585	No
Glyma.06G200200	PQ loop repeat Nodulin MtN21 Laminin G domain	952-1152 2719-2853 710-778	No
Glyma.06G200400	Cytochrome_b_N super family	1447-1836	THR > SER
Glyma.06G200800	Laminin G domain PQ loop repeat	4-72 246-446	No
Glyma.06G202300	P450 super family	5988-6629 4644-5141 213-539	No
Glyma.06G220000	Plant-specific B3-DNA binding domain	2062-2319 772-999	No
Glyma.06G233300	Bromo Adjacent Homology domain	5225-5551	No
Glyma.06G239100	No conserved domain	N/A	N/A
Glyma.06G242200	The WRKY DNA-binding domain	3251-3361 552-617	No

3.3 Computational analysis of candidate genes

The candidates identified via sequence analysis, were determined based on NVs identified to be consistently different between the E7/e7 genotypes. Among these NVs, included those found within the coding region that resulted in an amino acid mutation, as well those found within the non-coding region that may potentially have an effect on intron splicing and should therefore not be eliminated from discussion (Table 3). To further support the involvement of candidate genes in time of flowering and maturity, -omics investigation was used, including determining the PIPE

interacting partners, paired with GO and literature. The PIPE data was analyzed using a concept referred to as *guilt by association*, where if the gene of interest is found to interact with other genes known to be involved in time of flowering, then it is most likely that that gene is also involved in time of flowering. The PIPE score for each candidate was determined by averaging the PIPE score of all the top interacting partners with PIPE scores (>0.5) with a maximum of 200 interacting partners (Appendix Table 6). The GO for these top interacting partners was determined using the "GO Term Enrichment Tool" on www.soybase.org, with an emphasis placed on interacting partners involved in processes related to time of flowering and maturity, including embryo development, flowering, etc. (Appendix Figure 1). Candidates identified as having higher total interactions, are hypothesized to have a greater chance to be involved in time of flowering and maturity.

Taking into consideration the evidence provided by GQ data (Table 3), and the databases for LOF, SNP, and RNA seq. (Appendix Table 2), conserved domains (Table 4) and PIPE analysis, were used to rank the candidates, based on most to least likely being the underlying gene for the E7 locus (Table 5). The candidates were ranked into three subgroups, (1) NVs found in the exon region, (2) NVs in the form of INDELs found in the intron region, and (3) NVs in the form of SNPs found in the intron region.

Table 5. Gene candidates ranked on priority-based sequence variations between E7/e7 genotypes, with additional evidence provided from computational analysis using PIPE paired with GO. The biological processes for the top 200 interacting partners for each gene identified by PIPE was determined, with interacting partners related to biological processes known to be involved in time of flowering and maturity reported as total interactions. The candidates were ranked, as follows (1) – NVs, in the form of SNPs and INDELs found within the exon, (2) – NVs, in the form of INDELs found within the intron, and (3) – NVs, in the form of SNPs found within the intron.

Rank	**Gene**	**Total interactions (GO)**
1	*Glyma.06G200400*	23
	Glyma.06G200800	12
	Glyma.06G220000	32
2	*Glyma.06G199800*	17
	Glyma.06G233300	3
	Glyma.06G239100	N/A
	Glyma.06G242200	19
3	*Glyma.06G180300*	16
	Glyma.06G200200	N/A
	Glyma.06G202300	15

3.5 Expression analysis with RT-qPCR and dPCR

While mRNA transcript level analysis of all 10 candidates will be performed, greater emphasis has been placed on candidates identified as having NVs found within the exon region, including: *Glyma.06G200400*, *Glyma.06G200800* and *Glyma.06G220000*. Followed by a second group of candidates identified as having NVs in the form of INDELs in the intron region: *Glyma.06G233300*, *Glyma.06G239100*, and *Glyma.06G242200*. Considering the E7 locus was previously identified as controlling photoperiod response [57], it is hypothesized that the underlying gene for the E7 locus should exhibit difference in gene expression among the samples being grown in SD and LD conditions. To further determine the expression differences, samples were also analyzed at various developmental stages including before, during and after flowering. Among the 10 candidates, thus far expression analysis for 4 have been conducted, including

Glyma.06G199800 (Figure 8A), *Glyma.06G180300* (Figure 8B), *Glyma.06G233300* (Figure 8C), and *Glyma.06G239100* (Figure 8D).

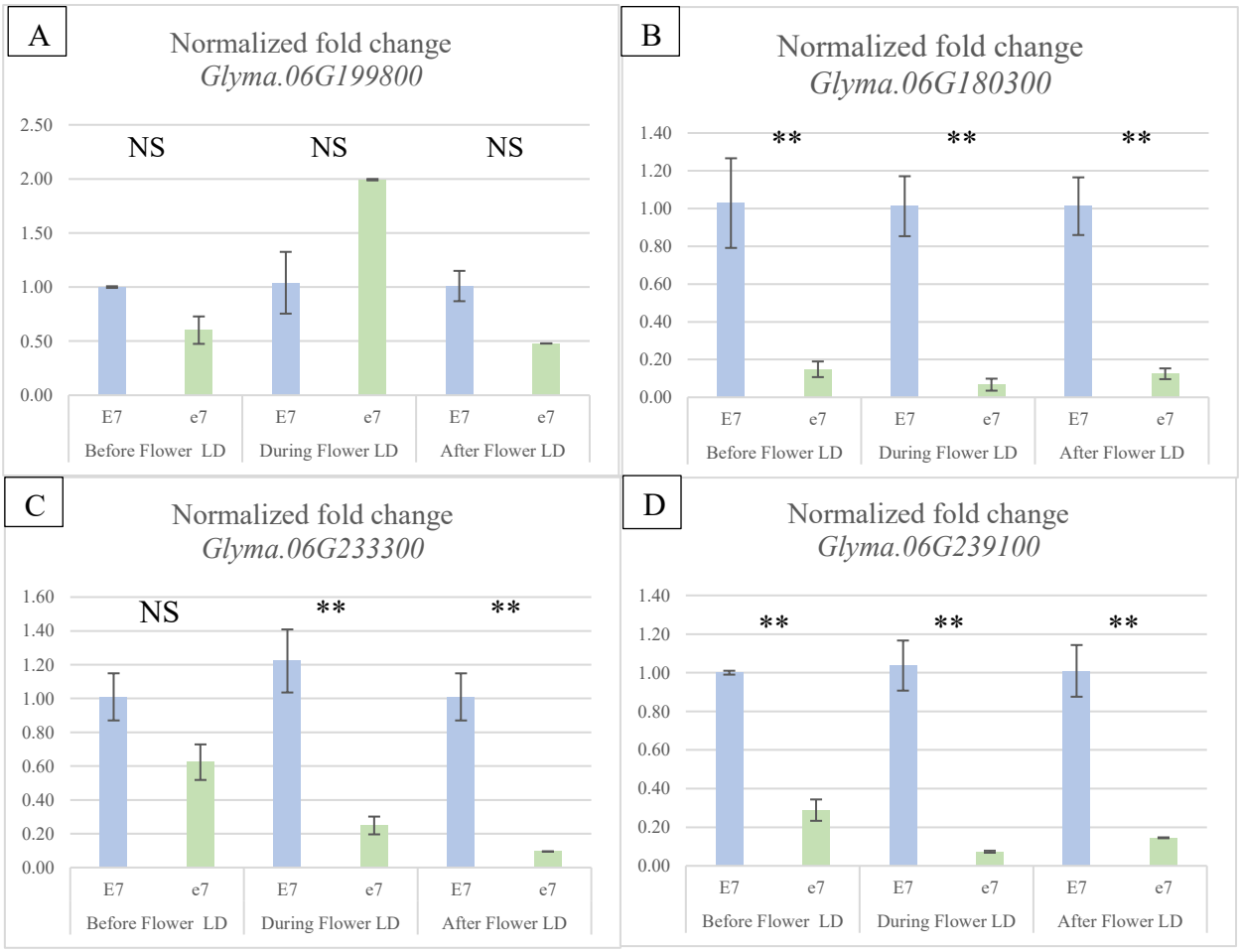


Figure 8. Relative (normalized) fold change for 4 of the candidate genes across 3 different conditions. All experiments were performed in triplicate (three biological replicates), using Pfaffl method with TUBB4 housekeeping gene. A. B. C. and D. represent the normalized fold change for *Glyma.06G199800*, *Glyma.06G180300, Glyma.06G233300* and *Glyma.06G239100,* respectively. P-value ≤ 0.05*, P-value ≤ 0.01 **, and P-value>0.05 (NS). *Glyma.06G180300*, and *Glyma.06G239100* has the most significant change between the E7/e7 genotypes across all three developmental stages, while *Glyma.06G233300* had the most significant change during and after flowering.

3.6 Candidate gene summary

Glyma.06G220000 is a 2373 bp gene with a transcript size of 921 bp, spanning four exons and 3 introns (Figure 9A). *Glyma.06G220000* is annotated as REDUCED VERNALIZATION RESPONSE 1 (VRN1) in Arabidopsis, playing a key role in vernalization. Figure 9B shows the

conserved plant-specific B3-DNA binding domain, identified to be conserved among many TFs. The TF that share the B3 domain include auxin response factors (ARF), LEAFY COTYLEDON2-ABI3-VAL (LAV), cold-responsive transcription factor RAV1, and REPRODUCTIVE MERISTEM (REM) families. The function of ARF and LAV families are linked to phytohormones, auxin and abscisic acid responsible for controlling development of various organs, and regulating seed maturation/germination respectively, in addition to being lined to stress response and vernalization [96, 97]. Sequence data of *Glyma.06G220000* shows several NVs, including 2 SNPs in the exon region, resulting in a change from (valine > phenylalanine), and (alanine >glycine) (Table 3). To execute its function, RNA is able to form stable two-dimensional (2D), then to three-dimensional (3D) structures by folding back on themselves [69] [70]. These structures are key to biological functions, as key interacting partners will target specific regions (with specific structure) to execute their function. NVs, in the form of SNPs and INDELs are known to have a potential effect on RNA folding and biological function. To determine whether these amino acid substitutions present within the coding region had an effect on the 2D RNA folding, the 2D RNA structure of the candidates with contrasting E7/e7 genotypes were analyzed. The difference between the genotypes was determined by comparing the structural differences and the minimum free energy (MFE) to determine major differences in the thermodynamic stability. Comparing the 2D RNA structure for *Glyma.06G220000*, it was determined that MFE for E7, and e7 was -493.80 kcal/mol, and -502.90kcal/mol, respectively, with significant structural changes seen between the wild and mutant alleles (Figure 10). In addition, *Glyma.06G220000* was found to have 32 total interactions involved in time of flowering and maturity based on PIPE data (Table 5). The LOF data predicted no mutations for this gene, and the RNA seq data shows no expression

of the gene across all tissue samples, with only minimal expression identified in tissue samples 10 days after flowering.

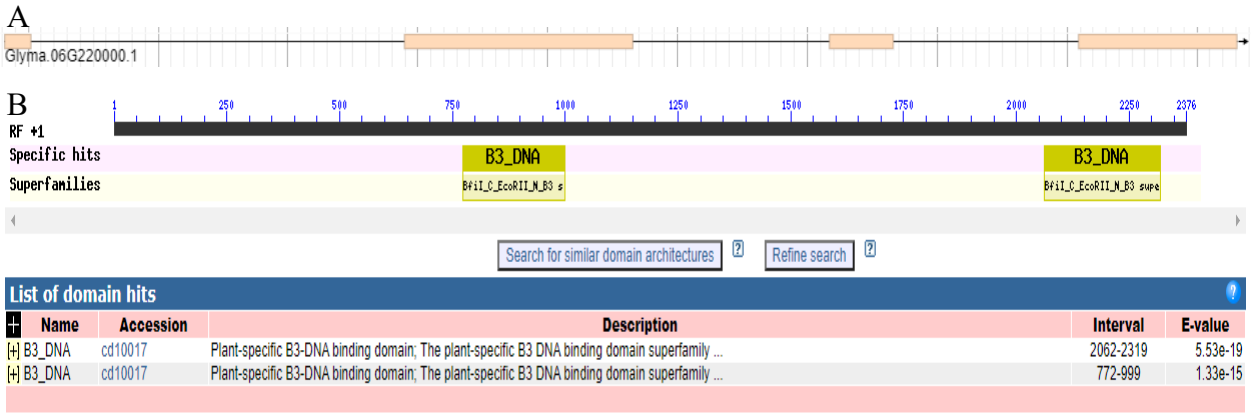


Figure 9. A) *Glyma.06G220000* intron (black spaces)/exon (beige) map. Figure taken from Phytozome.org B) plant-specific B3-DNA binding conserved domain within *Glyma.06G220000* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

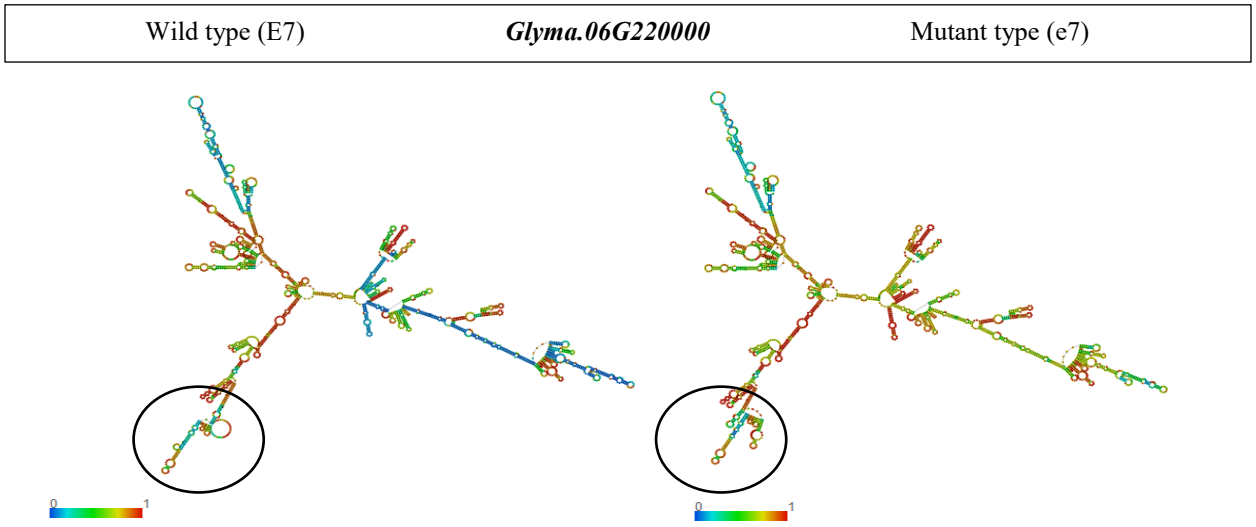


Figure 10. 2D RNA structure prediction of *Glyma.06G220000* generated using the RNAfold WebServer from Vienna University (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi). The MFE structure was generated with base-pairing probabilities, 0 (blue) and 1 (red). MFE for E7 calculated at -493.80 kcal/mol, and e7 calculated at -502.90kcal/mol. With major structural changes identified in the circle.

Glyma.06G200400 is a 1974 bp gene with a transcript size of 1974 bp, spanning one exon (Figure 11A). *Glyma.06G200400* is annotated as an EXOCYST SUBUNIT EXO70 FAMILY PROTEIN B1 (AtExo70B1) in Arabidopsis. Figure 11B shows the location of an Exo70 and a cytochrome B super family. The Exo70 is only one subunit of the exocysts complex, involved in

cell polarity, regulation of actin polarity and transport of exocystic vesicles [98]. The cytochrome B superfamily is involved in the biosynthesis of primary and secondary metabolites, including the synthesis and catabolism of plant growth regulators and signaling molecules [99]. Sequencing data of *Glyma.06G200400* shows 1 SNP (T>A) within the coding region, resulting in an amino acid substitution (threonine > serine) (Table 3). *Glyma.06G200400* did not have any major 2D structural changes (Figure 12), with the MFE for E7 and e7 identified as 425.65 kcal/mol and -421.95 kcal/mol, respectively. In addition, *Glyma.06G200400* had 23 total interacting partners involved in time of flowering and maturity based on PIPE data (Table 5). There is no available RNA seq data , and no LOF mutations predicted for this gene.

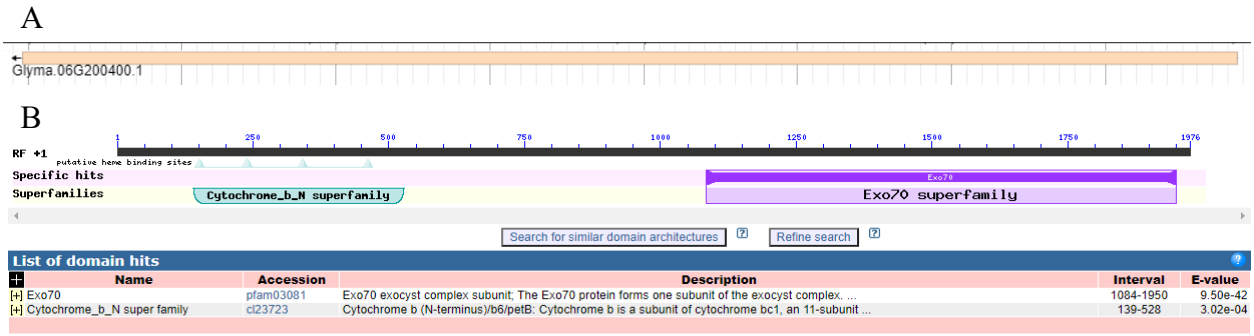


Figure 11. A) *Glyma.06G200400* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) Exo70 and Cytochromse b subunit conserved domains within *Glyma.06G200400* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

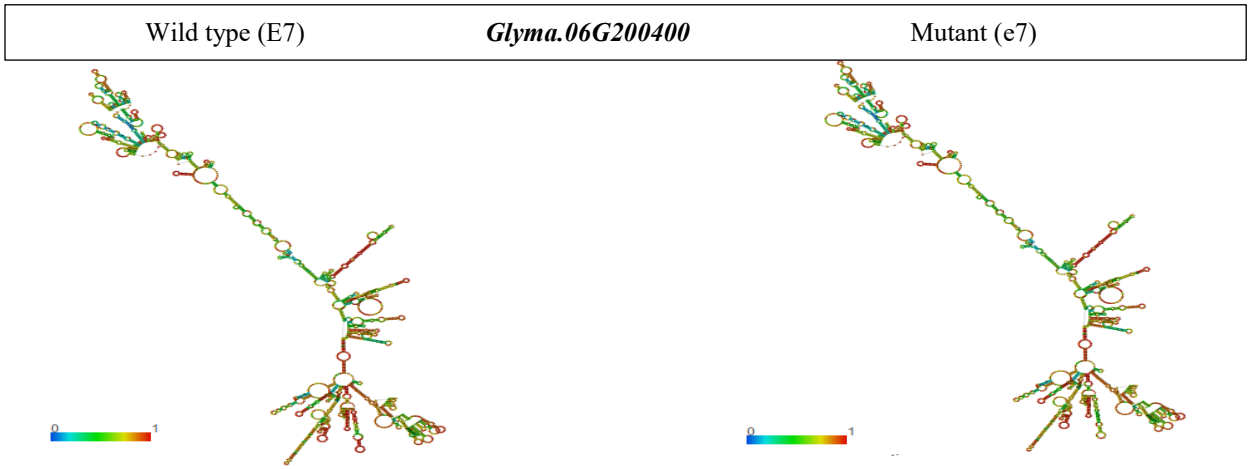


Figure 12. 2D RNA structure prediction of *Glyma.06G200400*, generated using the RNAfold WebServer from Vienna University (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi). The MFE structure was generated with base-pairing probabilities, 0 (blue) and 1 (red). MFE for E7 calculated at –425.65 kcal/mol, and e7 calculated at -421.95 kcal/mol.

Glyma.06G200800 is a 1583 bp gene with a transcript size of 555 bp, spanning 4 exons and 3 introns (Figure 13A). *Glyma.06G200800* is annotated as SWEET17 in Arabidopsis, a vacuolar fructose transporter [100]. (Figure 13B) shows the location of a Laminin Globular domain (LamG) and a PQ-super family. The LamG domains vary in their function, with roles in cell adhesion, signalling, and migration [101]. The PQ loop repeat family are membrane bound with a pair of repeats each spanning two transmembrane helices connected by a loop. Members of this family are believed to be transporters, as two members cystinosin and PQLC2 transport cysteine and cationic amino acid acids, respectively across the lysosomal membrane [102]. Sequencing data of *Glyma.06G200800* shows 1 SNP (A>C) and an INDEL, both in the coding region that result in a (valine > glycine) and a 13 amino acid deletion, respectively. With slight changes seen in the 2D RNA structure, correlated to the amount of variations present, with the MFE for E7 calculated at –1407.82 kcal/mol, and e7 calculated at -1400.17 kcal/mol (Figure 14). *Glyma.06G200800* was also predicted to have 12 interacting partners related to time of flowering and maturity according to PIPE data (Table 5). There is no RNA seq data available for this gene. There is however, a LOF mutant present within this gene (Table 2).

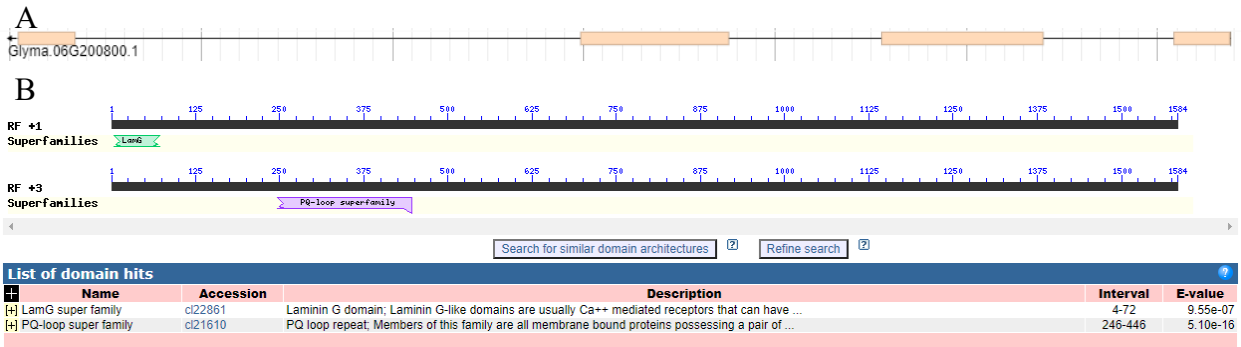


Figure 13. A) *Glyma.06G200800* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) Laminin G and PQ loop repeat conserved domains within Glyma.06G200800 in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021)

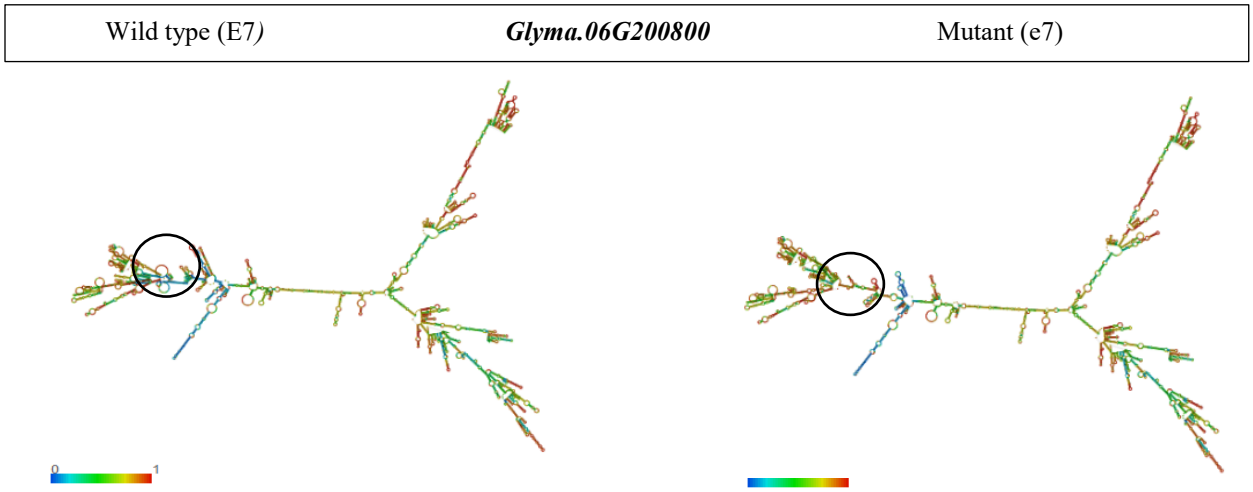


Figure 14. 2D RNA structure prediction of *Glyma.06G200800*, generated using the RNAfold WebServer from Vienna University (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi). The MFE structure was generated with base-pairing probabilities, 0 (blue) and 1 (red). MFE for E7 calculated at –1407.82 kcal/mol, and e7 calculated at -1400.17 kcal/mol. With major structural changes indicated within the circle.

Glyma.06G199800 is a 12101 bp gene with a transcript size of 3771 bp, spanning 16 exons and 15 introns (Figure 15A). *Glyma.06G199800* is annotated as an SNF2 domain-contain protein/helicase domain-containing protein/HNH endonuclease domain-containing protein in Arabidopsis, described as having a role in helicase activity, DNA, Adenosine triphosphate (ATP), and nucleic acid binding, and endonuclease activity [103]. (Figure 15B) shows the conserved Ribosomal protein S2 (RPS2) super family and Dead-like helicase C super family domains. The RPS2 superfamily is involved in the translation initiation complex, in Arabidopsis, RPS2 is thought to be involved in nucleotide triphosphate bindings, PPIs and have a function in disease resistance [104] [105]. The DEAD-box protein family all share the motif (Asp-Glu-Ala-Asp), and considered to be helicases, involved in transcription, pre-mRNA splicing, RNA transport, etc [106]. *Glyma.06G199800* was predicted to have 17 interacting partners involved in time of flowering and maturity based on PIPE data (Table 5). The RNA seq data has shown that it was expressed at a higer level in the young leaf and flowering tissue relative to all other tisse samples. There are no LOF mutations predicted for this gene. Expression analysis for *Glyma.06G199800* under LD conditions (Figure 8A) found there to be no significant change in expression between

the E7/e7 genotypes across the differing flowering periods, including before, during and after flowering.

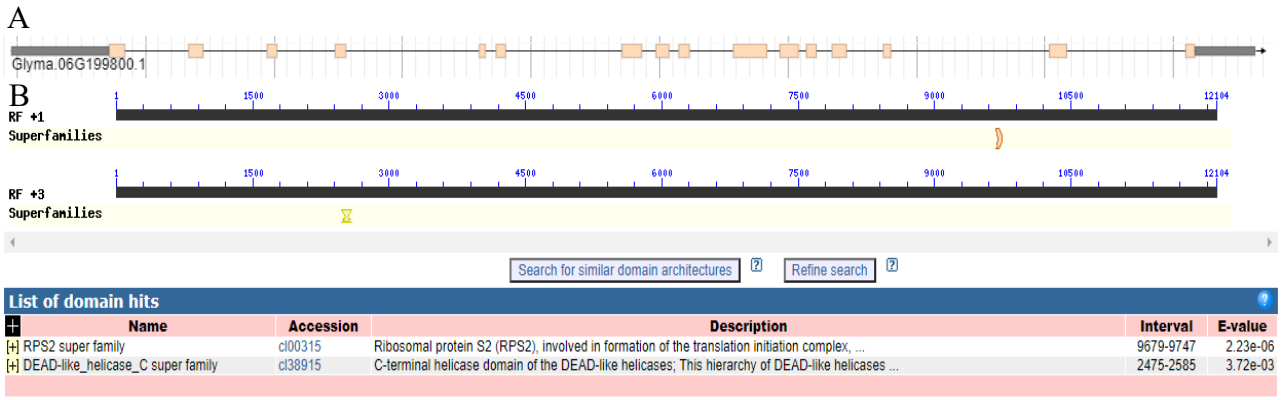


Figure 15. A) *Glyma.06G199800* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) Ribosomal protein S2 and the C-terminal helicase conserved domain within *Glyma.06G199800* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G233300 is a 5988 bp gene with a transcript size of 1153 bp, spanning 5 exons and 4 introns (Figure 16A). *Glyma.06G233300* is annotated as an EARLY BOLTING IN SHORT DAYS (EBS) gene in Arabidopsis, that encodes a chromatin remodelling factor that regulates time of flowering, specifically accelerating flowering during SD photoperiods [107]. (Figure 16B) shows the conservation of a Bromo-adjacent homology (BAH) super family domain within *Glyma.06G233300*. This BAH superfamily appears to be involved in DNA methylation, replication and transcriptional regulation events [108]. Sequence data of *Glyma.06g233300* includes 2 INS all within the intron region. *Glyma.06G233300* was predicted to have 3 interacting partners involved in time of flowering and maturity based on PIPE data (Table 5). The RNA seq data has shown that it was expressed at relatively high levels in the young leaf and flowering tissue relative to all other tisse samples. There was no LOF mutations predicted for this gene. Expression analysis during LD conditions (Figure 8C) identified significant expression change beftween E7 and e7 lines during and after flowering, with a normalized fold change of 1.22±0.09 and 0.91±0.10, respectively.

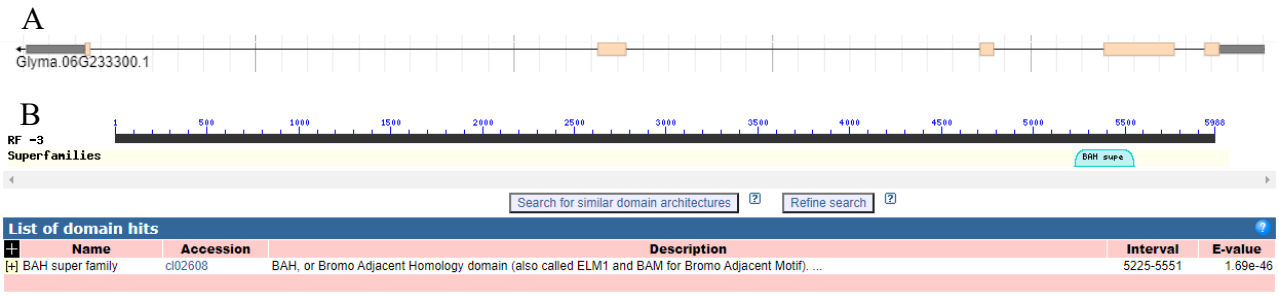


Figure 16. A) *Glyma.06G233300* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) Bromo Adjacent Homology conserved domain within *Glyma.06G233300* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G242200 is a 3736 bp gene with a transcript size of 639 bp, spanning 4 exons and 3 introns (Figure 17A). *Glyma.06G242200* is annotated as a member of the WRKY Transcription factor; Group I (WRKY20) in Arabidopsis. (Figure 17B) shows the location of a conserved WRKY DNA-binding domain. This domain consists of two motifs, including the WRKYGQK domain at the N-terminus, that the domain has been named after, and the zinc figure structure (C_2-H_2) or (C_2-HC) at their C-terminal [109]. These WRKY family members are categorized into three groups based on the number of WRKY domains and type of secondary motif either (C_2-H_2) or (C_2-HC) it possess. The WRKY family groups include, Group I that possesses two WRKY domains, Group II that posses a single WRKY domain and the C_2-H_2 motif, and Group III that possess a single WRKY domain and the C_2-HC motif [97]. This WRKY family of TF are believed to be involved in the stress response, both biotic and abiotic, including cold and hot temperature, drought, salinity and fungi, bacteria and nematodes. In addition the WRKY family of TF are reported to have a regulatory role in plant developmental and physiological processes, including plant growth, hormone signalling and leaf senescence [97] [109]. Sequence data of *Glyma.06G242200* showed 2 INDELs within the intron region. *Glyma.06G242200* was found to have 19 total interacting partners involved in time of flowering and maturity based on PIPE data.

The RNA seq data has shown this gene is not expressed across any tissue samples, with minimal expression 14 days after flowering. No LOF mutations were predicted for this gene.

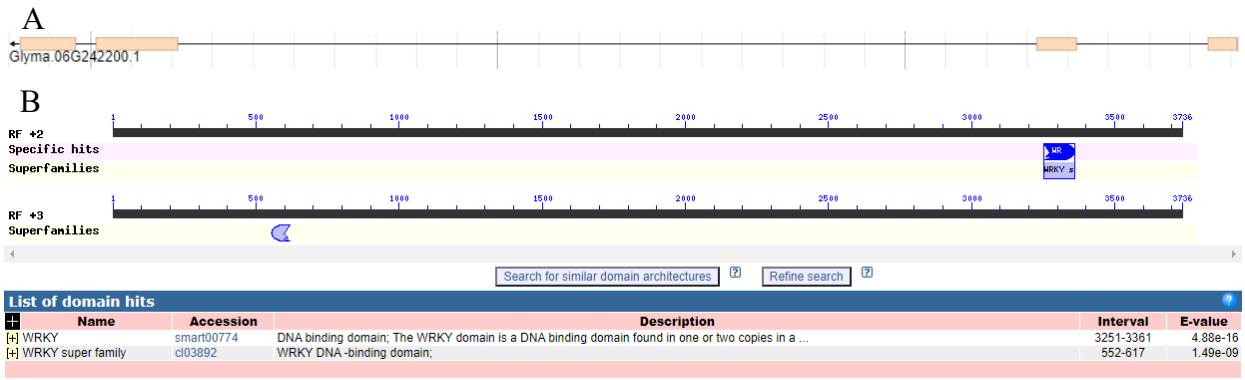


Figure 17. A) *Glyma.06G242200* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) WRKY DNA-binding conserved domains within *Glyma.06G242200* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G200200 is a 3149 bp gene with a transcript size of 795bp, spanning 7 exons and 6 introns (Figure 18A). *Glyma.06G200200* is annotated as a SWEET17 gene in Arabidopsis, a vacuolar fructose transporter [100]. (Figure 18B) shows the location of a PQ-loop, PLN00411 and LamG super family. The PQ-loop family of proteins belongs to the lysosomal cysteine transporter (LCT) family in the transporter/opsin/G protein-coupled receptor (TOG) superfamily. This family of proteins are all membrane bound, possessing a pair of PQ-loop repeats each spanning two transmembrane helices connected by a loop. It is believed that proteins that belong to the PQ-loop family act as membrane transporters, as members of this LCT family including cystinosin and PQLC2 transport cysteine and cationic amino acids respectively, across the lysosomal membrane [110]. The PLN00411 super family, described as the MtN21 family of proteins have been characterized in Arabidopsis, acting as a auxin export facilitator in vacuoles [111]. Lastly, the LamG superfamily, consists of the Laminin G domain that have Ca+++ mediated receptors. Proteins containing this domain are often found to have multipurpose roles in signal transduction, migration, etc. [112]. PIPE and GO data for *Glyma.06G200200* was inconclusive, therefore not

included in analysis. There is also no RNA seq data available for this gene. However, a LOF mutation was predicted (Table 2).

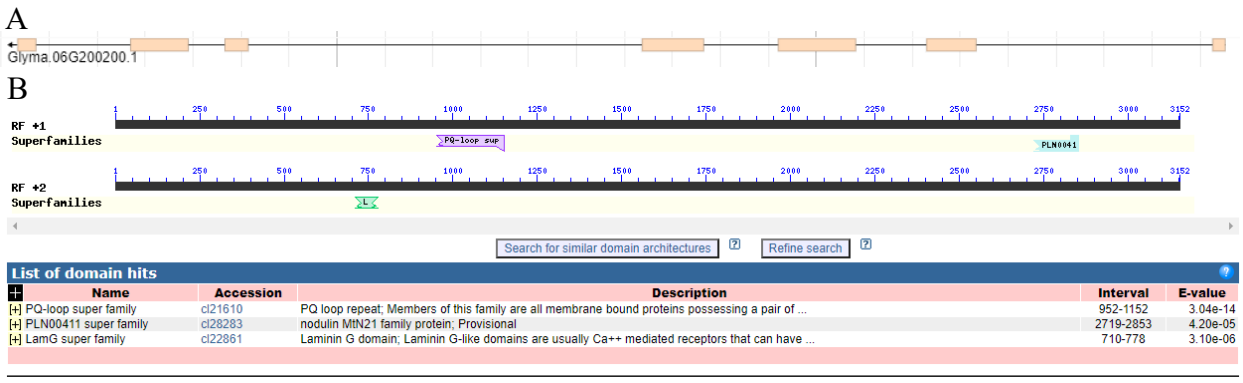


Figure 18. A) *Glyma.06G200200* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) PQ loop repeat, nodulin MtN21 family, and laminin G conserved domains within Glyma.06G200200 in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G180300 is a 5434 bp gene with a transcript size of 1377 bp, spanning 7 exons and 6 introns (Figure 19A). *Glyma.06G180300* is annotated as a LORD OF THE RINGS 1 gene in Arabidopsis, a protein of unknown function, it is however hypothesized to be involved Casparian strip formation [113]. Two conserved domains were identified, including Neprosin activation peptide, and Neprosin superfamily (Figure 19B). First identified in pitcher plants, this domain has been functionally annotated as a new class of prolyl endoprotease, which function in selectively cleaving after pro, ala motifs [114]. *Glyma.06G180300* is predicted to have 16 interacting partners involved in time of flowering and maturity based on PIPE data (Table 5). The RNA seq data showed steady expression in the young leaf and flower, but gradual decrease as the number of days after flowering increased, with minimal expression in other leaf tissue. There was LOF mutations predicted for this gene (Table 2). Expression analysis during LD conditions for *Glyma.06G180300* (Figure 8B) identified significant expression change beftween E7 and e7 lines before, during and after flowering that included a 0.88±0.14, 0.95±0.09, and 0.89±0.09 normalized fold change, respectively.

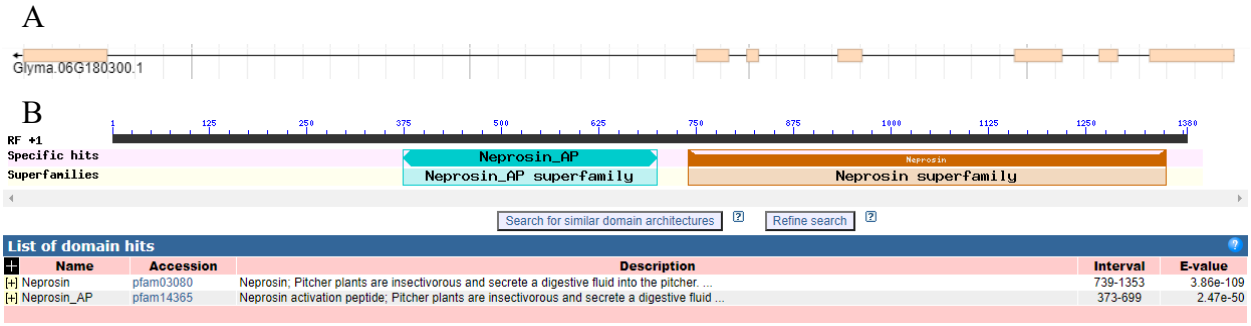


Figure 19. A) *Glyma.06G180300* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) Neprosin and Neprosin_AP conserved domains within *Glyma.06G180300* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G202300 is 6921 bp gene with a transcript size of 1914 bp, spanning 3 exons and 2 introns (Figure 20A). *Glyma.06G202300* is annotated as a Cytochrome P450 gene in Arabidopsis, regulated by the circadian clock pathway and involved in the synthesis of secondary compounds, including lignin, defense compounds, hormones and signaling molecules [115]. It has a conserved p450 superfamily domain (Figure 20B). The cytochrome P450 enzymes are a superfamily of haem-containing mono-oxygenases. While found in all kingdoms, in plants they play a role in the biosynthesis of compounds including hormones, defensive compounds and fatty acids [116]. *Glyma.06G202300* is predicted to have 15 interacting partners related to time of flowering and maturity based on PIPE data (Table 5). The RNA seq. data showed relatively high expression levels in the young leaf, followed by a significant drop in expression during flower formation, continued with an overall steady decline in expression as the days after flowering increased, with minimal expression in the root or nod. There was LOF mutations predicted for this gene (Table 2).

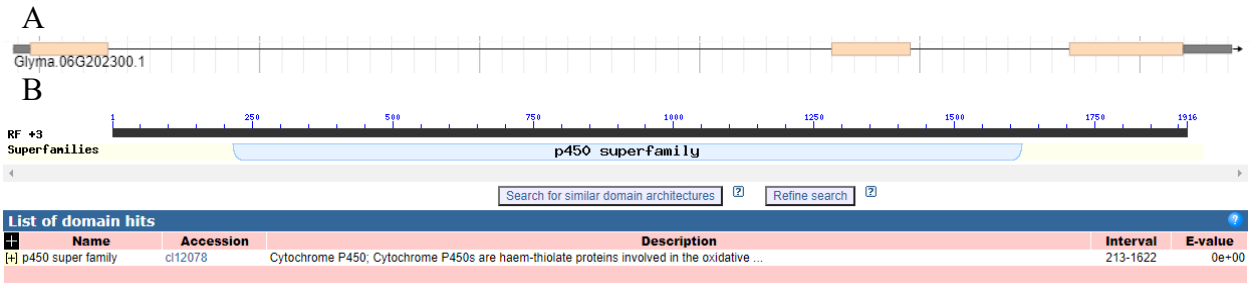


Figure 20. A) *Glyma.06G202300* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) p450 conserved domains within *Glyma.06G202300* in respect to the whole gene. Adopted from NCBI's conserved domain search (August 2021).

Glyma.06G239100 is 46755 bp gene with a transcript size of 5843 bp, spanning 17 exons and 16 introns (Figure 21A). *Glyma.06G239100* is annotated as an EMBRYO DEFECTIVE 1011 (emb1011) gene in Arabidopsis. EMB genes have found to have diverse functions, encoding components of protein complexes, including but not limited to protein import complexes, iron-sulfur clusters, and nuclear pore complexes. EMB genes have also been involved in histidine biosynthesis and chloroplast protein import motor [117]. There are no conserved domains identified for this gene. The PIPE and GO data for Glyma.06G239100 were found to be inconclusive. There are also no LOF mutations predicted for this gene (Table 2). Expression analysis of *Glyma.06G239100* during LD conditions (Figure 8D) identified significant expression change beftween E7 and e7 lines before, during and after flowering that included a 0.71±0.03, 0.96±0.09, and 0.86±0.09 normalized fold change, respectively.

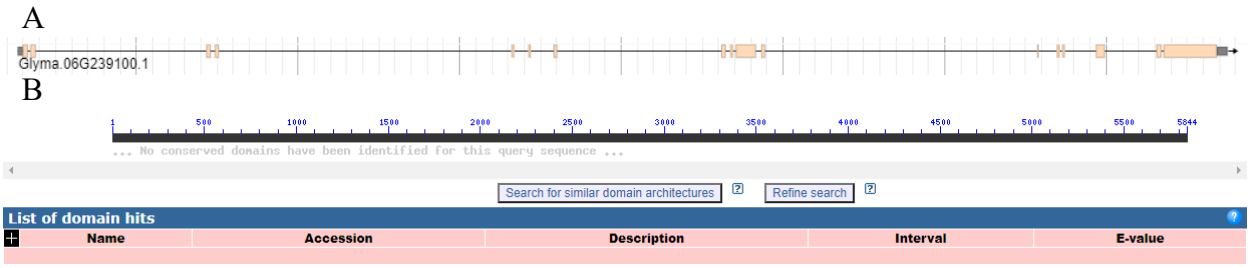


Figure 21. A) *Glyma.06G239100* intron (black spaces)/exon (beige) map with UTR's (grey). Figure taken from Phytozome.org B) no conserved domains were identified. Adopted from NCBI's conserved domain search (August 2021).

Chapter 4: Discussion

The expansion of soybean across Western and Northern Canada is strongly dependent on the development of early and ultra-early maturing cultivars that are better suited for the LD and shorter growing seasons of the region. To date, 10 maturity loci have been identified to be associated with time of flowering and maturity in soybean, while the molecular basis for E7, E8 and E11 are still unknown. To support breeding programs in developing suitable cultivars better adapted across Canada, this study was focused on identifying the underlying gene for the E7 locus by using next-generating sequencing, and functional genomics resources including a computational approach utilizing PIPE.

4.1 Analysis of candidate genes

The E7 region on chromosome 6 was previously narrowed 1350 genes. The E7 lines (Harosoy, OT89-5) and e7 lines (OT02-18, OT98-17) were sequenced by GQ, with the findings able to eliminate all but 10 potential candidates, as the rest did not contain NVs among the contrasting genotypes (E7/e7). In addition to sequencing analysis, candidates were assessed using computational analysis and literature to support their involvement in time of flowering and maturity, the 10 candidates were placed into 3-subgroups (Table 5). Ultimately, 3 candidates were ranked in subgroup 1, consisting of variations in the exon region: *Glyma.06G200400*, *Glyma.06G200800* and *Glyma.06G220000*. Four candidates were ranked in subgroup 2, consisting of variations in the form of INDELs in the intron region; *Glyma.06G199800, Glyma.06G233300*, *Glyma.06G239100* and *Glyma.06G242200*. Three candidates were ranked in sub-group 3, consisting of variations in the form of SNPs in the intron region; *Glyma.06G200200*, *Glyma.06G180300* and *Glyma.06G202300*. The placement of the 10 candidates in the 3 sub-groups was supported using PIPE paired with GO, candidates within sub-group 1 were found to

consist the most interacting partners that were involved in biological processes related to time of flowering and maturity (Table 5). The 2-D RNA structure of the mutant and wild lines further supported the involvement of the candidate genes found within sub-group 1 as potentially being involved in time of flowering and maturity. The NVs found within the exon region for *Glyma.06G220000* (Figure 10) and *Glyma.06G200800* (Figure 14) both resulted in significant structural variations that could result in altered function and lead to the differing time to flowering between genotypes. Further supporting the involvement of the candidates identified in time of flowering and maturity, the mRNA transcript level analysis for the candidates assessed thus far have shown significant changes in expression between the E7 and e7 for sub-group 2 candidates; Glyma*.06G233300*, *Glyma.06G239100* and sub-group 3 candidate *Glyma.06G239100* (Figure 10).

4.2 REDUCED VERNALIZATION RESPONSE 1 (VRN1) gene - (sub-group 1)

Through sequencing analysis, one of the genes identified to contrast in E7/e7 genotype was *Glyma.06G220000*, consisting of 2 SNPs within the exon region, resulting in a valine > phenylalanine, and an alanine > glycine (Table 3). It also consists of a conserved nuclear localized B3 DNA binding domain, and belongs to the AP2/B3-like transcriptional factor family (Figure 9B). It's homolog in Arabidopsis is the REDUCED VERNALIZATION RESPONSE 1 (VRN1) gene, that has previously been determined to be involved in vernalization, a process that is known to accelerate flowering in response to low temperatures [118] [119] [120].

While soybean does not have a vernalization requirement, temperature plays an important role in time of flowering and maturity, with cases reported of it both delaying as well as accelerating flowering [121] [122]. The detailed mechanism of how temperature influences time of flowering is still not fully understood [121]. Comparative analysis of the soybean and

Arabidopsis genome have found that they share similar flowering pathways, including genes involved in the vernalization pathway, despite not needing low temperatures for floral initiation [121]. The genes in the vernalization pathway in *Arabidopsis Thaliana*, includes FRIGIDA (FRI) and VERNALIZATION INSENSTIVIE 3 (VIN3) that regulate the expression of FLOWERING LOCUS C (FLC) a major gene that regulates vernalization requirements and response. In Arabidopsis, two flowering genes including FT and SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1) are positive regulators of flowering, their expression is thought to be regulated by the FLC target, with the decrease in expression correlated to later time of flowering phenotype. While the function of FLC has been conserved in some species, with the consistent function in acting as a floral repressor and being downregulated in response to cold. There are genes in other species, including wheat and barley that are responsible for vernalization instead of FLC, including VERNALIZATION 1 and 2 (VRN1 and VRN2) that are homologous to Arabidopsis APETALA 1 (AP1) and FT [123] [119].

Expression analysis of VRN1 in Arabidopsis has determined that it causes early flowering. In contrast *vrn1* alleles were found to not influence time of flowering but reduce vernalization response [120]. Including *vrn1-1* (TGG > TGA) and *vrn1-2* (1 bp deletion), both resulting in premature stop codons. Expression and phenotype analysis of another Arabidopsis VRN1 homolog, *Glyma.11G124200,* was responsive to photoperiod and low temperatures. Similar to VRN1, *Glyma.11G124200* was found to play a role in the vernalization pathway, in transgenic Arabidopsis the expression of DF and FLC decreased, while the expression of VIN3 and FT increased [121].

4.3 SUGARS WILL EVENTUALLY BE EXPORTED TRANSPORTER (SWEET16/17) - (sub-group 1)

Glyma.06G200800 (GmSWEET19) was also identified to contrast in E7/e7 genotype when sequenced, consisting of 1 SNP and 1 INDEL within the exon region, resulting in a valine > glycine and 10 amino acid deletions (Table 3). Consisting of conserved LamG and a PQ-super family domain, it's homolog in Arabidopsis is the SWEET16/17, a vacuolar fructose transporter (Figure 13B) [100].

Sugar compartmentation is among the factors that influence developmental processes including plant germination and development, as well as initiation of flowering. While SWEET proteins are known to be involved in the transport of sugars including sucrose, fructose and glucose, analysis of these SWEET genes in soybean have determined they are up-regulated during reproductive development, including seed and flower development [124]. The SWEET genes have not been explored in soybean to the extent seen in Arabidopsis and rice, however through whole-genome re-sequencing, *Glyma.06G200800* was identified as *GmSWEET19* and annotated as Sweet 16/17 in Arabidopsis [124]. Vacuolar sugar transport has been identified as playing a key role in a plants ability to cope with environmental stresses, including their tolerance to low temperatures [125].

4.4 EXOCYST SUBUNIT EXO70 FAMILY PROTEIN - (sub-group 1)

Glyma.06G200400 (*GmExo70J4*) was also identified to contrast in E7/e7 genotype when sequenced, consisting of 1 SNP within the exon region, resulting in a threonine > serine (Table 3). Figure 8b shows the location of an exocyst 70 (Exo70) and a cytochrome B super family, its homolog in Arabidopsis is the EXO70B1. In soybean, *Glyma.06G200400* was identified as a member of the novel legume-specific *Exo70J* subfamily, consisting of Golgi-targeting TM domains, and involved in vesicle trafficking [126].

The exocyst complex, in addition to consisting of the EXO70 protein, also consists of Sec3, Sec5, Sec6, Sec8, Sec10, Sec15, and Exo84. This complex has a function in exocytosis and cell-surface expansion. While proliferated greatly in plants, including 23 in Arabidopsis, in soybean *Exo70J* was phylogenetically distinct, only found in legumes, with 12 *GmExo70J* genes identified in soybean. Expression analysis of the *GmExo70J* genes identified that in addition to being involved in exocytosis during active cell growth, as is observed with *AtExo70* genes, the *GmExo70J* proteins are involved in a range of reproductive processes, based on its diverse expression pattern during floral organ and seed development. *GmExo70J* is also involved in the transport of nutrients and other biological processes, due to the genes expression observed in mature leaves, roots and floral organs [127].

4.5 EARLY BOLTING IN SHORT DAYS (EBS) – (sub-group 2)

Through sequencing analysis of contrasting E7/e7 genotypes, *Glyma.06G233300* was identified as consisting of 2 INDELs that resulted in a 14 amino acid deletion within the exon. *Glyma.06G233300* was identified has having a plant homeodomain *(*PHD) finger family protein/ bromo-adjacent homology (BAH) domain-containing protein (Figure 16B). Its homolog in Arabidopsis is the EBS gene that encodes a nuclear protein that contains a BAH and a PHD domain. Both motifs are thought to be involved PPIs and have been found in chromatin-mediated transcriptional regulators [128].

In plants, PHD domains are involved in the control of a variety of biological mechanisms including a number of developmental processes, including time of flowering. While floral-promotion pathways in Arabidopsis, including the photoperiod, autonomous and gibberellin pathway are known of, the plants ability to control when flowering should be initiated is strongly dependent on signalling that occurs between these regulator pathways. As part of the mechanisms

known to regulate flowering includes many chromatin remodelling factors, including PHD-containing proteins. Under non-inductive photoperiods (SD) flowering inhibition is strongly dependent on genes involved in light perception, including phytochrome B and signal transduction. Mutants defective of these genes have shown to reduce the photoperiodic inhibition of flowering. The presence of the novel EBS locus involved in regulating time of flowering was first identified, through screening of two early-flowering in short-days mutants, to show similar early-flowering phenotype. Further analysis determined the recessive mutation to be at a single locus responsible for an early-bolting phenotype specifically under short-days. EBS was confirmed to be involved in repressing flowering under SD conditions, as the *ebs* mutants were found to reduce the time of flowering by reducing the duration of the adult vegetative phase. EBS was also found to play a role in the repression of FT, required for floral initiation, as the double mutants *ebs ft* were found to have delayed time to flowering [107].

Through analyzing the double mutants *ebs ft*, and *ebs-1 fwa-1* it was suggested that EBS interacts with key floral initiators including FT and FWA under SD to delay flowering. This is corroborated by other studies found that FT overexpression alone is enough to promote flowering under SD and that FWA could act as a floral repressor downstream of EBS. It was also suggested that EBS and CO may regulate FT transcript levels independently, as the double mutant *ebs-2 co-2* had an intermediate-time of flowering phenotype.

4.6 WRKY family transcription factor family protein - (sub-group 2)

From sequencing analysis, *Glyma.06G242200* was identified to contrast in E7/e7 genotype with 2 INDELs within the intron region identified (Table 3). *Glyma.06G242200* is annotated as a member of group I WRKY transcription factor family, specially WRKY20. These transcription factor genes have been continuously proven to be involved in developmental and physiological

processes and various abiotic and biotic stress responses. The WRKY transcription factor family is named after the WRKYGQK conserved motif, containing either one or two WRKY domains that are comprised of 60 amino acids at the N-terminal end and a zinc-finger-like motif either C2H2 or C2HC at the C-terminal. The WRKY proteins regulate the expression of genes by recognizing and binding to the W-box (TTGACC/T) present in their promoters or non-W box sequences [129]. Expression analysis of the wild soybean WRKY20 in *Arabidopsis* has identified the gene to have a regulatory role in abscisic acid (ABA) signalling [130].

ABA is referred to as a stress-related hormone, playing a role in a plants growth and development in response to environmental cues, including drought, and salinity [131]. A correlation between drought stimuli, photoperiodic pathway and flowering were made, with ABA signalling playing a large part in mediating it [131]. While investigating the relationship between time of flowering and drought conditions, it was found that Arabidopsis is able to regulate time of flowering based on watering conditions, also referred to as drought escape (DE). Further screening of mutants impaired in DE found them to also be defective in the photoperiodic response. When exposed to water deficit conditions, crops grown under LD were found to activate DE and promote flowering. In contrast crops grown under SD were found to not cause DE, and in contrast to promoting flowering seen under LD conditions, under SD conditions it was found to delay flowering [131]. This supports that flowering is regulated by cues provided by drought signals and the photoperiodic pathway, further supported by the involvement of key genes included in the floral network, including FT, TWIN SISTER OF FT (TSF) whose expression increased in response to water deficit and only under LD. In addition, GI was also found to play a role in the DE under LD, conveying drought-derived cues upstream of FT in parallel to CO.

The interlinkage seen between drought stimuli, photoperiodic response and floral network is largely mediated by ABA. GI is a key gene involved in coordinating multiple regulatory mechanisms, it has also recently been identified to play a key regulatory role in ABA signalling. It was found that GI relates ABA signals to FT with minor contribution from CO, supported by genetic evidence where by impairing ABA signalling, and maintaining GI function, there was a decrease in FT and TSF accumulation, and moderately reduced CO levels [131]. While the mechanism behind GI-dependent florigen regulation promoted by ABA signalling is still not fully understood, it is known that the complex between GI and CO is necessary to activate FT and for DE to occur [131].

4.6 PHOTOPERIOD-INDEPENDENT EARLY FLOWERING1 (PIE1) - (sub-group 2)

Glyma.06G199800 encodes for an SNF2 domain-containing protein/ helicase domain-containing protein. In Arabidopsis, the PHOTOPERIOD-INDEPDENDENT EARLY FLOWERING1 (PIE1) gene encodes for a member of the SNF2 ATP-dependent, chromatin-remodelling proteins, that have been previously identified to be involved in time of flowering by regulating FLC expression.

In Arabidopsis accessions, allelic variations at FRIGIDA (FRI) and FLOWERING LOCUS C (FLC) determine if the crop is winter-annual or summer-annual. The main difference being that the summer-annual accessions have a non-functional *fri* allele, and thus lack the requirement for vernalization for flowering. FRI increases the transcript level of FLC, a novel protein identified to encode for a MADS-domain transcription factor. It has been identified to act as a floral repressor by down-regulating genes including SOC1 and FT that have been previously identified to promote flowering. In the winter-annual accessions, vernalization regulates the expression of FLC by FRI.

PIE1 has been reported to play a key regulatory role in the expression of FLC at levels that inhibit flowering through activating FRI. In addition to its regulatory role involved in the vernalization pathway, it is also found to play a role in the photoperiod pathway to some extent. The *pie1* mutants flowered earlier in LD and continuous light, compared to when grown in SD conditions [132]. While evidence supports the involvement of PIE1 in time of flowering in Arabidopsis, screening of soybean orthologs for Arabidopsis flowering genes had identified two PIE1, SNF2 domain-containing protein/helicase domain containing proteins in soybean to also be involved in flowering [133].

Chapter 5: Conclusion and future direction

5.1 Conclusion

Expanding soybean production across Northern regions of Western Canada requires the development of cultivars that are able to withstand environmental stimuli including abiotic and biotic stresses of the region. The adaptation of soybean to various geographic latitudes is strongly dependent on the transition from the vegetative to reproductive stage. Gaining a greater understanding of the genetics underlying key pathways, including time of flowering and maturity, can yield the development of ultra-early maturing soybean cultivars that are better suited for Western and Northern regions of Canada. Recent advancements in identifying novel genes involved in time of flowering and maturity have been made possible due to the availability of functional genomics approaches, including transcriptome studies (RNA-Seq), advancement in high throughput techniques such as NGS, and the increasing capabilities of computational tools, such as Protein-protein Interaction Prediction Engine (PIPE), and Scoring Protein INTeractions (SPRINT), to name a few [65]. In this research, the combination of sequence analysis, combined

with computational tools and functional genomics approaches were key in narrowing the E7 region to a manageable short-list of candidate genes.

Based on sequence analysis, among the genes within the E7 locus, 10 candidates were identified to contrast in E7/e7 genotype, with the SNPs and INDELs identified resulting in an amino acid variation, and/or open reading frame alterations. These candidates were further categorized into 3-subgroups, based on the sequencing analysis. That included those identified in having NVs in the form of SNPs and INDELs between the E7/e7 genotypes and present within the exon region (sub-group 1). Those identified in having INDELs present within the intron region (sub-group 2), and those identified as having NVs in the form of SNPs in the intron region (sub-group 3). The candidates categorized in sub-group 1, included *Glyma.06G200800* (*GmSWEET19*), *Glyma.06G200400* (*GmExo70J4*), and *Glyma.06G220000* (VRN1). In *Glyma.06G200800,* the one SNP and one INDEL resulted in an amino acid substitution, and 13 amino acid deletion, respectively. This variations resulted in major structural changes, with the MFE for E7 calculated at -1407.82 kcal/mol, and for e7 calculated at -1400.17 kcal/mol. *Glyma.06G200400* consisted of a single SNP, resulting in a single amino acid substitution that did not have any significant structural changes, with the MFE for E7 calculated at -425.65 kcal/mol, and for e7 calculated at -421.95 kcal/mol. Lastly, *Glyma.06G220000* consisted of 2 SNPs that resulted in two amino acid substitutions, with major structural changes identified, with the MFE for E7 calculated at -493.80 kcal/mol, and for e7 calculated at -502.90 kcal/mol. While expression analysis for these candidates continue to be performed to determine their involvement in time of flowering and maturity. Candidates in sub-group 2: *Glyma.06G199800, Glyma.06G233300*, *Glyma.06G239100*, and *Glyma.06G242200* and sub-group 3; *Glyma.06G180300*, *Glyma.06G200200*, and *Glyma.06G202300*, may also be likely candidates and will remain as potential candidates, in the

case those identified in sub-group 1 are identified as not the candidate through expression and compensation analysis.

5.2 Future direction

5.2.1 Expression analysis

The approach used to identify the short-list of candidate genes underlying the E7 locus (Figure 7) was successful at identifying candidates that were prioritized into sub-groups based on contrasting sequence between the E7/e7 genotypes, categorized as (sub-group 1) consisting of SNPs and INDELs within the exon region, (sub-group 2) consisting of INDELs within the intron region, and (sub-group 3) consisting of SNPs within the intron region. While currently in the process of performing expression analysis on all 10 candidate genes using RT-qPCR. Thus far, expression results have been obtained for *Glyma.06G199800*, *Glyma.06G233300*, *Glyma.06G239100*, and *Glyma.06G180300* among the contrasting E7 and e7 lines under LD conditions, during 3 developmental stages. In addition, dPCR will also be performed on tissue samples, as it has been proven to have greater sensitivity and will provide insight to any minor expression changes identified between the contrasting lines.

5.2.2 Vernalization experiment and photoperiod induction

Among the 5 major pathways involved in time of flowering that include photoperiod, vernalization, autonomous, ageing and gibberellin pathway, the two that are triggered by environmental cues such as light and temperature include photoperiod and vernalization. While light is known as one of the most important environmental cues that regulates time of flowering and maturity, temperature has also found to be an important environmental cue regulating time of flowering, via two ways, (1) some species must undergo a period of time where they are exposed

to cold temperatures in order to be capable of floral transition, also known as vernalization. (2) some winter annual, biennial, and perennial plants respond to physiological temperatures during the vegetative stage that are deemed non-stressful, with vernalization requirements ranging from (1 to 7°C).

Among the 10 candidates identified as potentially underlying the E7 locus, are 4 involved in time of flowering in response to environmental cues based on literature findings. Among the candidates identified include, sub-group 1 candidate *Glyma.06G220000* (VRN1), and sub-group 2 candidates *Glyma.06G233300* (EBS), *Glyma.06G199800* (PIE1), and *Glyma.06G2422000* (WRKY20). Among the 4 candidates, all were found to play a role in plant development in response to environmental cues that ranged from exposure to low temperatures, as well as various abiotic and biotic stress including drought and salinity. Recent studies found the expression of these respective Arabidopsis homologs to be responsive to photoperiod and temperature, influencing time of flowering. The photoperiod and vernalization pathways are considered separate when influencing time of flowering, with the vernalization pathway regulating the expression of a key gene FLC, that regulates FT and SOCI. Evidence supports that genes previously determined to play a pivotal role in regulating time of flowering through the vernalization pathway are also responsive to photoperiod [121]. While soybean is a non-vernalized plant, this supports that temperature plays an important role in time of flowering and maturity, as low temperatures could delay the time of flowering in some cultivars. The E7 maturity locus is believed to be involved in time of flowering and maturity, strongly influenced by the photoperiod pathway. However, evidence suggests that the underlying gene for the E7 locus may be responsive to photoperiod and vernalization. Therefore, to confirm the involvement of the candidate genes to time of flowering and maturity, and their function as the underlying gene for the E7 locus, the

E7/e7 lines should be grown under cold acclimation and control conditions in SD and LD photoperiods to determine if there is a difference in time of flowering and maturity in response to photoperiod and temperature. If there is a distinct difference in time of flowering and maturity, this would lead to evidence supporting that the underlying gene for the E7 locus is involved in both the photoperiod and vernalization pathway.

5.2.3 Compensation analysis

Compensation analysis will be performed in parallel to determine if the underlying gene for the E7 locus is responsive to photoperiod and vernalization through phenotypic analysis. The top candidates identified, will be further functionally characterized using Arabidopsis and/or soybean. This approach will follow similar to the transformation protocol previously used in our lab to transform E8 candidates into Arabidopsis, including initially amplifying transcript sequences of the candidate genes using Platinum® Pfx taq from ThermoFisher (cat no. 11708-013), and cloned into the Zero Blunt™ TOPO™ cloning vector from ThermoFisher (cat no. 451245). Using Blunt end restriction enzymes, the blunt-end PCR product is inserted into the vector, and transformed into chemically competent cells, using kanamycin to select successfully transformed cells.

Prior to transforming into Arabidopsis, the vectors will be sequenced to identify the presence of the respective NV for the respective candidate. Following transformation, the transformed mutant genotypes will be grown in inductive and non-inductive photoperiods to phenotypically assess time to flowering and maturity.

5.2.4 Allele-specific marker development (CAP/dCAP and KASP)

To allow for soybean breeding programs to accelerate their breeding of ultra-early cultivars suitable for Western and Northern regions of Canada, it is necessary for the development of allele-

specific markers as a diagnostic tool. Following the identification of the E7 and e7 alleles using sequencing and computational analysis, and further confirming the candidates' involvement in time of flowering and maturity using expression analysis and compensation validations, it is necessary to develop allele-specific markers. While traditional DNA marker-based approaches including RADP, RFLP, AFLP, and SSR, have been used in the past, an alternative, KASP method developed by LGC Genomics Ltd., is high throughput that uses SNP genotyping and requires minimal SNP markers to genotype various samples based on dual FRET (Fluorescent Resonance Energy Transfer) [134]. To develop allele-specific markers for the identified and validated E7 gene, KASP markers will be designed following the KASP manual, including the design of KASP primers and probes [135]. Another reliable allele-specific marker system, CAPS is cost effective, but is low throughput in comparison [136]. CAPS and derived CAPS (dCAPS) are PCR based markers that amplify genomic sequences around polymorphic endonuclease cleaving sites, with the fragments observed via agarose gel electrophoresis [137]. CAPS markers are used to detect polymorphisms that occur in restriction sites, in contrast to dCAPS markers that introduce restriction sites at a SNP and INDEL site using primers. The CAPS markers will be designed following Primer 3 (http://primer3.ut.ee), while the dCAPS markers will be designed using dCAPS finder 2.0 (http://helix.wustl.edu/dcaps/).

References

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

[2] United States Department of Agriculture, "World Agricultural Production," Foreign Agricultural Service, 2021.

[3] C. V. Montania, T. Fernandez-Nunez and M. A. Marquez, "The role of the leading exporters in the global soybean trade," *Agricultural Economics,* vol. 67, no. 7, pp. 277-285, 2021.

[4] Statistics Canada, "Seeding decisions harvest opportunties for Canadian farm operators," 2017.

[5] Agriculture and Agri-Food Canada, "CANADA: OUTLOOK FOR PRINCIPAL FIELD CROPS," 2021.

[6] Statistics Canada, "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," 2021.

[7] United States Department of Agriculture, "World Agriculture Production," 2019. [Online]. Available: https://apps.fas.usda.gov/psdonline/circulars/production.pdf.

[8] E. Dorff, "The soybean, agriculture's jack-of-all-trades, is gaining ground across Canada," *Canadian Agriculture at a Glance,* pp. 1-13, 26 October 2007.

[9] A. a. A.-F. Canada, Artist, *Canada: Soybean Production.* [Art]. Foreign Agricultural Service U.S. Department of Agriculture, 2018.

[10] L. C. Purcell, M. Salmeron and L. Ashlock, "Soybean Growth and Development," in *Arkansas Soybean Production Handbook*, University of Arkansas System Division of Agriculture, 2014.

[11] C. A. Knott and C. Lee, "Identifying Soybean Growth Stages," 2016.

[12] W. R. Fehr and C. E. Caviness, "Stages of soybean development," Experiment Station Publications, 1977.

[13] eKonomics , "Soybean Development and Growth Staging," [Online]. Available: https://nutrien-ekonomics.com/latest-fertilizer-research/soybean-development-and-growth-staging/.

[14] X. Liu, J.-a. Wu, H. Ren, Y. Qi, C. Li, J. Cao, X. Zhang, Z. Zhang, Z. Cai and J. Gai, "Genetic variation of world soybean maturity date and geographic distribution of maturity groups," *Breeding Science,* vol. 67, no. 3, pp. 221-232, 30 May 2017.

[15] G. Wolfgang and Y.-q. Charles 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. 8, 2017.

[16] P. Grassini, N. C. La Menza, J. I. R. Edreira, J. P. Monzon, F. A. Tenorio and J. E. Specht, "Soybean," in *Crop Physiology Case Histors for Major Crops*, Academic Press, 2021, pp. 282-319.

[17] G. S. Golembeski, H. A. Kinmonth-Schultz, Y. Hun Song and T. Imaizumi, "Photoperiodic flowering regulation in Arabidopsis thaliana," *Advances in Botanical Research,* pp. 1-14, 2014.

[18] Z. Xia, H. Zhai, B. Liu, F. Kong, X. Yuan, H. Wu, E. R. Cober and K. Harada, "Molecular identification of genes controlling flowering time,," *Plant Systematics and Evolution,* pp. 1217-1227, 2012.

[19] 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 ,* 2016.

[20] D. Thakare, S. Kumudini and R. R. Dinkins, "The alleles at the E1 locus impact the expression pattern of two soybean FT-like genes shown to induce flowering in Arabidopsis," *Planta,* vol. 234, pp. 933-943, 17 June 2011.

[21] W. Rensink and C. R. Buell, "Arabidopsis to Rice. Applying Knowledge from a Weed to Enhance Our Understanding of a Crop Species1," *Plant Physiology,* vol. 135, no. 2, pp. 622-629, 2004.

[22] K. H. Wolfe, M. Gouy, Y.-W. Yang, P. M. Sharp and W.-H. Li, "Date of the monocot-dicot divergence estimated from chloroplast," *PNAS,* vol. 86, no. 16, pp. 6201-6205, 1989.

[23] M. Johansson and D. Staiger, "Time to flower: interplay between photoperiod and the circadian clock," *Journal of Experimental Botany,* vol. 66, no. 3, pp. 719-730, 2015.

[24] T. Imaizumi, A. Kubota and J. S. Shim, "Circadian Clock and Photoperiodic Flowering in Arabidopsis: CONSTANS Is a Hub for Signal Integration," *Plant Physiology,* vol. 173, pp. 5-15, 2017.

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

[26] R. Hayama, S. Yokoi, S. Tamaki, M. Yano and Shimamoto, "Adaptation of photoperiodic control pathways produces short-day flowering in rice," *Nature,* vol. 422, pp. 719-722, 2003.

[27] E. J. Sedivy, F. Wu and Y. Hanzawa, "Soybean domestication: the origin, genetic architecture and molecular," *New Phytologist,* vol. 214, pp. 539-553, 2017.

[28] B. Valliyodan, S.-H. Lee and H. T. Nguyen, "Sequencing, Assembly, and Annotation of the Soybean Genome," in *The Soybean Genome*, 2017, pp. 73-82.

[29] M. Y. Kim, S. Lee, K. Van, T.-H. Kim, S.-C. Jeong, I.-Y. Choi, D.-S. Kim, Y.-S. Lee, D. Park, J. Ma, W.-Y. Kim, B.-C. Kim, S. Park, K.-A. Lee, D. H. Kim, K. H. Kim, J. H. Shin, Y. E. Jang, K. D. Kim, W. X. Liu, T. Chaisan, Y. J. Kang, Y.-H. Lee, K.-H. Kim, J.-K. Moon, J. Schmutz, S. A. Jackson, J. Bhak and S.-H. Lee, "Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 107, no. 51, pp. 22032-22037, 2010.

[30] P. B. Cregan, "Soybean Molecular Genetic Diversity," in *Genetics and Genomics of Soybean*, New York, Springer Science+Business Media, 2008, pp. 17-32.

[31] M. Y. Kim, K. Van, Y. J. Kang, K. H. Kim and S.-H. Lee, "Tracing soybean domestication history: From nucleotide to genome," *Breeding Science,* vol. 61, no. 5, pp. 445-452, 2012.

[32] J. A. O'Rourke, A. M. Graham and S. A. Whitham, "Soybean Functional Genomics: Bridging the Genotype-to-Phenotype Gap," in *Plant Pathology and Microbiology, Agronomy*, 2017, pp. 151-170.

[33] J. Schmutz, S. B. Cannon and S. A. Jackson, "Genome sequence of the palaeopolyploid soybean," *Nature,* vol. 463, pp. 178-183, 2010.

[34] E. Gaspersakaja and V. Kuchinskas, "The most common technologies and tools for functional genome analysis," *Acta medica Lituanica,* vol. 24, no. 1, pp. 1-11, 2017.

[35] R. L. Bernard, "Two Major Genes for Time of Flowering and Maturity in Soybeans," *Crop Science,* vol. 11, no. 2, 1971.

[36] 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.

[37] T. R. Copley, M.-O. Duceppe and L. S. O'Donoughue, "Identification of novel loci associated with maturity and yield traits in early maturity soybean plant introduction lines," *BMC Genomics,* vol. 19, no. 1, p. 167, 1 Marc 2018.

[38] B. Liu, S. Watanabe, T. Uchiyama, F. Kong, A. Kanazawa, Z. Xia, A. Nagamatsu, M. Arai, T. Yamada, K. Kitamura, C. Masuta, K. Harada and J. Abe, "The soybean stem growth habit gene Dt1 is an ortholog of arabidopsis TERMINAL FLOWER1," *Plant Physiology,* vol. 153, no. 1, pp. 198-210, 2010.

[39] 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,* vol. 68, no. 8, pp. 1873-1884, 2017.

[40] E. O. Tanaka, T. Shimizu, M. Hajika, A. Kaga and M. Ishimoto, "Highly multiplexed AmpliSeq technology identifies novel variation of flowering timerelated genes in soybean (Glycine max)," *DNA Research,* vol. 26, no. 3, pp. 243-260, 11 March 2019.

[41] R. Takeshima, H. Nan, K. Harigai, L. Dong, J. Zhu, S. Lu, M. Xu, N. Yamagishi, N. Yoshikawa, B. Liu, T. Yamada, F. Kong and J. Abe, "Functional divergence between soybean FLOWERING LOCUS T orthologues FT2a and FT5a in post-flowering stem growth," *Journal of Experimental Botany,* vol. 70, no. 15, pp. 3941-3953, 30 Apr 2019.

[42] J. L. Weller and R. Ortega, "Genetic control of flowering time in legumes," *frontiers in Plant Science,* 2015.

[43] H. Zhai, S. Lu, H. Wu, Y. Zhang, X. Zhang, J. Yang, Y. Wang, G. Yang, H. Qiu, T. Cui and Z. Xia, "Diurnal Expression Pattern, Allelic Variation, and Association Analysis Reveal Functional Features of the E1 Gene in Control of Photoperiodic Flowering in Soybean," *PLos ONE,* vol. 10, no. 8, 2015.

[44] 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," *PNAS,* vol. 109, no. 32, pp. 2155-2164, 2012.

[45] M. Xu, N. Yamagishi, C. Zhao, R. Takeshima, M. Kasai, S. Watanabe, A. Kanazawa, N. Yoshikawa, B. Liu, T. Yamada and J. Abe, "The Soybean-Specific Maturity Gene E1 Family of Floral Repressors Controls Night-Break Responses through Down-Regulation of FLOWERING LOCUS T Orthologs," *Plant Physiology,* vol. 168, pp. 1735-1746, 26 August 2015.

[46] J. Zhu, R. Takeshima, K. Harigai, M. Xu, F. Kong, B. Liu, A. Kanazawa, T. Yamada and J. Abe, "Loss of Function of the E1-Like-b Gene Associates With Early Flowering Under Long-Day Conditions in Soybean," *Frontiers in Plant Science ,* vol. 9, 08 January 2019.

[47] 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 the GIGANTEA Gene Is Involved in Soybean Maturity and Flowering," *Genetics Soceity of America,* June 2011.

[48] Y. Tsubokura, S. Watanabe, Z. Xia, H. Kanamori, H. Yamagata, A. Kaga, Y. Katayose, J. Abe, M. Ishimoto and K. Harada, "Natural variation in the genes responsible for maturity loci E1, E2, E3 and E4 in soybean," *Annals of Botany,* vol. 113, no. 3, pp. 429-441, 2014.

[49] H. D. Buzzell and R. I. Voldeng, "Inheritance of insensitivity to long daylength," *Soybean Genetics Newsletter,* vol. 7, no. 13, 1980.

[50] J. Miladinovic, M. Ceran, V. Dordevic, S. Balesevic-Tubic, K. Petrovic, V. Dukic and D. Miladinovic, "Allelic Variation and Distribution of the Major Maturity Genes in Different Soybean Collections," *frontiers in Plant Science,* 04 Sept 2018.

[51] 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. 91, 2017.

[52] 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.

[53] 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.

[54] A. Dissanayaka, T. M. 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.

[55] N. Nissan, E. R. Cober, M. Sadowski, M. Charette, A. Golshani and B. Samanfar , "Identifying new variation at the J locus, previously identified as e6, in long juvenile 'Paranagoiana' soybean," *Theoretical and Applied Genetics,* vol. 134, no. 4, pp. 1007-1014, 2021.

[56] E. Bonato and N. Vello, "E6, a dominant gene conditioning early flowering and maturity in soybeans," *Genetics and Molecular Biologu,* vol. 22, no. 2, pp. 229-232, 1999.

[57] E. Cober and H. Voldeng, "Low R : FR light quality delays flowering of E7E7 soybean lines," *Crop Science,* vol. 41, no. 6, pp. 1823-1826, 2001b.

[58] E. R. Cober and H. D. Voldeng, "A New Soybean Maturity and Photoperiod-Sensitivity Locus Linked to E1 and T," *Crop Science Society of America,* vol. 41, no. 3, pp. 698-701, 2001.

[59] 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.

[60] 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,* 2018.

[61] 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.

[62] 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.

[63] 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," *Crop Science,* vol. 54, no. 6, pp. 2529-2535, 2014.

[64] 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. 20, 2016.

[65] 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, Feb 2017.

[66] Z. Gizlice, T. Carter and J. Burton, "CROP BREEDING, GENETICS & CYTOLOGY," *Crop Science,* vol. 34, no. 5, pp. 1143-1151, 1994.

[67] V. C. Concibido, D. A. Lange, R. L. Denny, J. H. Orf and N. D. Young, "Genome Mapping of Soybean Cyst Nematode Resistance Genes in 'Peking', PI 90763, and PI 88788 Using DNA Markers," *Crop Science,* vol. 37, no. 1, pp. 258-264, 1997.

[68] K. S. Lewers, E. H. Crane, C. R. Bronson, J. M. Schupp, P. Keim and R. C. Shoemaker, "Detection of linked QTL for soybean brown stem rot resistance in 'BSR 101' as expressed in a growth chamber environment*," *Molecular Breeding,* vol. 5, pp. 32-42, 1999.

[69] M. A. Nadeem, M. A. Nawaz, M. Q. Shahid, Y. Dogan, G. Comertpay, M. Yildiz, R. Hatipoglu, F. Ahmad, A. Alsaleh, N. Labhane, H. Ozkan, 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, 2018.

[70] B. C. Collard and D. J. Mackill, "Collard BC, Mackill DJ.. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century," *Philosophical Transactions of The Royal Society B Biological Sciences,* vol. 363, no. 1491, pp. 557-572, 2008.

[71] U.S. National Library of Medicine , "Derived Cleaved Amplified Polymorphic Sequences (dCAPS)," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/probe/docs/techdcaps/. [Accessed 2019].

[72] C. He, J. Holme and J. Anthony, "SNP Genotyping: The KASP Assay," in *Crop Breeding. Methods in Molecular Biology (Methods and Protocols)*, vol. 1145, New York, Humana Press, 2014, pp. 75-86.

[73] H. Verdeprado, T. Kretzschmar, H. Begum, C. Raghavan, P. Joyce, P. Lakshmanan and B. C. Collard, "Association mapping in rice: basic concepts and perspectives for molecular breeding," *Plant Production Science,* vol. 21, no. 3, pp. 158-176, 2018.

[74] B. C. Collard, M. Z. Jahufer, J. B. Brouwer and E. C. Pang, "An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts," *Euphytica,* vol. 142, pp. 169-196, 2005.

[75] A. Korte and A. Farlow, "The advantages and limitations of trait analysis with GWAS: a review," *Plant Methods,* vol. 9, no. 29, 2013.

[76] U. K. S. Kushwaha, V. Mangal, A. K. Bairwa, S. Adhikari, T. Ahmed, P. Bhat, A. Yadav, N. Dhaka, D. R. Prajapati, A. Gaur, R. Tamta, I. Deo and N. K. Singh, "Association Mapping, Principles and Techniques," *Journal of Biological and Environmental Engineering,* vol. 2, no. 1, pp. 1-9, 2017.

[77] S. Challa and N. R. Neelapu, "Chapter 9 - Genome-Wide Association Studies (GWAS) for Abiotic Stress Tolerance in Plants," in *Biochemical, Physiological and Molecular Avenues for Combating Abiotic Stress Tolerance in Plants*, 2018, pp. 135-150.

[78] C. Caragea and V. Honavar, "Machine Learning in Computational Biology," 2008.

[79] A. Oulas, G. Minadakis, M. Zachariou, K. Sokratous, M. N. Bourdakou and G. M. Spyrou, "Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches," *Briefings in Bioinformatics,* vol. 20, no. 3, pp. 806-824, 2019.

[80] S. Pitre, M. Alamgir, J. R. Green, M. Dumontier, F. Dehne and A. Golshani, "Computational Methods For Predicting Protein–Protein Interactions," *Advances in Biochemical Engineering/Biotechnology,* vol. 111, pp. 247-267, 2008.

[81] A. X. Jones, Y. Cao, Y.-L. Tang, J.-H. Wang, Y.-H. Ding, H. Tan, Z.-L. Chen, R.-Q. Fang, J. Yin, R.-C. Chen, X. Zhu, Y. She, N. Huang, F. Shao, K. Ye, R.-X. Sun, S.-M. He, X. Lei and M.-Q. Dong, "Improving mass spectrometry analysis of protein structures with arginine-selective chemical cross-linkers," *Nature Communications,* vol. 10, no. 3911, 2019.

[82] Z. Ding and D. Kihara, "Computational identification of protein-protein interactions in model plant proteomes," *Scientific Reports,* 19 June 2019.

[83] K. Dick, A. Pattang, J. Hooker, N. Nissan, M. Sadowski, B. Barnes, L. H. Tan, D. Burnside, S. Phanse, H. Aoki, M. Babu, F. Dehne, A. Golshani, E. R. Cober, J. R. Green and B. Samanfar, "Human–Soybean Allergies: Elucidation of the Seed Proteome and Comprehensive Protein–Protein Interaction Prediction," *Journal of Proteome Research,* vol. 20, no. 11, pp. 4925-4927, 2021.

[84] K. Dick and J. R. Green, "Reciprocal Perspective for Improved Protein-Protein Interaction Prediction," *Scientific Reports,* vol. 8, no. 1, 2018.

[85] K. Dick, B. Samanfar, B. Barnes, E. R. Cober, B. Mimee, L. H. Tan, S. J. Molnar, K. K. Bigger, A. Golshani, F. Dehne and J. R. Green, "PIPE4: Fast PPI Predictor for Comprehensive Inter- and Cross-Species Interactomes," *Scientific Reports,* vol. 10, no. 1390, 2020.

[86] SoyCanada, "10 Million Acres of Opportunity - Planning for a decade of sustainable growth and innovation in the Canadian soybean industry," 2017.

[87] ThermoFisher Scientific, "BigDye™ Terminator v3.1 Cycle Sequencing Kit," Appliedbiosystems, 2016.

[88] D. Torkamaneh, J. Laroche, B. Valliyodan, L. O'Donoughue, E. Cober, I. Rajcan, R. Abdelnoor, A. Sreedasyam, J. Schmutz, H. T. Nguyen and F. Belzile, "Soybean Haplotype Map (GmHapMap): A Universal Resource for Soybean Translational and Functional Genomics," *Plant biotechnology journal,* vol. 19, no. 2, pp. 324-334, 2019.

[89] S. Humira, M. Bastien, E. Iquira, A. Tardivel, G. Legare, B. Boyle, E. Normandeau, J. Laroche, S. Larose , M. Jean and F. Belzile, "An Improved Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and Efficiency of SNP Discovery and Genotyping," *PLOS One,* vol. 8, no. 1, 2013.

[90] A. Schoenrock, F. K. Dehne, J. R. Green, A. Golshani and S. Pitre, "MP-PIPE: A Massively Parallel Protein-Protein Interaction Prediction Engine," pp. 327-337, 2011.

[91] 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, "PIPE: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs," *BMC Bioinformatics,* vol. 7, no. 365, 2006.

[92] M. W. Pfaffl, "Chapter 3.Quantification strategies in real-time PCR," in *A-Z of quantitative PCR*, La Jolla, 2004, pp. 87-112.

[93] P.-L. Quan, M. Sauzade and E. Brouzes, "dPCR: A Technology Review," *Sensors (Basel),* vol. 18, no. 4, p. 1271, 2018.

[94] D. Torkamaneh, J. Laroche, A. Tardivel, L. O'Donoughue, 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, 2018.

[95] 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.

[96] K. Yamasaki, T. Kigawa, M. Inoue, M. Tateno, T. Yamasaki, T. Yabuki, M. Aoki, E. Seki, T. Matsuda, Y. Tomo, N. Hayami, T. Terada, M. Shirouza, T. Osanai, A. Tanaka, M. Seki, Shinozaki and S. Yokoyama, "Solution Structure of the B3 DNA Binding Domain of the Arabidopsis Cold-Responsive Transcription Factor RAV1," *Plant Cell,* vol. 16, no. 12, pp. 3448-3459, 2004.

[97] K. Yamasaki, "Structures, Functions, and Evolutionary Histories of DNA-Binding Domains of Plant-Specific Transcription Factors," in *Plant Transcription Factors*, Academic Press, 2016, pp. 57-69.

[98] D. R. TerBush, T. Maurice, D. Roth and P. Novick, "he Exocyst is a multiprotein complex required for exocytosis in Saccharomyces cerevisiae," *The EMBO journal,* vol. 15, no. 23, pp. 6483-6494, 1996.

[99] J. Yu, S. Tehrim, L. Wang, K. Dossa, X. Zhang, T. Ke and B. Liao, "Evolutionary history and functional divergence of the cytochrome P450 gene superfamily between Arabidopsis thaliana and Brassica species uncover effects of whole genome and tandem duplications," *BMC Genomics,* vol. 18, no. 733, 2017.

[100] M. Valifard, R. Le Hir, J. Muller, D. Scheuring, H. E. Neuhaus and B. Pommerrenig, "Vacuolar fructose transporter SWEET17 is critical for root development and drought tolerance," *Plant Physiology,* 2021.

[101] G. Beckmann, J. Hanke, P. Bork and J. G. Reich, "Merging extracellular domains: fold prediction for laminin G-like and amino-terminal thrombospondin-like modules based on homology to pentraxins," *Journal of Molecular Biology,* vol. 275, no. 5, pp. 725-730, 1998.

[102] S. Cherqui, V. Kalatzis, G. Trugnan and C. Antignac, "The targeting of cystinosin to the lysosomal membrane requires a tyrosine-based signal and a novel sorting motif," *The Journal of biological chemistry,* vol. 276, no. 16, pp. 13314-13321, 2001.

[103] G. J. Narlikar, R. Sundaramoorthy and T. Owen-Hughes, "Mechanisms and Functions of ATP-Dependent Chromatin-Remodeling Enzymes," *CellPress,* vol. 154, no. 3, pp. 490-503, 2013.

[104] Y. G. Yu, G. R. Buss and S. M. Maroof, "Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 93, pp. 11751-11756, 1996.

[105] A. F. Bent, B. N. Kunkel, D. Dahlbeck, K. L. Brown, R. Schmidt, J. Giraudat, J. Leung and B. J. Staskawicz, "RPS2 of Arabidopsis thaliana: a Leucine-Rich Repeat Class of Plant Disease Resistance Genes," *Science,* vol. 265, no. 5180, pp. 1856-1860, 1994.

[106] S. Rocak and P. Linder, "DEAD-box proteins: the driving forces behind RNA metabolism," *Nature Reviews Molecular Cell Biology,* vol. 5, pp. 232-241, 2004.

[107] C. Gomez-Mena, M. Pineiro, J. M. Franco-Zorrilla, J. Salinas, G. Coupland and J. M. Martinez-zapater, "early bolting in short days: An Arabidopsis Mutation That Causes Early-Flowering and Partially Suppresses the Floral Phenotype of leafy," *The Plant Cell,* vol. 13, pp. 1011-1024, 2001.

[108] I. Callebaut, J.-C. Courvalin and J.-P. Mornon, "The BAH (bromo-adjacent homology) domain: a link between DNA methylation, replication and transcriptional regulation," *FEBS Letters,* vol. 446, no. 1, pp. 189-193, 1999.

[109] A. Singh, P. K. Singh, A. K. Sharma, N. K. Singh, h. Sonah, R. Deshmukh and T. R. Sharma, "Understanding the Role of the WRKY Gene Family under Stress Conditions in Pigeonpea (Cajanus Cajan L.)," *Plants,* vol. 8, no. 7, p. 214, 2019.

[110] B. Liu, H. Du, R. Rutkowski, A. Gartner and X. Wang, "LAAT-1 is the Lysosomal Lysine/Arginine Transporter that Maintains Amino Acid Homeostasis," *Science,* vol. 337, no. 6092, pp. 351-354, 2012.

[111] N. Denance, B. Szurek and L. D. Noel, "Emerging Functions of Nodulin-Like Proteins in Non-Nodulating Plant Species," *Plant & Cell Physiology,* vol. 55, no. 3, pp. 469-474, 2014.

[112] S. Lu, J. Wang, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, M. Yang, D. Zhang, C. Zheng, C. J. Lanczycki and A. Marchler-Bauer, "CDD/SPARCLE: the conserved domain database in 2020," *Nucleic acids research,* vol. 48, no. D(1), pp. D265-D268, 2020.

[113] B. Li, T. Kamiya, L. Kalmbach, M. Yamagami, K. Yamaguchi, S. Shigenobu, S. Sawa, J. M. Danku, D. E. Salt and N. F. Geldner, "Role of LOTR1 in Nutrient Transport through Organization of Spatial Distribution of Root Endodermal Barriers," *Current Biology,* vol. 27, no. 5, pp. 758-765, 2017.

[114] C. U. Schrader, L. Lee, M. Rey, V. Sarpe, P. Man, S. Sharma II, V. Zabrouskove II, B. Larsen and D. C. Schriemer, "Neprosin, a Selective Prolyl Endoprotease for Bottom-up Proteomics and Histone Mapping," *Molecular & Cellular Proteomics,* vol. 16, no. 6, pp. 1162-1171, 2017.

[115] T. Pan, T. P. Michael, M. E. Hudson, S. A. Kay, J. Chory and M. A. Schuler, "Cytochrome P450 Monooxygenases as Reporters for Circadian-Regulated Pathways," *Plant Physiology,* vol. 150, no. 2, pp. 858-878, 2009.

[116] A. W. Munro, H. M. Girvan and K. J. McLean, "ytochrome P450--redox partner fusion enzymes," *Biochimica et Biophysica Acta,* vol. 1770, no. 3, pp. 345-359, 2007.

[117] D. W. Meinke, "Genome-wide identification of EMBRYO-DEFECTIVE (EMB) genes required for growth and development in Arabidopsis," *New Phytologist,* vol. 226, no. 2, pp. 289-291, 2019.

[118] A. Mouriz, L. Lopez-Gonzalez, J. A. Jarillo and M. Pineiro, "PHDs govern plant development," *Plant signaling & behavior,* vol. 10, no. 7, 2015.

[119] J. Lyu, Z. Cai, Y. Li, H. Suo, R. Yi, S. Zhang and H. Nian, "The Floral Repressor GmFLC-like Is Involved in Regulating Flowering Time Mediated by Low Temperature in Soybean," *International Journal of Molecular Sciences,* vol. 21, no. 4, 2020.

[120] Y. Y. Levy, S. Mesnage, J. S. Mylne, A. R. Gendall and C. Dean, "Multiple Roles of Arabidopsis VRN1 in Vernalization and Flowering Time Control," *Science,* vol. 297, no. 5579, pp. 243-246, 2002.

[121] J. Lu, H. Suo, R. Yi, Q. Ma and H. Nian, "Glyma11g13220, a homolog of the vernalization pathway gene VERNALIZATION 1 from soybean [Glycine max (L.) Merr.], promotes flowering in Arabidopsis thaliana," *BMC Plant Biology,* vol. 15, no. 232, pp. 1-12, 2015.

[122] E. R. Cober, D. W. Steward and H. D. Voldeng, "Photoperiod and Temperature Responses in Early-Maturing, Near-Isogenic Soybean Lines," *Crop Science,* vol. 41, no. 3, pp. 721-727, 2001.

[123] W. Deng, M. Casao, P. Wang, K. Sato, P. M. Hayes, J. Finnegan and B. Trevaskis, "Direct links between the vernalization response and other key traits of cereal crops," *Nature Communications,* vol. 6, no. 5882, 2015.

[124] G. Patil, B. Valliyodan, R. Deshmukh, S. Prine, B. Nicander, M. Zhao, H. S.-L. Song, L. Lin, J. Chaudhary, Y. Liu, T. Joshi, D. Xu and H. T. Nguyen, "Soybean (Glycine max) SWEET gene family: insights through comparative genomics, transcriptome profiling and whole genome re-sequence analysis," *BMC Genomics,* vol. 16, no. 1, 2015.

[125] P. A. Klemens, K. Patzke, J. Deitmer, L. Spinner, R. Le Hir, C. Bellini, M. Bedu, F. Chardon, A. Krapp and H. Neuhaus, "Overexpression of the Vacuolar Sugar Carrier AtSWEET16 Modifies Germination,," *Plant Physiology,* vol. 163, no. 3, pp. 1338-1352, 2013.

[126] Y. Chi, Y. Yang, G. Li, F. Wang, B. Fan and Z. Chen, "Identification and characterization of a novel group o flegume-specific, Golgi apparatus-localized WRKY and Exo70 proteins from soybean," *Journal of Experimental Botany,* vol. 66, no. 11, pp. 3055-3070, 2015.

[127] Z. Wang, P. Li, Y. Yang, Y. Chi, B. Fan and Z. Chen, "Expression and Functional Analysis of a Novel Group of Legume-specific WRKY and Exo70 Protein Variants from Soybean," *Scientific Reports,* 2016.

[128] M. Pineiro, C. Gomex-Mena and R. Schaffer, "EARLY BOLTING IN SHORT DAYS Is Related to Chromatin Remodeling Factors and Regulates Flowering in Arabidopsis by Repressing FT," *The Plant Cell,* vol. 15, no. 7, pp. 1552-1562, 2003.

[129] R. Huang, D. Liu, M. Huang, J. Ma, Z. Li, M. Li and S. Sui, "CpWRKY71, a WRKY Transcription Factor Gene of Wintersweet (Chimonanthus praecox), Promotes Flowering and Leaf Senescence in Arabidopsis," *International journal of molecular sciences,* vol. 20, no. 21, 2019.

[130] X. Luo, X. Bai, X. Sun, D. Zhu, B. Liu, W. Ji, H. Cai, L. Cao, J. Wu, M. Hu, X. Liu, L. Tang and Y. Zhu, "Expression of wild soybean WRKY20 in Arabidopsis enhances drought tolerance and regulates ABA signalling," *Journal of Experimental Botany,,* vol. 64, no. 8, pp. 2155-2169, 2013.

[131] D. Martignago, B. Siemiathkowska, A. Lombardi and L. Conti, "Abscisic Acid and Flowering Regulation: Many Targets, Difference Places," *International Journal of Molecular Sciences,* vol. 21, no. 24, pp. 1-14, 2020.

[132] Y.-S. Noh and R. M. Amasino, "PIE1, an ISWI Family Gene, Is Required for FLC Activation and Floral Repression in Arabidopsis," *The Plant cell,* vol. 15, no. 7, pp. 1671-1682, 2003.

[133] S. Watanabe, K. Harada and J. Abe, "Genetic and molecular bases of photoperiod responses of flowering in soybean," *Breeding Science,* vol. 61, no. 5, pp. 531-543, 2012.

[134] K. Bok Ma, S.-J. Yang, Y.-S. Jo, S. S. Kang and M. Nam, "Development of Kompetitive Allele Specific PCR markers for identification of persimmon varieties using genotyping-by-sequencing," *Electronic Journal of Biotechnology,* vol. 49, pp. 72-81, 2021.

[135] BioSearch Technologies Genomic Analysis by LGC, "KASP genotyping manual," 2021. [Online]. Available: https://biosearch-cdn.azureedge.net/assetsv6/KASP-genotyping-chemistry-User-guide.pdf.

[136] P. Lestari and H. J. Joh, "Development of New CAPS/dCAPS and SNAP Markers for Rice Eating Quality," *HAYATI Journal of Biosciences,* vol. 20, no. 1, pp. 15-23, 2013.

[137] K. Ilic, T. Berleth and N. J. Provart, "BlastDigester – a web-based program for efficient CAPS marker design," University of Toronto, Toronto, 2004.

Appendix

Appendix 1. The following, indicate the location of amino acid variations between the wild type (E7) and mutant type (e7) genotypes for all candidates in subgroups 1, 2 and 3. Pairwise alignment was done using Expasy SIM-alignment tool (https://web.expasy.org/sim/). * indicates consensus alignment between the E7 and e7 genotypes, while empty spots indicate the amino acid at that position is not consistent between the E7 and e7 genotypes.

Glyma.06G180300

```
180300-MT    481 LEULEUASNSERTHRLYSLEUTYRILETYRILETYRILETYRILETYRILETYRILETYR
180300-WT    481 LEULEUASNSERTHRLYSILETYRILETYRILETYRILETYRILETYRILETYRILETYR
                 ******************   ****************************************
```

Glyma.06G199800

```
199800-MT   1981 SNSTOPLYSGLUGLUPROPROLEUPROPROPROPROPROPROPROASNSTOPPROGLYPH
199800-WT   1981 SNSTOPLYSGLUGLUPROPROLEUPROPROPROPROPROPROPROILEASNPROVALLEU
                 ********************************************      *    *

199800-MT   2041 EVALSTOPASNPHELYSASNLEULEUILEARGARGILELYSILEPHEPHEPHESERPHES
199800-WT   2041 PHELYSILESERARGILECYSSTOPSERVALGLUSTOPLYSTYRPHESERPHEPROPHEA
                    *     *                    * *       ***   ***   ***   ***

199800-MT   2101 ERSERLYSLYSILEPHEPHEASNPROLEUPROLYSILEILESERPHEARGLEULYSARGL
199800-WT   2101 RGGLNLYSLYSPHEPHELEUTHRARGPHEPROLYSSTOP-PHELEUSERASPSTOPARGG
                      ******  ****       *    ******       *     *       ***

199800-MT   2161 YSLYSARGGLUGLYSTOPSTOPTHRARGGLUGLUV-ALARGGLUARGGLUGLYLYSARGG
199800-WT   2160 LULYSARGGLUARGALAAS--PARGARGGLUARGARGSTOPGLUARGGLUARGGLUARGG
                   *********       *  *   ******          *********      ****

199800-MT   2220 LUVALARGALAARGPHEPROHISLYSPROASNLEULEUSERLEUTYRLEUPHESERTHRG
199800-WT   2218 LUARGTYRGLUARGASPSERHISTHRASNGLNILESERTYRLEUTYRILETYRPHEPROH
                 **       * ***      ***        *    *   *******

199800-MT   2280 LNLYSILESERHISPROHISSERGLYPROGLYVALL-EULYSARGGLYGLUGLYGLULEU
199800-WT   2278 ISLYSLYSSERHISILEHISTHRLEUALAARGVALCYSSTOPARGGLUGLULYSGLYSER
                   ***   ******   ***  *          ***       ***** ***   **  *

199800-MT   2339 VALCYSVALALAPHESERLEULEULEUTHRVALCYSGLNARGSTOPTHRGLYTHRILELY
199800-WT   2338 TRPPHEVALT-RPHISP-----HELEUTYR---------PHESTOP---GLNPHEVALLY
                      ***   ** *       **** *            ****   **  *    **

199800-MT   2399 SHISTRPSTOPILETHRASPLEUVALGLNGLNGLUGLUASNTYRARGARGGLUVALASNL
199800-WT   2380 ---SASPLYSP-ROVALPROLEUASNTHRGLYLYSSERLEUTHRSERTYRASNLYSARGL
                   *  *   *          ***      **          * *             *  *

199800-MT   2459 YSLEUTHRGLYCYSLEUSTOPILES---TOPILELYSTYRTYRILEPROASNARGLYSTH
199800-WT   2436 YSTHRTHRGLUGLUARGSTOPILESERSTOPGLNGLYVALTYRLYSTYRARGLEUASNTH
                 **   *****       ********   *** *       ***      *        **

199800-MT   2516 RTRPSTOPVALHISLYSASNTHRLEUPHEPROTYRSERLYSHISALASTOPLEUPROLEU
199800-WT   2496 RILEPHEGLNTHRGLUARG----LEUGLYLYSSERTHRLYSILEHISSERPHEPROILE-
                 *          *  *        ***        *  ****      *  * *    **
```

Glyma.06G200200

```
200200-MT   1861 LEVALVALASNCYSTYRLEUTHRSTOPASNILEPROLEUASNSTOPSERLEUTHRGLUTY
200200-WT   1861 LEVALVALASNCYSTYRLEUSERSTOPASNILEPROLEUASNSTOPSERLEUTHRGLUTY
                 ********************  ****************************************
```

Glyma.06G200400

```
200400-MT   1621 ETRPPHETYRSTOPVALARGTHRILESERVALSERVALLEUSERPHETHRSERCYSTHRL
200400-WT   1621 ETRPPHETYRSTOPVALARGTHRILESERVALSERVALLEUTHRPHETHRSERCYSTHRL
                 ****************************************  ****************
```

Glyma.06G200800

```
200800-MT    781 STYRLYSPHEGLYARGGLYTYRCYSGLYCYSGLYGLYGLY-------------------
200800-WT    781 STYRLYSPHEVALARGGLYTYRCYSGLYCYSGLYGLYGLYTHRARGARGGLUTHRLEUAR
                 **********   ****************************

200800-MT    821 -------------ALAARGARGGLUTHRLEUARGLEUTHRHISHISSTOPSTOPARGSTO
200800-WT    841 GLEUTHRHISGLYALAARGARGGLUTHRLEUARGLEUTHRHISHISSTOPSTOPARGSTO
                         ************************************************
```

Glyma.06G200300

```
202300-MT   1741 TYRSTOPLEUMETPHETYRASNTYRLEULEUSERILEASPSTOPILETHRPHETYRSTOP
202300-WT   1741 TYRSTOPLEUMETPHETYRASNTYRLEULEUASNILEASPSTOPILEALAPHETYRSTOP
                 *******************************   ************   **********
```

Glyma.06G220000

```
220000-M      61 YRVALASNLEUASNARGGLUVALARGLYSASNASNGLUVALHISGLUILEGLNLYSSTOP
220000-WT     61 YRVALASNLEUASNARGGLUVALARGLYSASNASNASPVALHISGLUILEGLNLYSSTOP
                 ***********************************   ********************

220000-WT   1081 EUILEPHEPHELEUPHEGLYGLYLEUPROLEUPHEASNLEUILELEULYSGLUVALGLYT
                 *****   ********* * ****************************************

220000-M    1261 ARGVALPHEILEGLUASNLYSGLUPHESERTRPHISARGVALPHEPHEPHELEUSERVAL
220000-WT   1261 ARGVALPHEILEGLUASNLYSGLUSERSERTRPHISARGVALPHEPHEPHELEUSERVAL
                 ************************   ********************************

220000-M    1981 GLNSTOPHISLYSVALLEULYSALALEUTHRPROCYSTYRILETYRILETYRILETYRIL
220000-WT   1981 GLNSTOPHISLYSVALLEULYSALALEUTHRPROSERTYRILETYRILETYRILETYRIL
                 ***********************************   ********************
```

Glyma.06G233300

```
233300-MT   4801 HESERGLYGLYARGGLYGLYGLYGLYGLYPHEPHEPHEPHECYSPROTRPGLYSERGLNA
233300-WT   4801 HESERGLYGLYARGGLYGLYGLYGLYGLYPHEPHEPHEPHEPHEPROTRPGLYSERPROA
                 ****************************************   ************   *

233300-MT   4861 SPPHEGLYPHE-----------------------------------------TYRGLYS
233300-WT   4861 SPPHEGLYPHEPHEGLYGLUGLYGLYTYRPHELEUGLYGLYASPPHEGLYTYRTYRGLYS
                 **********                                             *******
```

Glyma.06G242200

```
242200-MT   1741 EUARGTHR-TYRLEUA-RGLEU-----ASPSTOPSER-LEUVALPROASNHISGLNILEA
242200-WT   1741 EUARGTHRSTOPLEUSERGLNTHRVALPROSTOPLEUGLNILEILELYSSERASPVALAL
                 ******** *   ***  ***          **** *  *         *        *

242200-MT   1793 RGCYSARGILELEUCYSPHEVALLEUTHRASNLEUPHEVALVALILETYRVALALAGLUG
242200-WT   1801 ASERCYSVALSERTYRSTOPLEUILESERLEUSTOPSTOPTYRMETSTOPPROASNLYSG
                       *      *           *      *           *     *     *

242200-MT   1853 LNTHRILEPHELEUGLYCYSLYSASNGLYSTOPSTOPGLNLEUGLUSTOPMETASNVALL
242200-WT   1861 LNSERPHETRPVALVALLYSMETASPAS-PASPSERSTOPASNGLUSTOPTHRSTOPASN
                 **  *  *           **   **       **          *******       *

242200-MT   1913 YSSTOPGLNLYSSERGLULEUASNGLNILELYSILELYSLEUGLYSERSTOPSTOPLEUG
242200-WT   1920 ASNLYSASNARGASNSTOPILELYSLEUARGSTOPASNTRPVALVALASNSERTYRLYSG
                  *      *                 *                   *    *    * *  *

242200-MT   1973 LNARGLYSTYR-LEUARGASNLYSASPLEUASNTRPASPARGLEUSTOPSERARGASPSE
242200-WT   1980 LUSERILETYRGLYILELYSTHRSTOPILEGLYTHRASP-------TYRASPGLNGLYTH
                 *       *** *       *   *  *       *  ***       *

242200-MT   2032 RASNTYRGLUSERLYSSTOPALALYSLEULYSPHEMETASNCYSSTOPVALTHRGLNLYS
242200-WT   2033 RGLNTHRMETSERLEUSERARGGLNASNSTOPASNLEUSTOPTHRALAARGSERARGLYS
                 *  ** *   ****  *       *             *                  *   ***

242200-MT   2092 V--ALVALSTOPHISLYSGLYASNLEUTYRLYSHISILESERPHESTOPPROGLYPHEPH
242200-WT   2093 ARGSERPHEASPILELYSGLYTHRTYRILEASNILEPHELEUPHEASNL-EUALAPHESE
                       *   ******                  * * ***        * ***

242200-MT   2150 EILEASPTYRPROPHEARGLEULEUPHESTO--PLEUSERPHEILEHISG------LNLE
242200-WT   2152 RLEUILETHRPROLEUGLYCYSTYRSERASNCYSLEULEUTYRILEASNSERSERTYRLY
                      * ****                       *** *    ***            *

242200-MT   2202 UILEVALSTOPASNARGASNVALA--LAPROSERSTOPTYRSERTYRLEULEUIL-EVAL
242200-WT   2212 SILEVALTHRSERLEUHISHISSERILEARGTHRTYRSTOPLEUTYRPHEVALARGSERL
                  ******           *       *  *   *    *   * ***            *

242200-MT   2259 LEUCYSSERGLU--STOPVALVALPROASPSERCYSALAASNSTOPASNSTOPGLYTHRL
242200-WT   2272 YSSTOPPHEGLNILEARGVALGLNTHRLYSILELYSVALARGLYSPHEVALLYSVALPHE
                       **       ***             **   *     *
```

Appendix 2. RNA sequence data provided by Severin et al., RNA atlas (https://soybase.org/soyseq/). Digital gene expression counts across 14 tissues, including young leaf, flower, root and nod, at different growth stages, including pod shell development 10 days after flowering to 14 days after flowering, and seed formation from 10 days after flowering to 42 days after flowering.

Wm82.a2.v1	**young_leaf**	**flower**	**one cm pod**	**pod shell 10DAF**	**pod shell 14DAF**	**seed 10DAF**	**seed 14DAF**	**seed 21DAF**	**seed 25DAF**	**seed 28DAF**	**seed 35DAF**	**seed 42DAF**	**root**	**nod**
Glyma.06g180300	36	14	19	31	27	6	14	4	15	4	10	4	17	2
Glyma.06g199800	72	60	50	48	38	27	21	7	23	14	27	20	43	31
Glyma.06g202300	437	122	215	273	307	144	271	204	412	261	59	10	8	0
Glyma.06g233300	95	74	77	49	47	29	38	26	57	39	63	42	32	49
Glyma.06g242200	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Glyma.06g196400	28	30	30	48	56	13	14	8	60	28	63	14	39	53

Appendix 3. Housekeeping genes and primers used for expression analysis. * primers have been used to date.

Gene	**Locus Name**	**Primer Sequence F (5'-3')**	**Primer Sequence R (5'-3')**
TUB4	*Glyma.03G124400*	AGCTGGTCAATGTGGAAACC	AAGCACAGCTCGAGGAACAT
N/A*	*Glyma.06G180300*	CTTGCAGGGCACTAACCACA	AATTGTCTTGACGGGAGGCT
N/A*	*Glyma.06G199800*	ATGTGCTGAACGACGCATAG	GTGCCAGTAGCACCCTTAATAC
N/A*	*Glyma.06G233300*	CTCCGACCACTACGATGTGC	CAGCACCCACATTCTCAAGC
N/A*	*Glyma.06G239100*	GACCCTCAATTACAGTCCTACTA	GAGAAACTGAGACAACATTGAGATCAG
N/A	*Glyma.06G200800*	GAGAAGCTCGTGGTACTCAATTA	GGCAGAGAGAGGTGAGAAATAC
N/A	*Glyma.06G220000*	TTTCAGGATCATAATTGCTCCC	GAGTATTTGGTAAGCCTTCTCCG
N/A	*Glyma.06G200400*	AGGGGTCCATTCCCTTTGAG	CGGGCAATCCAGGATAACTG
N/A	*Glyma.06G242200*	GATGACGATGCAAAGAACTG	ACCTCACTCAAAGTGCAACC

Appendix Table 4. Primers used for sequencing

Primer Name	***Sequence (5' – 3')***
260100-F1	TGCACTTACGACAGTCAACA
260100-F2	GGAAGCACCGACGAAATGAT
260100-F3	GTTTTCCACGAACCTCCCTT
260100-SR1	CAGGCACCAAGTTGAGGAAG
260100-SF2	CTTCCTCAACTTGGTGCCTG
260100-SR2	TCCCAACCAGAGAGATTGCC
260100-SF3	GGCAATCTCTCTGGTTGGGA
260100-SR3	GGAACATCATTAACCGGCCC
260100-SF4	GGGCCGGTTAATGATGTTCC
260100-SR4	CGAAACCAATGCACTCCTCC
260100-SF5	GGAGGAGTGCATTGGTTTCG
260100-SR5	CAACCTGCTCCACTTCCATG
260100-SF6	CATGGAAGTGGAGCAGGTTG
260100-SR6	ACTACCTTTGTGCCTTCCCA
260100-SF7	TGGGAAGGCACAAAGGTAGT
260100-SR7	TCGTGAAAGAACAACGGACAG
260100-SF8	CTGTCCGTTGTTCTTTCACGA
260100-SR8	CTTCTTGAGGTTGGGCAGTG
260100-SF9	CACTGCCCAACCTCAAGAAG
260100-SR9	CACCTGGTACTGCAGAGACA
260100-SF10	TGTCTCTGCAGTACCAGGTG
260100-SR10	CCGTATTTCTTCACCTCCGC
260100-SF11	GCGGAGGTGAAGAAATACGG
260100-R1	AAGCCACAAGTAACAGATAATAGATTT
260100-R2	AACACGAAAAATAATGTAGGGAAT
260100-R3	TTTTTAATTTGTGAATCGGAAAAA
208300-F1	TGCGCGTCACCCCATATATA
208300-F2	TCCCTCGTCCGCATATAACA
208300-F3	TGCATAAACAACACCTGGTCG
208300-SR1	GAGCGCTACAATGGTGAAGG
208300-SF2	CCTTCACCATTGTAGCGCTC
208300-SR2	CATGCGTGGATGATAGCTGG
208300-SF3	CCAGCTATCATCCACGCATG
208300-SR3	CACATTGCAACGGTGGCTTA
208300-SF4	TAAGCCACCGTTGCAATGTG
208300-SR4	CTGGGTGGTGTAGTTGGAGT
208300-SF5	ACTCCAACTACACCACCCAG

208300-R1	CGTACCTAAGCCGATTCAAACA
208300-R2	TTGAGAAAGAGCATGTAGGGTTC
208300-R3	AGTCTTGTCTTTAAGGGAGAGTG
258500-F1	AGTATCACAAGGACCGCTGC
258500-F2	TTGTGGCGGCTAAGACTGTA
258500-F3	GCGCTGTACCCTTCCTCTTA
258500-SR1	TGTTGAGACCCGTTGCTACT
258500-SF2	AGTAGCAACGGGTCTCAACA
258500-SR2	GCCATAAGCGAAGTGACCAG
258500-SF3	CTGGTCACTTCGCTTATGGC
258500-SR3	CCGGTTTTGTGGCGATTTTG
258500-SF4	CAAAATCGCCACAAAACCGG
258500-SR4	CAATGCCCTCGAGTTCCTCT
258500-SF5	AGAGGAACTCGAGGGCATTG
258500-SR5	TCAAATGATCTCGGTCGGGT
258500-SF6	ACCCGACCGAGATCATTTGA
258500-SR6	AACAGAGGTGCAAAGAACGAG
258500-SF7	CTCGTTCTTTGCACCTCTGTT
258500-SR7	AATGCTTGCGACCACAACC
258500-SF8	GGTTGTGGTCGCAAGCATT
258500-SR8	ATACGCATGCCACCAAACAA
258500-SF9	TTGTTTGGTGGCATGCGTAT
258500-SR9	AGCATCCCATGTTACAAGCT
258500-SF10	AGCTTGTAACATGGGATGCT
258500-SR10	TGCCGATGTATCTGAAAGCC
258500-SF11	GGCTTTCAGATACATCGGCA
258500-SR11	CTCCTCTACACACTCGAGCT
258500-SF12	AGCTCGAGTGTGTAGAGGAG
258500-R1	GTGTGCATAGTCCTTTTGAGCT
258500-R2	GGTTGGAAATTGGGTGGTTGA
258500-R3	GGTCCCAAAAGTCTCAAGCC
258600-F1	AGTCTTTGAGCTTGCACCAC
258600-F2	GGGCCATTGTGATGTGAGAC
258600-F3	AGATGAGCCGGAAATAGCCA
258600-SR1	GGTGTTGGTCATGGATTGCA
258600-SF2	TGCAATCCATGACCAACACC-
258600-SR2	GTGCAGATAATGAAGGGGCG
258600-SF3	CGCCCCTTCATTATCTGCAC
258600-SR3	CTTAGCCGCCACAATGTCAG

258600SF2B	TCCCCCAAGGAAGGAATAAC
258600SF2C	GACAATACACGCAAGGTCCA
258600SF2D	CTGAACGCTATCCCCCAAG
258600SR2B	GCAGCGGTCCTTGTGATACT
258600SR3C	TACAGTCTTAGCCGCCACAA
258600-SF4	CTGACATTGTGGCGGCTAAG
258600-R1	GGTGAGTGTGTCCGTTTCAG
258600-R2	TTCCTAGGTAGCGGTTGGTG
258600-R3	TGGTGTTGGTGTTGTTGGTG
205800-F1	CCGAGGAAGTAAACACGCAA
205800-F2	GCCTTTGTCCAATCCGAACT
205800-F3	ACCACCTTCTCACTTTCACTT
205800-SR1	TTCAACTGCACCCTTCCTCT
205800-SF2	AGAGGAAGGGTGCAGTTGAA
205800-SR2	TGCCTTTGGTGGAGAAGACT
205800-SF3	AGTCTTCTCCACCAAAGGCA
205800-SR3	TGACAGAGACCCATCCATAAGC
205800-SF4	GCTTATGGATGGGTCTCTGTCA
205800-SR4	TTTGTGTCTGAGGCTGAGGT
205800-SF5	ACCTCAGCCTCAGACACAAA
205800-SR5	AGTGGTGCATATACTCGAGTGT
205800-SF6	ACACTCGAGTATATGCACCACT
205800-SR6	TGACTTGCCTGATTTCAAACTCT
205800-SF7	AGAGTTTGAAATCAGGCAAGTCA
205800-SR7	GGACCTCGCCTCTTGAAAATC
205800-SF8	GATTTTCAAGAGGCGAGGTCC
205800-SR8	CGACTGGGATATATGAGGCG
205800-SF9	CGCCTCATATATCCCAGTCG
205800-SR9	CATGGTGGGATGGGTAACTT
205800-SF10	AAGTTACCCATCCCACCATG
205800-SR10	TGGTCAACATCACAGTCATGC
205800-SF11	GCATGACTGTGATGTTGACCA
205800-SR11	AAAATGATCGTAGGTGGCGC
205800-SF12	GCGCCACCTACGATCATTTT
205800-SR12	AATGAGAGCTCGTGGGTTCA
205800-SF13	TGAACCCACGAGCTCTCATT
205800-SR13	GTCTGCCTTTCCTACTGTGC
205800-SF14	GCACAGTAGGAAAGGCAGAC
205800-SR14	GCACTGCATAGCCACTTCAA

205800-SF15	TTGAAGTGGCTATGCAGTGC
205800-SR15	TCCATTTCATCACCACGCTG
205800-SF16	CAGCGTGGTGATGAAATGGA
205800-SR16	ACTGACCCATTGATCGTGCT
205800-SF17	AGCACGATCAATGGGTCAGT
205800-R1	GTCGTCGTAAACATCCACACT
205800-R2	AGAAGCGGTATACTCCAAGGA
205800-R3	TGACCAAACCCGACTTCCAA
202100-F1	GACATGTCTAAATCTCCCACCA
202100-F2	TTCCAAGACCCTACCCAACC
202100-F3	TCAGCCACGCCAATTTACAA
202100-SR1	GGTTAGGGTTCGTGGAATCG
202100-SF2	CGATTCCACGAACCCTAACC
202100-SR2	TTCGTCGACCTCTTCTTCCC
202100-SF3	GGGAAGAAGAGGTCGACGAA
202100-SR3	CCGTAACTCCCTTCCTCCTC
202100-SF4	GAGGAGGAAGGGAGTTACGG
202100-SR4	TTCACGCACACACACTACAC
202100-SF5	GTGTAGTGTGTGTGCGTGAA
202100-SR5	AAGGCAAGAAGTTACGCACA
202100-SF6	TGTGCGTAACTTCTTGCCTT
202100-SR6	ATGATAAGAACGCGTGCACC
202100-SF7	GGTGCACGCGTTCTTATCAT
202100-SR7	TTTGCGCGGATATTTTGGCT
202100-SF8	AGCCAAAATATCCGCGCAAA
202100-SR8	AGGCCTCTCATCTTCCTCCT
202100-SF9	AGGAGGAAGATGAGAGGCCT
202100-SR9	ACAAGCACACAGAGGACACA
202100-SF10	TGTGTCCTCTGTGTGCTTGT
202100-SR10	CACTATCGTACACAACCGCC
202100-SF11	GGCGGTTGTGTACGATAGTG
202100-SR11	TGTCCCAGGCTCTGCTTTAT
202100-SF12	ATAAAGCAGAGCCTGGGACA
202100-R1	TTTGGCCGTTTCCCTTGAAC
202100-R2	TGTTGTTGAAGGTGCACTCG
202100-R3	GCGAGTGATGGCATTGACAA
178200-F1	TCACCGGATGCAAAATATGTTGA
178200-F2	CCCCTTTGAAACAAAGTTCCAG
178200-F3	TCTGCATAAGATGTGGCTACTG

178200-SR1	CTGTCACCACCCTCTTGCTA
178200-SF2	TAGCAAGAGGGTGGTGACAG
178200-SR2	GGGAGAAAGGAAGAGGGGAG
178200-SF3	CTCCCCTCTTCCTTTCTCCC
178200-SR3	GCAGCACGTTTTGGCATTTT
178200-SF4	AAAATGCCAAAACGTGCTGC
178200-SR4	TACCACTCATAGACCGCACC
178200-SF5	GGTGCGGTCTATGAGTGGTA
178200-SR5	TTCTTCCTGCAACAGTCGTG
178200-SF6	CACGACTGTTGCAGGAAGAA
178200-SR6	TGACAGTTTGATGCCCAGGT
178200-SF7	ACCTGGGCATCAAACTGTCA
178200-SR7	CCTTGTCGTTGTAGCTTGCA
178200-SF8	TGCAAGCTACAACGACAAGG
178200-SR8	CATGCAGTCCCAAATCCCTG
178200-SF9	CAGGGATTTGGGACTGCATG
178200-SR9	GCAGACACCAGACCCTAGAC
178200-SF10	GTCTAGGGTCTGGTGTCTGC
178200-SR10	TGTGTGCATGGGTTGATTCA
178200-SF11	TGAATCAACCCATGCACACA
178200-SR11	ACACGCTTGTCTAACAGGGA
178200-SF12	TCCCTGTTAGACAAGCGTGT
178200-SR12	TATGAGCCGGCAACTGTCAA
178200-SF13	TTGACAGTTGCCGGCTCATA
178200-SR13	TGAACACCAATGCCTCAGGA
178200-SF14	TCCTGAGGCATTGGTGTTCA
178200-SR14	TGAGAAACCCCAAATGCTCC
178200-SF15	GGAGCATTTGGGGTTTCTCA
178200-SR15	GCACCTGGTCTCTTGGGAAT
178200-SF16	ATTCCCAAGAGACCAGGTGC
178200-SR16	ACTGTCCACTGCTTCCTTGA
178200-SF17	TCAAGGAAGCAGTGGACAGT
178200-R1	GCTCCAAAGGTGTTAGTGATTGA
178200-R2	CAATTTTGGGCAGTGTCTAGC
196200-F1	TTGGGCGGAGTGAGAAAGAT
196200-F2	TGAGAAGTGGGAGAATATGTCAC
196200-SR1	CCTCACCAGACTCCATACCC
196200-SF2	GGGTATGGAGTCTGGTGAGG
196200-SR2	AAACCAAACCACCTCAGCAC

196200-SF3	GTGCTGAGGTGGTTTGGTTT
196200-SR3	GGAAGGTCAAGTTCAGCCAG
196200-SF4	CTGGCTGAACTTGACCTTCC
196200-SR4	CCAATACAAACCATGCGCCT
196200-SF5	AGGCGCATGGTTTGTATTGG
196200-SR5	GGGTAACTAACAGTGTTCGGC
196200-SF6	GCCGAACACTGTTAGTTACCC
196200-SR6	TTTGCAAAGTACGCCCCAAG
196200-SF7	CTTGGGGCGTACTTTGCAAA
196200-SR7	AGGAGACACACCACCACTTT
196200-SF8	AAAGTGGTGGTGTGTCTCCT
196200-SR8	AGACGGGCTAAAAGGTGAGA
196200-SF9	TCTCACCTTTTAGCCCGTCT
196200-SR9	GGATGCATCACACCCTGTTG
196200-SF10	CAACAGGGTGTGATGCATCC
196200-SR10	TCTCTTGCTCCCTGGTCATC
196200-SF11	GATGACCAGGGAGCAAGAGA
196200-R1	GCCTTGGATGGTGGTCAAAT
196200-R2	ACTAACATCAACCGGGTAAACA
196200-R3	CAAAGCTTTCCTCTCCTAAGTCC
215400-F1	AATCGGACGGTTCACATTGC
215400-F2	TGTCCTTGGTGCATTTGAAAAG
215400-F3	GTCCCATCCTAACAGTCTCTCT
215400-SR1	AAATGGCGTCGAAGTAAGCG
215400-SF2	CGCTTACTTCGACGCCATTT
215400-SR2	GCTTGTGCCGACCATATAGC
215400-SF3	GCTATATGGTCGGCACAAGC
215400-SR3	ATGGACACAAGAACCGGAGA
215400-SF4	TCTCCGGTTCTTGTGTCCAT
215400-SR4	ATGTGCGTATCAAGAAGCCAC
215400-SF5	GTGGCTTCTTGATACGCACAT
215400-SR5	TGCATTCCTGACTCCTTTGC
215400-SF6	GCAAAGGAGTCAGGAATGCA
215400-SR6	ACTGACAGACATCCCAACCA
215400-SF7	TGGTTGGGATGTCTGTCAGT
215400-SR7	AGGGCCTCTTTACGTCCAAA
215400-SF8	TTTGGACGTAAAGAGGCCCT
215400-SR8	AGCAGAACACTTCAGAGGCA
215400-SF9	TGCCTCTGAAGTGTTCTGCT

215400-SR9	TCTTCCATCGACTTAGCCCA
215400-SF10	TGGGCTAAGTCGATGGAAGA
215400-SR10	TTTCTTTTCATATCTGCAAGCCA
215400-SF11	TGGCTTGCAGATATGAAAAGAAA
215400-SR11	ACACCTTCTCTCTCCTGCTT
215400-SF12	AAGCAGGAGAGAGAAGGTGT
215400-SR12	AACCACCCTCCAAATCCACA
215400-SF13	TGTGGATTTGGAGGGTGGTT
215400-SR13	AGGAGGAGGGTTGAGGAAGA
215400-SF14	TCTTCCTCAACCCTCCTCCT
215400-SR14	GGAAACGATCAACCAAGGCA
215400-SF15	TGCCTTGGTTGATCGTTTCC
215400-SR15	CAAACCATGTGAGCCGTTCC
215400-SF16	GGAACGGCTCACATGGTTTG
215400-R1	CAACATCTCTATTGTCTCGTCGT
215400-R2	GGAAGATCGTGTTCATCGGT
215400-R3	AGGGTGTCAAACGGTACAAA
214900-F1	CACACTTCACCAGCCATACCT
214900-F2	AATTTTGACGAACAAGTAAAATGA
214900-F3	TAAAAGAAGGGCTCCAGGTG
214900-SR1	CCCCTTTCCCTGACATCCTT
214900-SF2	AAGGATGTCAGGGAAAGGGG
214900-SR2	ACAACAGGAACAAGAGGTCCA
214900-SF3	TGGACCTCTTGTTCCTGTTGT
214900-SR3	ATGTGGTTCTGCTTCTGTGT
214900-SF4	ACACAGAAGCAGAACCACAT
214900-SR4	GAATCACTCCTCCTCCAGCA
214900-SF5	TGCTGGAGGAGGAGTGATTC
214900-SR5	CAACAGCAGCAACTCAGTCA
214900-SF6	TGACTGAGTTGCTGCTGTTG
214900-R1	ACCCAAATACCGTCTCCTCT
214900-R2	TATCGCAATCAAAGGGGACG
214900-R3	CCATGTCGCTACTACTTCGC
195700-F1	TCGAGAAACAGGGGAATTTG
195700-F2	TGTTTGCATTGCTTAGGACA
195700-F3	CCAACGTGGATTATGATGACC
195700-SR1	AGCGTGTAGGCTCTGACGAT
195700-SF2	ATCGTCAGAGCCTACACGCT
195700-SR2	ATCAAAAACTTCCACAGCAGA

195700-SF3	TCTGCTGTGGAAGTTTTTGAT
195700-SR3	ATAACGCTTGCTGGCACTCT
195700-SF4	AGAGTGCCAGCAAGCGTTAT
195700-SR4	TTGGAAAATTGAGCACAACG
195700-SF5	CGTTGTGCTCAATTTTCCAA
195700-SR5	ATAACGCTTGCTGGCACTCT
195700-SF6	AGAGTGCCAGCAAGCGTTAT
195700-SR6	TTGGAAAATTGAGCACAACG
195700-SF7	CGTTGTGCTCAATTTTCCAA
195700-SR7	GCAAGGTATATGGCGAGCTC
195700-SF8	GAGCTCGCCATATACCTTGC
195700-SR8	ATTTGGCACCTGAAATTTCG
195700-SF9	CGAAATTTCAGGTGCCAAAT
195700-SR9	TATCACTTACGCTCCCCACC
195700-SF10	GGTGGGGAGCGTAAGTGATA
195700-SR10	AGGATCACAGCCTAGCTCCA
195700-SF11	TGGAGCTAGGCTGTGATCCT
195700-SR11	ATGTTTCTTGGACCCTGCAC
195700-SF12	GTGCAGGGTCCAAGAAACAT
195700-SR12	GGAATATCAGTCACACCCCG
195700-SF13	CGGGGTGTGACTGATATTCC
195700-SR13	TTCTGTGTCATCCGGAACA
195700-SF14	TGTTCCGGATGACACAGAA
195700-SR14	GCATAAGGGGTATGTGCCTC
195700-SF15	GAGGCACATACCCCTTATGC
195700-R1	TTCCTCCCAATCTCAATTCC
195700-R2	TCATAGGAACCCAAGTTGAAAAA
195700-R3	TGGAGGAAACAAGGAGATGG
255600-F1	CCGCATGAATTCACCTTCTT
255600-F2	CACAGCAACGCGAATTACAA
255600-F3	TGGTAACGAAAATGCCTTCA
255600-SR1	TCTTCCAAGTGCAACTGCAG
255600-SF2	CTGCAGTTGCACTTGGAAGA
255600-SR2	CTGATTTCCCTCCACTGCAT
255600-SF3	ATGCAGTGGAGGGAAATCAG
255600-R1	AAGGCATGTCTTGACCCCTA
255600-R2	ATGGCGAACTGTGCATTGTA
255600-R3	AGGCCCTGGTCTATCCTTTG

Appendix Table 5. Genome Quebec sequencing analysis for the E7 lines (OT93-26 and OT89-9) and e7 lines (OT02-18 and OT98-17) on chromosome 6, using Illumina NovaSeq 600 S4 PE150. Note: 0/0 (the sample is homozygous reference), 0/1: the same is heterozygous, carrying 1 copy of each of the REF and ALT alleles, 1/1 (the sample is homozygous alternative). ./. were identified as being inconclusive.

CHROM	POS	REF	ALT	Harosoy	OT89-5	OT02-18	OT98-17	Gene (Wm82.a2.v1)	Region
6	15289734	T	A	./.	./.	'1/1	'1/1	*GLYMA.06G180300*	intron
6	17632800	C	CT	./.	./.	1/1	1/1	*GLYMA.06G196400*	UTR
6	18285668	A	AG	./.	'0/0	1/1	1/1	*GLYMA.06G199800*	intron
6	18329753	A	T	./.	0/0	'1/1	'1/1	*GLYMA.06G200200*	intron
6	18349606	T	A	./.	0/0	'1/1	'1/1	*GLYMA.06G200400*	exon
6	18404037	A	C	./.	0/0	'1/1	'1/1	*GLYMA.06G200800*	exon
6	18404061	CACCA…	C	./.	'0/0	1/1	1/1	*GLYMA.06G200800*	exon
6	18732840	C	T	./.	0/0	'1/1	'1/1	*GLYMA.06G202300*	intron
6	18732854	T	C	./.	0/0	'1/1	'1/1	*GLYMA.06G202300*	intron
6	25283550	C	G	'1/1	'1/1	0/0	0/0	*GLYMA.06G220000*	intron
6	25284517	C	A	'1/1	'1/1	0/0	0/0	*GLYMA.06G220000*	exon
6	25284530	G	C	'1/1	'1/1	0/0	0/0	*GLYMA.06G220000*	exon
6	25284716	A	G	'1/1	'1/1	0/0	0/0	*GLYMA.06G220000*	intron
6	25285427	C	G	'1/1	'1/1	0/0	0/0	*GLYMA.06G220000*	intron
6	37299885	A	C	./.	0/0	'1/1	'1/1	*GLYMA.06G233300*	intron
6	37299895	C	A	./.	./.	'1/1	'1/1	*GLYMA.06G233300*	intron
6	37299898	G	A	./.	./.	'1/1	'1/1	*GLYMA.06G233300*	intron
6	37299899	GGCC	G	./.	./.	1/1	1/1	*GLYMA.06G233300*	intron
6	37299903	TGAAA…	T	./.	./.	1/1	1/1	*GLYMA.06G233300*	intron
6	37299954	T	A	./.	./.	'1/1	'1/1	*GLYMA.06G233300*	intron
6	39218016	CT	C	./.	./.	1/1	1/1	*GLYMA.06G239100*	intron
6	39218028	C	G	./.	./.	'1/1	'1/1	*GLYMA.06G239100*	intron

Appendix Table 6. Representative PIPE raw data for candidate *Glyma.06G200400* (*Glyma06G21600*), along with its top 200 predicted interacting partners (protein-b), and their respective PIPE score.

protein_b	PIPE_score
Glyma17g03740	0.830
Glyma01g25810	0.791
Glyma12g29300	0.789
Glyma18g46050	0.787
Glyma11g29790	0.783
Glyma18g12970	0.782
Glyma20g34900	0.781
Glyma14g06550	0.777
Glyma09g23230	0.773
Glyma10g27740	0.772
Glyma16g22900	0.770
Glyma08g09880	0.759
Glyma17g34280	0.758
Glyma01g03490	0.756
Glyma02g04150	0.755
Glyma03g24430	0.755
Glyma01g03490	0.754
Glyma02g04150	0.753
Glyma04g35880	0.750
Glyma16g29160	0.750
Glyma01g05260	0.749
Glyma03g20130	0.748
Glyma11g03080	0.747
Glyma01g42280	0.746
Glyma08g09880	0.746
Glyma09g25140	0.737
Glyma12g14490	0.737
Glyma08g23010	0.736
Glyma04g15020	0.736
Glyma05g02370	0.735
Glyma17g09530	0.734
Glyma19g40530	0.732
Glyma18g42810	0.728
Glyma18g16760	0.727
Glyma14g34560	0.725
Glyma12g20210	0.724
Glyma18g03250	0.722
Glyma03g20390	0.722
Glyma17g25690	0.721
Glyma10g12950	0.721
Glyma19g22190	0.721
Glyma18g46100	0.720
Glyma18g35850	0.719
Glyma14g27600	0.719
Glyma08g16220	0.719
Glyma18g40890	0.718
Glyma12g10700	0.717
Glyma15g09360	0.716
Glyma19g26040	0.715
Glyma03g21560	0.714
Glyma06g39510	0.712
Glyma18g24250	0.711
Glyma04g02360	0.709
Glyma06g38040	0.709
Glyma03g07080	0.708
Glyma10g24280	0.707
Glyma01g34140	0.706
Glyma02g42600	0.706
Glyma06g28320	0.706
Glyma09g34840	0.704
Glyma18g16270	0.702
Glyma05g33640	0.702
Glyma16g32660	0.702
Glyma20g20110	0.701
Glyma07g16470	0.701
Glyma02g02540	0.700
Glyma02g02540	0.700
Glyma02g02540	0.700
Glyma05g01360	0.700
Glyma05g01360	0.700
Glyma01g04950	0.700
Glyma04g12440	0.700
Glyma17g10520	0.700
Glyma17g10520	0.700
Glyma05g01360	0.699
Glyma09g27670	0.699
Glyma02g02500	0.699
Glyma05g33640	0.699
Glyma13g05120	0.699
Glyma13g05120	0.699

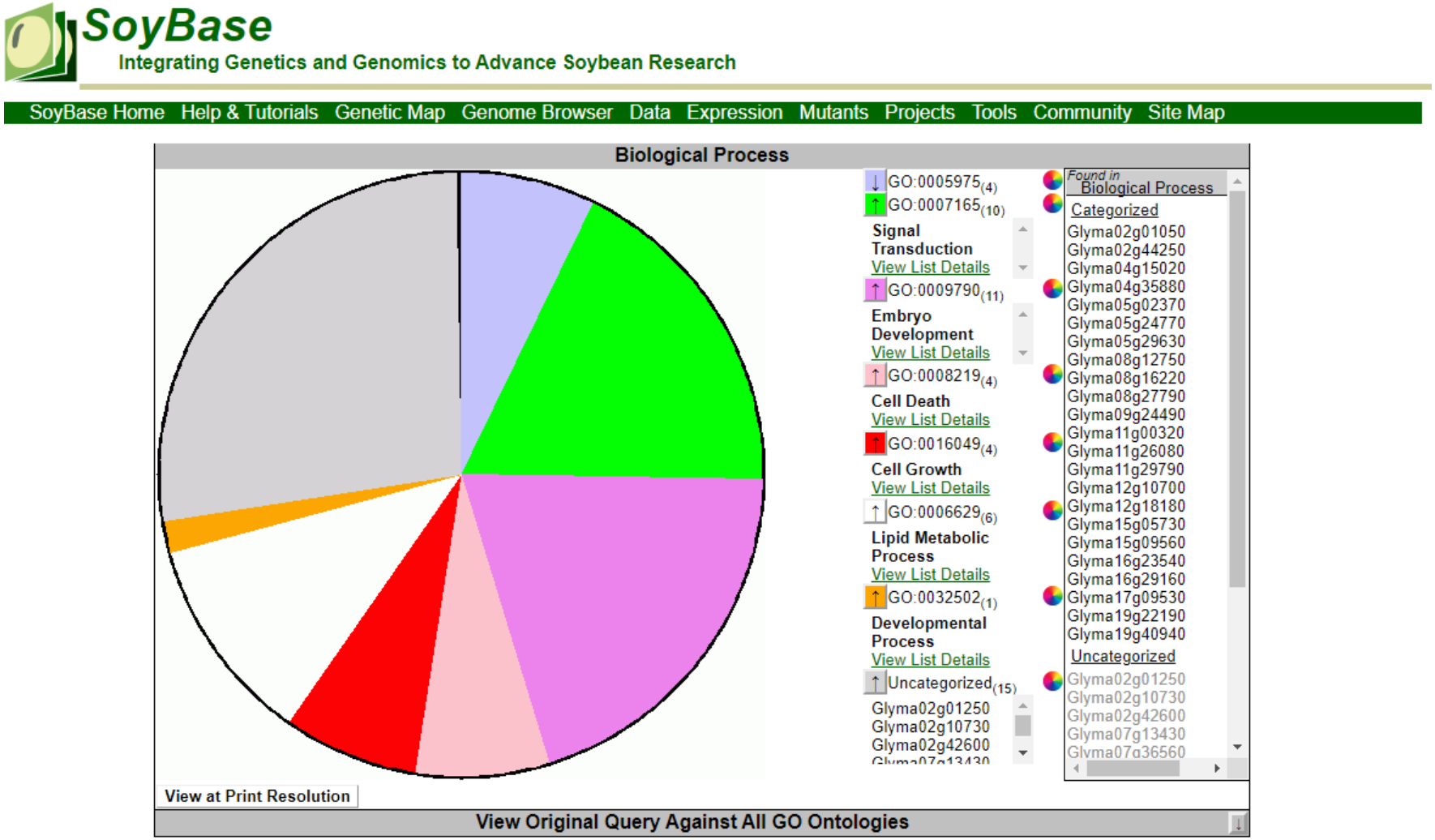
Glyma19g02370	0.699
Glyma19g02370	0.699
Glyma19g02370	0.699
Glyma08g06090	0.698
Glyma02g02550	0.697
Glyma05g33640	0.695
Glyma06g40460	0.695
Glyma06g34930	0.694
Glyma05g33630	0.693
Glyma01g43240	0.693
Glyma02g19850	0.693
Glyma12g18180	0.692
Glyma02g44250	0.692
Glyma02g01250	0.692
Glyma14g17830	0.691
Glyma08g06090	0.691
Glyma14g37220	0.691
Glyma05g33640	0.690
Glyma05g33640	0.690
Glyma09g09290	0.689
Glyma12g34600	0.689
Glyma06g20840	0.689
Glyma01g10220	0.689
Glyma17g21230	0.688
Glyma20g29100	0.688
Glyma01g10100	0.688
Glyma02g42350	0.687
Glyma13g07060	0.687
Glyma02g14160	0.687
Glyma08g27790	0.685

Glyma10g38650	0.683
Glyma08g09910	0.683
Glyma10g20150	0.683
Glyma11g35090	0.683
Glyma07g34960	0.683
Glyma02g10730	0.682
Glyma08g18850	0.682
Glyma08g18850	0.682
Glyma11g32170	0.681
Glyma15g06170	0.681
Glyma05g33630	0.681
Glyma08g39750	0.681
Glyma18g51330	0.680
Glyma17g04060	0.680
Glyma09g36460	0.679
Glyma19g05200	0.679
Glyma17g21210	0.677
Glyma10g23860	0.675
Glyma11g31590	0.674
Glyma09g24490	0.673
Glyma08g06090	0.673
Glyma14g21310	0.673
Glyma14g09210	0.672
Glyma05g25830	0.671
Glyma12g00890	0.670
Glyma18g11690	0.668
Glyma13g07060	0.667
Glyma20g31320	0.666
Glyma0349s00200	0.665
Glyma11g26080	0.665

Glyma08g28380	0.663
Glyma08g09900	0.660
Glyma15g37900	0.659
Glyma08g09910	0.658
Glyma10g36490	0.657
Glyma05g24770	0.655
Glyma02g13320	0.655
Glyma16g23540	0.654
Glyma07g13430	0.654
Glyma14g24170	0.654
Glyma05g26930	0.653
Glyma07g01590	0.653
Glyma15g09560	0.652
Glyma13g16540	0.652
Glyma08g21100	0.652
Glyma17g06140	0.652
Glyma05g29630	0.651
Glyma12g33290	0.651
Glyma09g02120	0.651
Glyma16g08550	0.651
Glyma17g28670	0.651
Glyma08g12750	0.650
Glyma07g36560	0.650
Glyma07g36560	0.650
Glyma20g05140	0.649
Glyma17g00570	0.649
Glyma18g05290	0.648
Glyma15g40320	0.648
Glyma13g16540	0.648
Glyma10g36280	0.648

Glyma20g19640	0.647
Glyma11g00320	0.647
Glyma1605s00200	0.647
Glyma20g04640	0.647
Glyma17g28240	0.646
Glyma05g01050	0.646
Glyma10g25440	0.645
Glyma19g40940	0.645
Glyma10g01300	0.645
Glyma18g49260	0.644
Glyma10g23080	0.644
Glyma09g23970	0.643
Glyma20g17080	0.643
Glyma16g24400	0.643
Glyma20g31080	0.643
Glyma02g01050	0.643
Glyma20g12090	0.643
Glyma10g28880	0.643
Glyma10g39340	0.642
Glyma09g19830	0.642
Glyma13g30050	0.642
Glyma06g34900	0.641
Glyma15g05730	0.641
Glyma11g16070	0.640
Glyma11g16070	0.640
Glyma11g16070	0.640
Glyma16g28780	0.640
Glyma18g52090	0.640
Glyma04g38900	0.640
Glyma02g08360	0.640

Appendix Figure 1. Representative GO identified for the top 200 interacting partners identified from the raw PIPE data for candidate *Glyma.06G200400* (*Glyma06G21600*) using www.soybase.org. Focused on GO related to Biological Processes including Multicellular Organismal Development, Flowering, Embryo Development and other developmental processes related to time of flowering and maturity.



Appendix Figure 2. I am the second author, and the lead author for the biological analysis of this Human-Soybean Allergies paper, contributing to 60% of the written portion.

Journal of proteome research

pubs.acs.org/jpr

Article

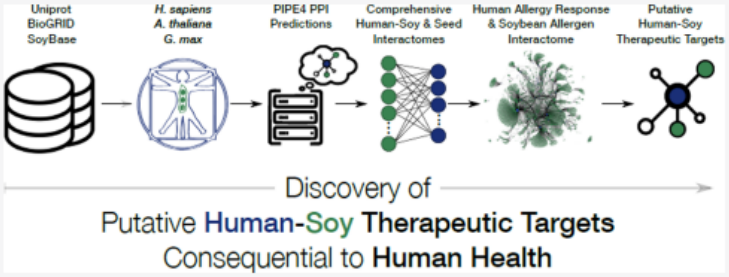
Human–Soybean Allergies: Elucidation of the Seed Proteome and Comprehensive Protein–Protein Interaction Prediction

Kevin Dick, Arezo Pattang, Julia Hooker, Nour Nissan, Michael Sadowski, Bradley Barnes, Le Hoa Tan, Daniel Burnside, Sadhna Phanse, Hiroyuki Aoki, Mohan Babu, Frank Dehne, Ashkan Golshani, Elroy R. Cober, James R. Green, and Bahram Samanfar*

Cite This: *J. Proteome Res.* 2021, 20, 4925–4947

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: The soybean crop, *Glycine max* (L.) Merr., is consumed by humans, *Homo sapiens*, worldwide. While the respective bodies of literature and *-omics* data for each of these organisms are extensive, comparatively few studies investigate the molecular biological processes occurring between the two. We are interested in elucidating the network of protein–protein interactions (PPIs) involved in human–soybean allergies. To this end, we leverage state-of-the-art sequence-based PPI predictors amenable to predicting the enormous comprehensive interactome between human and soybean. A network-based analytical approach is proposed, leveraging *similar interaction profiles* to identify candidate allergens and proteins involved in the allergy response. Interestingly, the predicted interactome can be explored from two complementary perspectives: *which soybean proteins are predicted to interact with specific human proteins* and *which human proteins are predicted to interact with specific soybean proteins*. A total of eight proteins (six specific to the human proteome and two to the soy proteome) have been identified and supported by the literature to be involved in human health, specifically related to immunological and neurological pathways. This study, beyond generating the most comprehensive human–soybean interactome to date, elucidated a soybean seed interactome and identified several proteins putatively consequential to human health.

KEYWORDS: *soybean, human allergies, seed proteome, mass spectrometry, protein–protein interactions, computational biology, PIPE4*