Moving Soybean to Western Canada and Northern Regions, an Attempt to Identify the Underlying Gene for the E7 Maturity Locus

Arezo Pattang[1,2], Elroy Cober[1], Stephen J. Molnar[1], Ashkan Golshani[2], Bahram Samanfar[1,2]

[1] Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, [2] Carleton University, Department of Biology

Abstract

Soybean is one of the largest sources of vegetable oil and protein globally, its multipurpose characteristics has made it of environmental and economic significance. While the expansion of soybean to North Western regions of Canada is dependent on many factors, time of flowering and maturity is the key to the adaptability of soybean into regions with longer days. Currently, among the 11 maturity loci identified to be controlling time of flowering and maturity in soybean, the underlying gene(s) for E7 (and E8) remain unknown. Through the integration of various "-omics" approaches, a short list of 17 candidate genes on chromosome 6 are to be further examined. So far, among investigated candidates, only one has shown to contain an amino acid variation in the form of a SNP between E7/e7 genotypes. While sequencing data for all candidates has yet to be released, follow-up experiments, including expression analysis, compensation analysis using *Arabidopsis*, etc., will be performed. Last but not least, this will lead to the development of allele-specific marker(s) for the underlying gene for E7 locus, which will significantly contribute to the soybean breeding programs through marker-assisted selection to further develop ultra-early cultivars.

Introduction

While Canada has seen more than a 100% increase in soybean production from 2009 to 2020, the majority of soybean production remains to be in southern Ontario and Quebec (Figure 1). To further increase soybean production across Canada, with an emphasis made on Western and Northern Canada the crop will need to overcome challenges and adapt to the region while also maintaining reasonably (high) yield and seed quality.

These cultivars would need to overcome challenges posed by the photoperiod of the region, as the long days attributed to Western and Northern Regions of Canada delay time of flowering and maturity preventing maturation from occurring before first frost. In addition these cultivars would need to withstand abiotic stresses including drought, flooding, iron deficiency chlorosis (IDC) and salinity, as well as biotic stresses including diseases, and pests

Figure 1: Soybean growing areas in Canada [1]

To date, eleven maturity loci (known as the E-series) have been identified to be associated with time of flowering and maturity in soybean, in addition to many other genes involved in time of flowering and maturity but not classified under the E-nomenclature. Among the E-series, the underlying genes for E1-E4, E6/J, E9-E11 have been identified, while the underlying gene(s) for E7 and E8 remain unknown. The dominant alleles at E1-E4, E7, E8 and E10 confer late flowering, whereas the dominant alleles at E6, E9, E11 and *J* confer late flowering.

Methodology

While the gap in our understanding of the mechanisms involved in time of flowering and maturity is one of the limitations to further enhance soybean expansion, this research project is focused on contributing to the continued efforts in understanding these pathways and mechanisms in more detail. Specifically, focusing on identifying the underlying gene for the E7 locus, and to develop user-friendly and accurate allele-specific marker for breeding programs developing ultra-early maturity soybeans. This project aims to incorporate various –omics approaches along with bioinformatics, computational approaches and molecular biology related practices to address the proposed objective. Including; PIPE (Protein-Protein Interaction Prediction Engine), gene ontology, SNP and RNA sequence database investigations.

The eight soybean lines used in this study (Table 1), are presumed to have E7 vs. e7 genotype. The late maturing E7 allele, sensitive to long photoperiods, contrasting to the early maturing e7 allele with less sensitivity towards long photoperiods, flowering ~13 to 17 days earlier.

Table 1: Genotypes of soybean lines used in this study

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

Methodology Cont'd

Experimental approach to identifying potential candidate gene within the E7 region (Figure 2), through the integration of various functional genomics approaches including a computational approach; PIPE (Protein-protein Interaction Prediction Engine) (Figure 3), paired with gene ontology (Figure 4) along with RNA sequence and SNP database analysis (Figure 5). A shortlist of candidate genes were further assessed through the integration of molecular biology practices, including sequencing analysis to identify consistent variations between E7 and e7 lines with additional expression analysis and compensation analysis to be performed to assist in confirming the candidate gene as the E7 candidate.

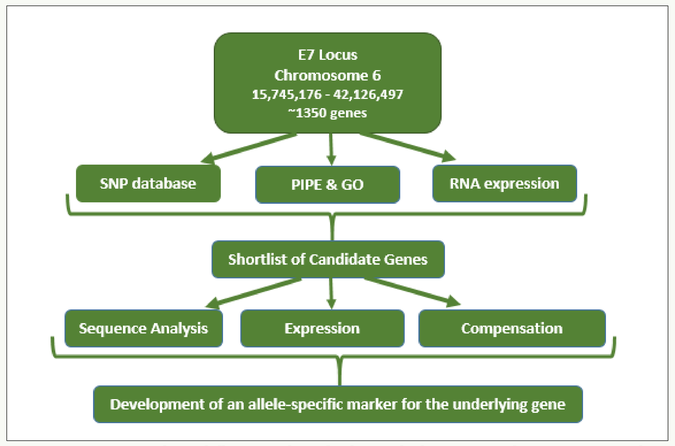

Figure 2: Experimental workflow to identify the underlying gene for E7 maturity locus

PIPE is a computational tool that predicts PPIs based on short reoccurring amino acid sequences and a known interaction database [2]. A simplified algorithm for PIPE is presented in (Figure 3), provided that a database of 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 the known interacting data set shifting by one amino acid every round until the end of protein A is reached. let's assume, it is 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 confirm if A and B are interacting or not (at a given specificity(99.9%) and sensitivity(23.3%)) [2].

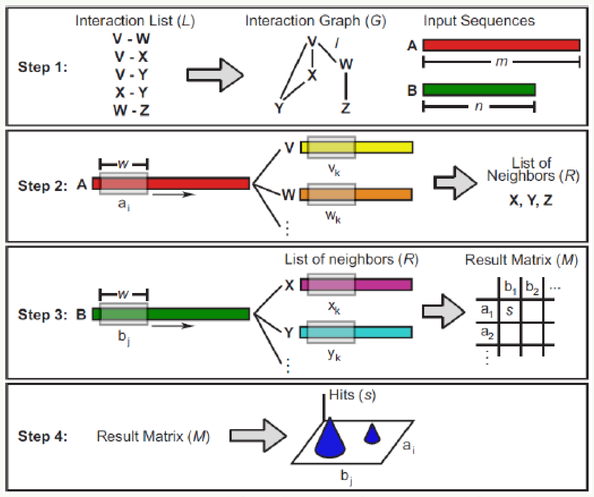

Figure 3: A simplified algorithm for PIPE workflow [3]

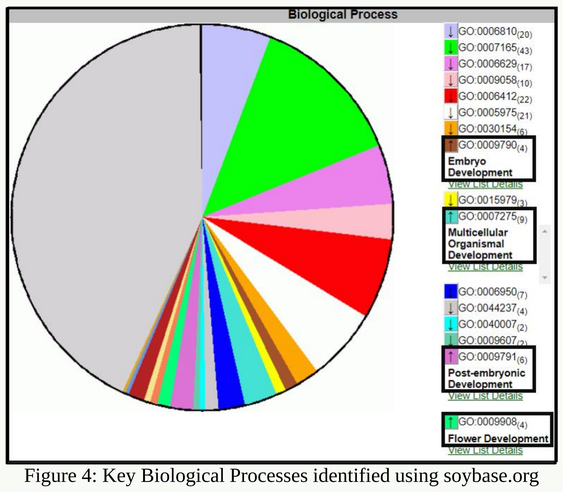

Figure 4: Key Biological Processes identified using soybase.org

	Location			
Corrected_Names	20,321,936	20,321,946	43,550,939	43,573,333
OT02-18 (e7)	G	C	A	C
OT94-47 (e7)	G	C	A	C
Harosoy (E7)	A	T	G	T
Harosoy (E7)	A	T	G	T
A	46.58%	0.00%	1.71%	0.00%
G	42.02%	0.00%	96.77%	0.00%
C	0.00%	42.02%	0.00%	1.52%
T	0.00%	46.58%	0.00%	96.96%
Harosoy63	A	T	G	T
Evans	A	T	G	T
OAC_Prudence	A	T	G	T
90A07	A	T	G	T
Lynx_RR	G	C	G	T
PR9368B25	G	C	G	T
etc.	A	T	G	T

Figure 5: Selected SNPs with consistent variation between E7 and e7 from the SNP database selected

Results & Discussion

The top candidates #1 - #14 were selected based on PIPE score, total interactions involved in development and flowering, the presence of SNPs and expression in vital organs prior and post flowering. In addition, candidates #15 - #17 were selected based on recommendations from collaborators and recent publications (Table 2).

To date, candidates #4, #8, #12, #14, #15, and #16 have been fully sequenced, while no variation between E7 and e7 alleles was seen for candidates #4, #8, #12, #15 and #16, a SNP was identified in the coding region of candidate #14. The e7 allele possesses a thymine, instead of a cytosine that found in the E7 allele (Figure 6).

While candidate #14 has shown to be a strong contender, additional sequencing data for the remaining candidates has yet to be released, and could be included in additional analysis to identify the candidate gene.

Table 2: Short-list of candidate genes for E7 maturity locus

	Gene	Function
1	Glyma.06g2*****	Leucine-Rich Repeat Receptor-Like Protein Kinase
2	Glyma.06g2*****	Leucine-Rich Repeat – Containing protein
3	Glyma.06g2*****	Tetraspanin
4	~~Glyma.06g2*****~~	~~Leucine-Rich Repeat Receptor-Like Protein Kinase~~
5	Glyma.06g2*****	MADS BOX PROTEIN
6	Glyma06g2*****	TATA-Binding protein-Associated phosphoprotein
7	Glyma.06g2*****	Leucine-Rich Repeat Receptor-Like Protein Kinase
8	~~Glyma.06g1*****~~	~~methyltransferase 1~~
9	Glyma.06g1*****	CCT motif -containing response regulator protein
10	Glyma06g2*****	actin-related protein 6
11	Glyma.06g2*****	protein arginine methyltransferase 10
12	~~Glyma.06g2*****~~	~~histone H2A protein 9~~
13	Glyma.06g1*****	N/A
14	Glyma.06g2*****	ANCIENT UBIQUITOUS PROTEIN
15	~~Glyma.06G3*****~~	~~Far-red elongated hypocotyls 3~~
16	~~Glyma.06G1*****~~	~~Phytochrome interacting factor7~~
17	Glyma.06G1*****	Protein of unknown function

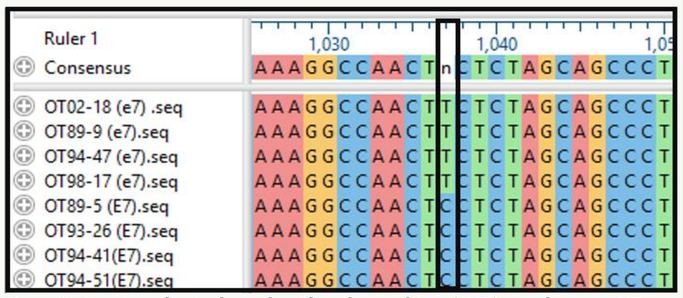

Figure 6: Consistently single nucleotide polymorphism (SNP) seen between E7 and e7 alleles for candidate #14

Future Direction

While all candidates have yet to be fully sequenced, additional analysis will be conducted on selected gene(s) found to consist an allelic variation between E7 and e7, these include:

- Expression Analysis: (qRT-PCR & ddPCR)
  On soybean grown in different photoperiod (LD and SD), with expression analysis conducted on tissue samples harvested before flowering, during flowering, and after flowering

- Compensation Analysis
  Transform candidate gene alleles into *Arabidopsis*, observe physiological traits associated with E7 locus, including time of flowering and maturity

Acknowledgement

I would like to thank Martin Charette, Doris Luckert and the present and past members of Samanfar lab for their support through out this project. Special thanks to Carleton University, Agriculture and Agri-Food Canada and CFCRA for funding opportunities.

References

[1] SoyCanada, "Canadian Soybean Seeded Hectares (1980 to current)," 2019. [Online]. Available: https://soycanada.ca/statistics/seeded-area-hectares/. [2] Z. Ding and D. Kihara, "Computational Identification of Protein-Protein Interactions in Model Plant Proteomes," *Scientific Reports*, 2019. [3] Pitre et al., "Computational Methods for Predicting Protein-Protein Interactions," *Advances in Biochemical Engineering/Biotechnology*, vol. 111, pp.247-267, 2008.